Effects of Nucleotide Sequence Alignment on Phylogeny Estimation: A Case Study of 18S rDNAs of Apicomplexa

David A. Morrison and John T. Ellis

Molecular Parasitology Unit, University of Technology Sydney

The reconstruction of phylogenetic history is predicated on being able to accurately establish hypotheses of character homology, which involves sequence alignment for studies based on molecular sequence data. In an empirical study investigating nucleotide sequence alignment, we inferred phylogenetic trees for 43 species of the Apicomplexa and 3 of Dinozoa based on complete small-subunit rDNA sequences, using six different multiple-alignment procedures: manual alignment based on the secondary structure of the 18S rRNA molecule, and automated similarity-based alignment algorithms using the PileUp, ClustalW, TreeAlign, MALIGN, and SAM computer programs. Trees were constructed using neighbor-joining, weighted-parsimony, and maximum-likelihood methods. All of the multiple sequence alignment procedures yielded the same basic structure for the estimate of the phylogenetic relationship among the taxa, which presumably represents the underlying phylogenetic signal. However, the placement of many of the taxa was sensitive to the alignment procedure used; and the different alignments produced trees that were on average more dissimilar from each other than did the different tree-building methods used. The multiple alignments from the different procedures varied greatly in length, but aligned sequence length was not a good predictor of the similarity of the resulting phylogenetic trees. We also systematically varied the gap weights (the relative cost of inserting a new gap into a sequence or extending an already-existing gap) for the ClustalW program, and this produced alignments that were at least as different from each other as those produced by the different alignment algorithms. Furthermore, there was no combination of gap weights that produced the same tree as that from the structure alignment, in spite of the fact that many of the alignments were similar in length to the structure alignment. We also investigated the phylogenetic information content of the helical and nonhelical regions of the rDNA, and conclude that the helical regions are the most informative. We therefore conclude that many of the literature disagreements concerning the phylogeny of the Apicomplexa are probably based on differences in sequence alignment strategies rather than differences in data or tree-building methods.

Introduction

One of the essential steps in the reconstruction of phylogenetic history is establishing hypotheses of character and character-state homology among the taxa being studied, and mistaken hypotheses of homology are thus a primary source of error in evolutionary studies. Concepts of homology are often intuitively obvious when dealing with phenotypic data. For example, characters and their states can be postulated as homologous on the basis of their structural, positional, ontogenetic, compositional, and/or functional correspondences, and they can be postulated between different taxa so as to maximize the number of one-to-one correspondences of their parts (Stevens 1984). For analyses using molecular sequence data, the assessment of homology involves alignment of the nucleotides or amino acids. Alignment of molecular sequences thus consists of a series of hypotheses of homology among the taxa, with one hypothesis of homology for each position (nucleotide or amino acid) in the sequences (i.e., we hypothesize that the nucleotides or amino acids at each position are descended from the same residue in a common ancestral sequence). The concepts of homology in molecular and in morphological studies are thus fundamentally the same (de Pinna 1991; Williams 1993). However, for molecular data it seems that there is little possibility of further investigations (such as ontogeny) to assess homology, and so in practice homology assessment is very different for molecular studies (Mindell 1991). In this paper, we investigate some of the consequences for phylogeny reconstruction of using different strategies for aligning nucleotide sequences, based on a specific empirical example.

Sequence Alignment

Positional homology in orthologous sequences can be represented by either identical character states (nucleotides or amino acids) in all of the sequences, substitutions in one or more of the sequences (representing point mutations), or insertions/deletions (indels) in one or more of the sequences. The most problematic aspect of sequence alignment is the positioning of indels, and this problem becomes more acute for more divergent taxa. It is worthwhile in this context to distinguish between gaps, which are introduced into the sequences during the alignment process, and indels, which are actual mutation events (Olsen 1988); clearly, the objective is to introduce into the sequences only gaps that truly represent indels (and this might almost be taken as a definition of the sequence alignment procedure).

There are two possible processes for sequence alignment. First, the alignment can be constructed by hand. This is possible, for instance, when there are apparently relatively few indels needed to align the sequences. This situation is shown, for example, by protein-coding parts of mtDNA (Miyamoto and Cracraft 1991) and the plastid *rbc*L gene (Chase et al. 1993). Alternatively, the sequences may have as their product a molecule for which there is an a priori biological mod-

Key words: phylogenetics, sequence alignment, neighbor joining, parsimony, maximum likelihood, Apicomplexa, rRNA.

Address for correspondence and reprints: David A. Morrison, Molecular Parasitology Unit, University of Technology Sydney, Westbourne Street, Gore Hill NSW 2065, Australia. E-mail: davidm@bio.uts.edu.au.

Mol. Biol. Evol. 14(4):428–441. 1997

el of secondary structure or function in which certain active sites must be maintained; the alignment can then be constrained by the model (Olsen 1988; Olsen and Woese 1993). Such a situation is shown, for example, by rRNA genes (Kjer 1995; Hickson et al. 1996).

Second, it may not be possible to produce a robust alignment by hand, usually due to the low percent identity between the sequences (<50%; Schulze-Kremer 1996). Under these circumstances, it is usual to use a mathematical algorithm to produce the alignment. These algorithms attempt to produce a sequence alignment that optimizes some chosen criterion of match between the individual sequences (cost). That is, the sequences are compared using a pattern-matching process that searches for correspondence between the elements of the sequences, introducing gaps into the sequences as required to optimize some criterion for correspondence (usually minimizing the cost). There are many algorithms currently available (see Waterman 1989; Doolittle 1990; Chan, Wong, and Chiu 1992; McClure, Vasi, and Fitch 1994) which optimize a variety of mathematical functions measuring the overall alignment cost. When there are more than two sequences, most of these algorithms use exact procedures (which are guaranteed to find the optimal solution) to align the sequences pairwise but then use heuristic procedures (computationally efficient strategies that should produce a solution that is at least close to the optimal one) to braid these pairwise alignments into a multiple alignment (Hirosawa et al. 1995). Thus, not all of these procedures are guaranteed to produce the globally optimal alignment. Furthermore, they do not guarantee that the optimal alignment (even if they could find it) represents the true alignment (Thorne and Kishino 1992), as their procedures are based on maximizing sequence similarity, which is not necessarily the same as sequence homology (similarity can be the result either of common ancestry or of chance convergence, parallelism, or reversal).

Irrespective of the alignment procedure used, when dealing with the problematic nature of sequence alignment, molecular biologists often delete parts of their sequences from the phylogenetic analysis (Olsen and Woese 1993). The rationale for this is that those parts of the sequences that cannot be reliably aligned should be excluded from the estimation of the phylogeny (Olsen 1988), as they are likely to be phylogenetically uninformative (Olsen and Woese 1993). This particularly occurs when alignment regions contain many gaps. Unfortunately, there are few objective criteria for deciding which parts of the alignment are ambiguous (Gatesy, DeSalle, and Wheeler 1993), and traditionally the exclusion of sequence regions has been done by hand (Olsen and Woese 1993).

Empirical Study

In the study presented here, we address the following specific questions regarding the effects of different sequence alignment procedures on molecular phylogenetic analysis. How sensitive are the phylogenetic trees to different sequence alignment strategies? How sensitive are the phylogenetic trees to different sequence alignment parameters? How sensitive are the phylogenetic trees to the exclusion of parts of the sequences? The use of an empirical data set to examine these questions can provide a valuable complement to simulation studies, as the data set exemplifies the real world (Cracraft and Helm-Bychowski 1991; Allard and Miyamoto 1992; Cummings, Otto, and Wakeley 1995; Wheeler 1995; Russo, Takezaki, and Nei 1996).

To address these questions, we analyzed complete 18S ribosomal RNA (rRNA) gene sequences from 43 taxa of the phylum Apicomplexa. The Apicomplexa is a group of diverse parasitic protozoa, characterized by the presence of an apical complex at the anterior end of their invasive life cycle stage (Levine 1988). This grouping is the apparently monophyletic remains of the once much larger phylum Sporozoa, which has had several recently erected phyla excised from it (Cox 1994). The phylogenetic relationships among the Apicomplexa are still the subject of considerable debate (Corliss 1994; Cox 1994), with several competing hypotheses. Thus, it is an appropriate group for testing how sensitive the different phylogenetic conclusions are to variation in alignment procedures. Note that our objective here is not necessarily to produce a definitive estimate of the phylogeny of the Apicomplexa, but to test how sensitive the phylogenetic estimates are to the alignment process (see Cracraft and Helm-Bychowski 1991).

Small-subunit rRNA has been widely used to infer the phylogeny of a broad range of organisms, as it is universal and abundant (Hillis and Dixon 1991; Olsen and Woese 1993). The rRNA molecule has a specific secondary structure necessary for the formation and functioning of ribosomes (Gutell, Larsen, and Woese 1994), and the primary and secondary structures are conserved even among very divergent taxa (Hillis and Dixon 1991). So, the sequences of the rRNA genes (rDNA) are constrained by the secondary structures of their products, and this allows knowledge of the secondary structure to be used for the alignment of the rDNA sequences (Kjer 1995; Hickson et al. 1996). Furthermore, the helix (or stem) and loop regions can be treated as a major division of nucleotide site change (Vawter and Brown 1993; Muse 1995), as the paired bases of the helices must result from compensatory mutations. This leads to an objective criterion for assessing the cladistic informativeness of different regions within the sequences (Wheeler and Honeycutt 1988; Smith 1989; Dixon and Hillis 1993; Ellis and Morrison 1995; Kjer 1995; Hickson et al. 1996). Therefore, the 18S rRNA data are also appropriate for testing how sensitive the different phylogenetic conclusions are to different parts of the sequences.

Our primary objective was to test whether (under realistic circumstances) variability in cladogram estimation as a result of differences in alignment procedure can be as large as variability resulting from differences in tree-building procedure. To do this, it is only necessary to find some combination of circumstances where this is true, rather than to test every theoretical possibility. Our analyses thus sample only part of the extremely large number of possible variations in the alignment procedure. Furthermore, it is important to recognize that the objective of phylogenetic analysis is to produce a phylogenetic tree (cf. Feng and Doolittle 1996), and thus it is necessary to compare the different alignment procedures by testing the robustness of the resulting phylogenetic hypotheses rather than by simply comparing the alignments directly (e.g., by computing average pairwise percent similarity).

Variation in the sequence alignment process can be achieved in a number of ways, and, consequently we tried two different strategies. First, a single alignment algorithm could be chosen and the values of the available parameters could be varied (e.g., Fitch and Smith 1983; Lake 1991; Mindell 1991; Gatesy, DeSalle, and Wheeler 1993). Probably the most important of these parameters are the alignment gap cost ratios or "gap weights" (Tyson 1992; Vingron and Waterman 1994; Wheeler 1995), which refer to the relative cost of inserting a new gap into a sequence or extending an already-existing gap. For each gap these parameters are usually in the form:

$$\text{Gap penalty} = \text{gap opening penalty} + (\text{gap extension penalty} \times \text{gap length})$$

where the penalties are relative to the cost of a substitution. There is no way of determining analytically what these weights should be (Rinsma-Melchert 1993), and the computer programs that implement the alignment algorithms usually have default values for the weights that are designed to produce "biologically interesting" results. To asses the effects of this type of variation, we varied the gap penalties in a systematic manner for the commonly used ClustalW computer program.

Alternatively, a number of different alignment algorithms could be chosen, with only one set of parameter values employed for each of these algorithms. In using this strategy for six different algorithms, our objective was not to provide a rigorous comparison of the different alignment methods (and their associated computer programs), and so no special attempt was made to optimize the performance of any of the mathematical methods. The default values (or those suggested in the instruction manual) were used for all of the parameters, as these are the ones that are most likely to be employed in practice.

For the phylogenetic analyses we employed three commonly used methods of cladistic inference that cover the range of available possibilities: neighbor joining, weighted parsimony, and maximum likelihood (Morrison 1996). This was to determine whether the sensitivity of the alignments depended on the choice of a phylogeny inference method, rather than for the purpose of comparing the effectiveness of these different tree-construction methods (cf. Cummings, Otto, and Wakeley 1995). So, little attempt was made to optimize the performance of these methods, but a similar model of character-state transformations (transversion/transition cost ratio) was used for each analysis (see Wheeler 1995). For the measurement of sensitivity we were only interested in the branching order of the phylogenetic trees, rather than in the branch lengths. Also, we made little attempt to estimate the support for any of the clades within each analysis, as we were interested solely in sensitivity to different alignments rather than robustness within an alignment.

Materials and Methods

Data Set

The data set consisted of the complete 18S rDNA sequences of all Apicomplexa lodged with GenBank (72 sequences covering 43 taxa; table 1). Recent morphological and molecular data suggest that the sister group to the Apicomplexa is likely to be found among the dinoflagellates (phylum Dinozoa) (e.g., Levine 1988; Barta, Jenkins, and Danforth 1991; Gajadhar et al. 1991; Schlegel 1991; Wolters 1991; Sadler et al. 1992; Cavalier-Smith 1993; Goggin and Barker 1993; Rodrigo, Bergquist, and Bergquist 1994; Escalante and Ayala 1995; Siddall, Stokes, and Burreson 1995); so, sequences of three divergent taxa (one symbiotic, two nonsymbiotic) from this phylum were used to root the cladograms (see Smith 1994).

Sequence Alignment Algorithms

Six different multiple-alignment procedures were used: manual alignment according to secondary structure, and computer-alignment using five mathematical procedures that cover the range of available possibilities.

The manual sequence alignment used was that described by Van de Peer et al. (1994), which defines the complete secondary structure of the 18S rRNA molecule. The alignment process is iterative, beginning with the juxtaposition of regions of extensive primary structural similarity, and then refinement by invoking higher-order structural constraints; higher-order structures are inferred by comparative analysis, relying on the search for compensatory base substitutions or positional covariance (Gutell 1996). This alignment is available from The SSU rRNA Database (contactable at http://www-rrna.uia.ac.be/), maintained by Y. Van de Peer, P. De Rijk, and R. De Wachter (Departement Biochemie, Universiteit Antwerpen). The aligned sequence length was 2,704 nucleotides, of which 1,229 positions (45%) were invariant across all taxa (a position with a single nucleotide aligned against a gap in the other sequences is treated as invariant, as are positions with identical nucleotides in all of the sequences).

The first computer alignment program used was PileUp, in the GCG 8.1 package (Genetics Computer Group; Devereux, Haeberli, and Smithers 1985). This method produces the final multiple alignment from a series of progressive pairwise alignments between sequences and clusters of sequences. The pairwise alignments use the method of Needleman and Wunsch (1970), while the clustering order of progressive sequence alignment (Feng and Doolittle 1987) is determined from a UPGMA guide tree. All of the program default values were used. The aligned sequence length was 2,509 nucleotides, of which 953 positions (38%) were invariant across all taxa.

The second computer alignment program used was ClustalW 1.5 (Thompson, Higgins, and Gibson 1994).

Table 1
The 18S rDNA Sequences Used in the Phylogenetic Analyses

Taxonomic Arrangement[a]	GenBank Accession Number(s)	Length(s)
Phylum Apicomplexa		
Class Perkinsidea		
Perkinsus marinus	X75762	1,793
Perkinsus sp.	L07375	1,795
Class Coccidea		
Cryptosporidium baileyi	L19068	1,733
Cryptosporidium muris	L19069, X64342,X64343	1,743–1,748
Cryptosporidium parvum	L16996, L16997, X64340, X64341, L25642	1,740–1,750
Eimeria acervulina	Anderson et al.[b]	1,795
Eimeria brunetti	Anderson et al.[b]	1,791
Eimeria maxima	Anderson et al.[b]	1,796
Eimeria mitis	Anderson et al.[b]	1,796
Eimeria necatrix	Anderson et al.[b]	1,796
Eimeria praecox	Anderson et al.[b]	1,794
Eimeria tenella	Anderson et al.[b]	1,803
Neospora caninum	L24380, U03069	1,789–1,792
Sarcocystis arieticanis	L24382	1,892
Sarcocystis fusiformis	U03071	1,881
Sarcocystis gigantea	L24384	1,900
Sarcocystis muris	M64244	1,809
Sarcocystis neurona	U07812	1,803
Sarcocystis tenella	L24383	1,837
Toxoplasma gondii	M97703, X68523, X75453, X75429, X75430, U00458, U03070, L24381, X65508	1,784–1,795
Class Hematozoea		
Order Haemosporida		
Plasmodium berghei	M14599, M19712	2,059
Plasmodium cynomolgi	L07559, L08241, L08242	2,065–2,167
Plasmodium falciparum	M19172, M19173	2,090–2,237
Plasmodium fragile	M61722	2,082
Plasmodium gallinaceum	M61723	2,120
Plasmodium knowlesi	L07560	2,111
Plasmodium lophurae	X13706, M14821	2,118
Plasmodium malariae	M54897	2,147
Plasmodium mexicanum	L11716	2,200
Plasmodium reichenowi	Z25819	2,093
Plasmodium vivax	X13926, U03079, U03080, U07367, U07368	2,032–2,147
Order Piroplasmida		
Babesia bigemina	X59604, X59605, X59607	1,693
Babesia bovis	L19077, L19078, M87566	1,574–1,653
Babesia caballi	Z15104	1,694
Babesia canis	L19079	1,711
Babesia divergens	U16370, U07885	1,721–1,724
Babesia equi	Z15105	1,748
Babesia rodhaini	M87565	1,745
Cytauxzoon felis	L19080	1,774
Theileria annulata	M64243, M34845	1,744
Theileria buffeli	Z15106	1,740
Theileria parva	L02366, L28999	1,742
Theileria taurotragi	L19082	1,736
Phylum Dinozoa		
Class Dinoflagellata		
Crypthecodinium cohnii	M64245, M34847	1,796
Prorocentrum micans	M14649	1,800
Symbiodinium pilosum	X62650	1,796

[a] Taxonomic arrangement follows Corliss (1994).

[b] J. W. Anderson, A. Elbrecht, M. Dashkevicz, S. D. Feighner, P. R. Chakraborty, P. A. Liberator, H. P-Juchelka, and A. Perkins-Barrow. 1992. Species-specific method for identifying infectivity of *Eimeria* species. European Patent Application 0-516-381-A2.

The algorithm used is very similar to that of PileUp, except that the guide tree is produced by neighbor joining. All of the program default values were used. The aligned sequence length was 2,529 nucleotides, of which 1,236 positions (49%) were invariant across all taxa.

The third computer alignment program used was TreeAlign (Nov. 90) (Hein 1990). The algorithm produces a phylogenetic tree as it aligns the sequences, using a combination of distance matrix and heuristic parsimony methods. Pairwise distances between sequences are used to construct a guide tree (Hein 1989*b*), which is optimized by rearrangements; a parsimony tree is then produced during the alignment (Hein 1989*a*), which is also optimized by rearrangements. All of the program default values were used, with gap weights set to 8+3k (as suggested in the manual). The aligned sequence length was 2,834 nucleotides, of which 1,562 positions (55%) were invariant across all taxa.

The fourth computer alignment program used was MALIGN 2.5 (Wheeler and Gladstein 1994). This program also produces a phylogenetic tree as it aligns the sequences. The algorithm uses a wide range of heuristic procedures to search for the combination of multiple alignment and phylogenetic tree that minimizes the total alignment score along the tree (i.e., there is a specified criterion for global optimality of the solution but no single specified search procedure for finding this solution). The following parameters were set (as suggested in the manual): internal 3, extragaps 1, leading 2, trailing 2, changecost 1, score 3, contig, iter, quick, alignswap, treeswap. The aligned sequence length was 3,549 nucleotides, of which 1,829 positions (52%) were invariant across all taxa.

The fifth computer alignment program used was SAM 1.1a (Hughey and Krogh 1996). The algorithm uses a linear hidden Markov model that estimates the probabilities of nucleotide change, which are then used as penalties in the alignment cost (Krogh et al. 1994). The model is trained (i.e., the probabilities calculated) using an expectation-maximization procedure, and the trained model is then used to create the multiple alignment. All of the program default values were used, the buildmodel module being trained on a representative selection of 22 of the sequences, followed by the addfims module. The aligned sequence length was 2,732 nucleotides, of which 1,288 positions (47%) were invariant across all taxa.

Due to computer memory constraints, it was not possible to use all 75 of the sequences in the alignment process. So, a consensus sequence (based on the secondary structure alignment) was derived from the individual sequences for each of those species with more then one sequence, using the MacClade 3.04 computer program (Maddison and Maddison 1992). The standard IUPAC ambiguity codes were used for those few nucleotide positions with more than one possible character state in a consensus sequence. This resulted in a data set of 46 sequences (one for each taxon), which were used for the alignments.

Sequence Gap Penalties

For the investigation of the relative effect of the alignment gap cost ratios (gap penalties), the multiple-sequence alignment parameters were varied for the ClustalW computer alignment program. The values of the gap opening penalty (GOP) and the gap extension penalty (GEP) were both varied in a logarithmic fashion (Wheeler 1995), testing all orthogonal combinations of these values. The gap opening penalty was varied from 0.5 to 64 times the cost of a substitution ($\log_2$GOP = -1, 0, 1, 2, 3, 4, 5, 6), and the gap extension penalty was varied from 0.031 to 8 times the cost of a substitution ($\log_2$GEP = -5, -4, -3, -2, -1, 0, 1, 2, 3). The PileUp, TreeAlign, and MALIGN programs were not investigated because of technical computer difficulties in running the required number of analyses, while the SAM program uses the hidden Markov model to estimate the gap penalties (which thus cannot be specified a priori). The phylogenetic trees were inferred using maximum likelihood (see below).

Sequence Subsets

For the investigation of subdivision of the sequences, all of the currently recognized helical regions within the 18S rRNA were identified from the alignment according to secondary structure (Van de Peer et al. 1994). Two separate data files were then created, one containing all of the helical (whether double- or single-stranded) positions (49% of the aligned sequence length) and one containing all of the nonhelical (single-stranded) positions (51% of the aligned sequence length), using the DCSE sequence editor (De Rijk and De Wachter 1993). These data sets were then analyzed separately.

Cladistic Analyses

Neighbor-joining analyses were performed using the TREECON 3.0 program (Van de Peer and De Wachter 1993). Distances were calculated using the Kimura two-parameter model, as modified by Jin and Nei to allow the nucleotide substitution rate to vary across sites as a gamma distribution with a = 0.5. Gaps were included in the distance calculations as described by Van de Peer, Neefs, and De Wachter (1990). Weighted-parsimony analyses were performed using the PAUP 3.1.1 program (Swofford 1993). A stepmatrix was used to give transversions twice the weight of transitions, and gaps were treated as missing data. Two heuristic search strategies were used for each analysis, the first with addseq=simple, and the second with addseq=random and 10 replicates. Branch swapping was by tree bisection-reconnection. Maximum-likelihood analyses were performed using the fastDNAml 1.0.6 program (Olsen et al. 1994). A transition : transversion ratio of 2:1 was used, with empirical base frequencies, and one rate class for nucleotide substitutions across sites. The heuristic search strategy used the quickadd option followed by local branch swapping. Adams consensus trees (which include all of the nested relationships that occur in all of the set of trees) and symmetric-difference distances between trees (which indicate how many nonshared

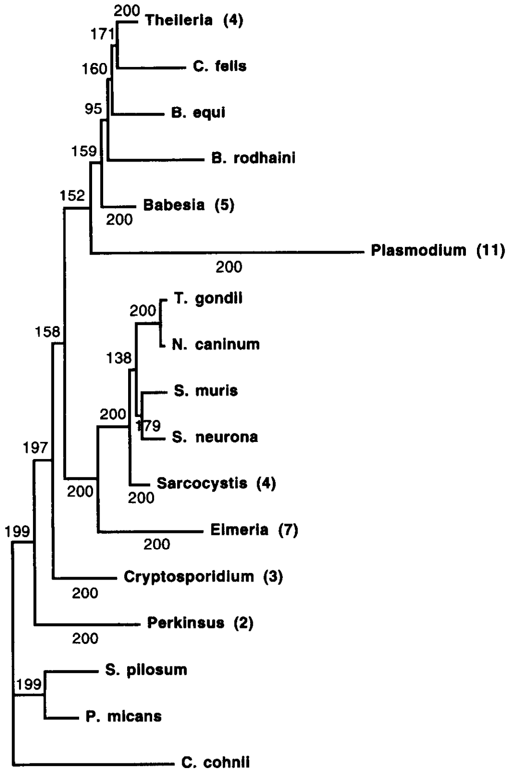

FIG. 1.—Phylogenetic relationships among the Apicomplexa as inferred from the structure alignment of the SSU rRNA and the neighbor-joining tree-building method. Only 17 of the 46 taxa analyzed are shown in the cladogram, the numbers in parentheses indicating how many species of each genus form a monophyletic group on that branch. Species names are as in table 1. The branch lengths are proportional to the amount of inferred evolutionary change; and the numbers on the branches are the number of times that the branch was supported in 200 bootstrap replicates.

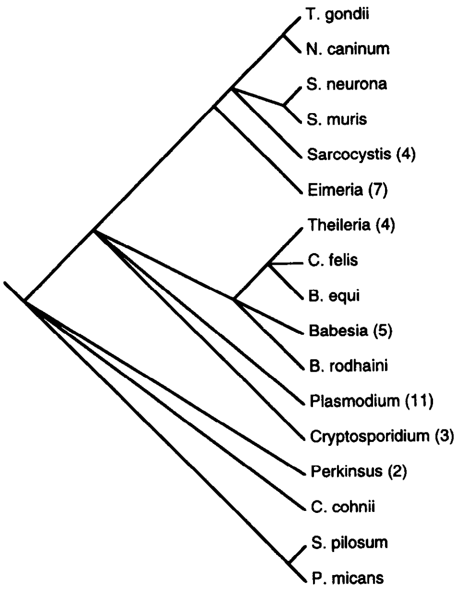

FIG. 2.—Adams consensus tree of the 18 trees inferred from SSU rRNA by different combinations of alignment procedure and cladistic method. Only 17 of the 46 taxa analyzed are shown in the cladogram, the numbers in parentheses indicating how many species of each genus form a monophyletic group on that branch. Species names are as in table 1.

clades there are for two trees) were calculated using the PAUP program.

Results

Sequence Alignment Algorithms

In total, 18 phylogenetic analyses were performed (three cladistic methods for each of six alignment methods). Most of the analyses produced only one phylogenetic tree, but some of the weighted-parsimony analyses produced several equally parsimonious trees: 3 for the structure alignment, 18 for the PileUp alignment, 2 for the Malign alignment, and 3 for the SAM alignment. In each of these cases, a 50% majority-rule consensus tree (which includes all of the clades that occur in >50% of the set of trees) was calculated using the PAUP program. Most of the variations among the multiple trees were due to local rearrangements of taxa (i.e., near the tips of the trees). Consequently, we ignored the minor rearrangements in comparing the results of the analyses, focusing instead on the basal parts of the trees as shown in the example neighbor-joining tree in figure 1.

The 18 final trees from these analyses showed a great deal of similarity, with their Adams consensus tree having considerable structure (fig. 2). This indicates that the underlying phylogenetic signal is present in all of the alignments, and that the phylogeny of the Apicomplexa is thus relatively robust to variation in the sequence alignment process. In particular, the coccidia (*Eimeria* + *Sarcocystis* + *Neospora* + *Toxoplasma*) always formed a monophyletic group, as did the piroplasms (*Babesia* + *Cytoauxzoon* + *Theileria*) (fig. 2). Furthermore, most of those genera represented by sequences from several species always formed monophyletic groups, notably *Cryptosporidium, Plasmodium, Theileria,* and *Eimeria.*

However, there were also prominent differences among the cladograms in the placement of particular species (table 2), indicating that both the sequence alignment process and the cladistic procedure do influence the phylogenetic inference. The 18 trees were compared using the symmetric-difference distance, with the distances ranging from 0 to 14 (the most dissimilar trees being those from the structure alignment with maximum likelihood and the MALIGN alignment with neighbor joining).

The alignments that were most sensitive to the tree-building method were the structure and MALIGN alignments, with average symmetric-difference distances among their three trees of 8.0 and 6.7 respectively, followed by the PileUp (average distance 3.0), TreeAlign (3.0), and SAM (2.7) alignments, and then the ClustalW (0.0) alignment. For example, the three trees from the structure alignment placed *B. rodhaini* in three different positions (table 2), while the ClustalW alignment produced three identical trees (which were also identical to the tree from the PileUp alignment with neighbor joining).

The alignment that produced trees that were on average most different from those of the other alignments

Table 2
Major Taxon Placements that Differ Among the Phylogenetic Trees Produced by the Different Combinations of Sequence Alignment and Cladistic Method

	Structure Alignment			PileUp Alignment			Clustal Alignment			TreeAlign Alignment			Malign Alignment			SAM Alignment		
Taxon Placement	N-j	W-p	M-l	N-j	W-p	M-l	N-j	W-p	M-l	N-j	W-p	M-l	N-j	W-p	M-l	N-j	W-p	M-l
Perkinsus																		
Sister to Apicomplexa	*	*	*			*				*	*	*	*	*	*	*		*
Within dinoflagellates				*	*		*	*	*								*	
Cryptosporidium																		
Sister to rest of Apicomplexa	*	*	*	*	*	*	*	*	*					*	*	*	*	*
Sister to piroplasms + *Plasmodium*										*	*	*		*				
Sister to *Plasmodium*													*					
Babesia rodhaini																		
Sister to *Babesia* + *C. felis* + *Theileria*		*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Sister to *B. equi* + *C. felis* + *Theileria*	*																	
Sister to *Plasmodium*			*															
Babesia equi																		
Sister to *C. felis*				*		*	*	*	*	*			*			*		
Sister to *Theileria*					*						*	*		*	*		*	*
Sister to *C. felis* + *Theileria*	*	*																
Sister to *Babesia*			*															
Cytauxzoon felis																		
Sister to *B. equi*				*		*	*	*	*	*			*			*		
Sister to *Theileria*	*	*																
Sister to *B. equi* + *Theileria*					*									*	*		*	*
Sister to *B. equi* + *Theileria* + *Babesia*			*								*	*						
Sarcocystis neurona and *Sarcocystis muris*																		
Sister to *T. gondii* + *N. caninum*	*			*	*		*	*	*	*	*		*			*	*	*
Monophyletic sister to *Sarcocystis*						*						*						
Paraphyletic sisters to *Sarcocystis*		*	*											*	*			

Note.—N-j: neighbor joining; W-p: weighted parsimony; M-l: maximum likelihood.

(when compared using the same tree-building methods) was the structure alignment, with average symmetric-difference distance to the other trees of 7.5, followed by the TreeAlign and MALIGN alignments (average distances 6.1), then the ClustalW alignment (5.9), and the PileUp and SAM alignments (4.7). For example, all three trees from the structure alignment placed *B. equi* in positions that do not appear in the trees from any of the other alignments (table 2), and for two of these three trees the placements of *B. rodhaini* and *C. felis* were similarly nonconformist (table 2). The two alignments that produced trees that were most similar were the PileUp and ClustalW alignments (average distance 2.0).

The computer alignment that produced trees that were most similar to those of the structure alignment (when compared using the same tree-building methods) was the MALIGN alignment, with average symmetric-difference distance of 6.0, followed by the PileUp (average distance 7.3), SAM (7.3), TreeAlign (8.0), and ClustalW (8.7) alignments. For example, the trees from the MALIGN alignment are the only ones that consistently placed *S. muris* and *S. neurona* in the same position as did the equivalent trees from the structure alignment (table 2).

The tree-building method that was least sensitive to the alignment procedure was the neighbor-joining method, with average symmetric-difference distance among its six trees of 4.1, followed by the weighted-parsimony (average distance 6.7) and maximum-likelihood (6.7) methods. The tree-building method that produced trees that were most similar to the other methods (when compared using the same alignments) was the weighted-parsimony method, with average symmetric-difference distance among trees of 3.5, followed by the maximum-likelihood (average distance 4.3) and neighbor-joining (4.8) methods.

Finally, and most importantly, the sequence alignments were responsible for more of the variation among the 18 trees than were the tree-building methods, as the average symmetric-difference distance among the trees when compared using the same tree-building method was 5.8, as opposed to 3.9 when compared using the same alignment.

Sequence Gap Penalties

In total, 72 alignment and subsequent phylogenetic analyses were performed (nine gap extension penalties for each of eight gap opening penalties). As expected,

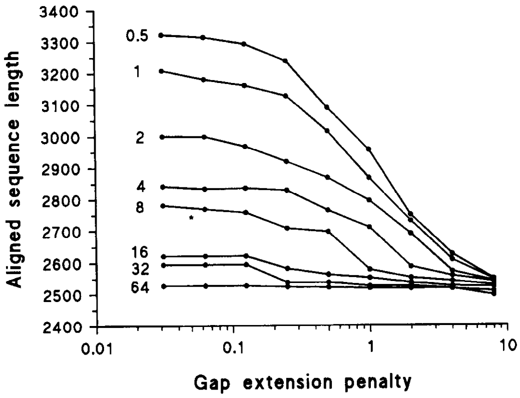

FIG. 3.—Lengths (number of nucleotide positions) of the aligned sequences produced by varying the values of the gap opening and gap extension penalties in the ClustalW alignment program. Each line represents a value for the gap opening penalty (as indicated). The approximate position of the alignment produced by the default penalty values is indicated by the asterisk.

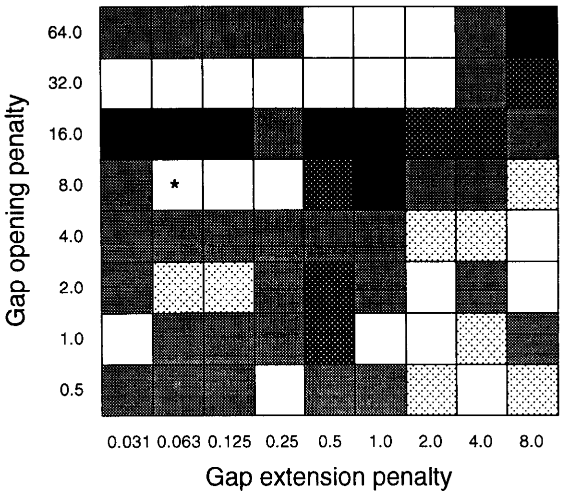

FIG. 4.—Symmetric-difference distances among the 72 maximum-likelihood cladograms produced by varying the values of the gap opening and gap extension penalties in the ClustalW alignment program. Increasing shading represents distances of 0, 2, 4, 6, and 8 from the tree produced by the default penalty values, the approximate position of which is indicated by the asterisk.

increasing the gap opening and extension penalties relative to the cost of a substitution decreased the resulting aligned sequence length, and there was convergence for this data set to an aligned length of ca. 2,500 nucleotide positions (fig. 3). The effect on aligned sequence length of varying the GEP decreased with increasing GOP, and the GEP had little effect below a value of 0.1 for all GOPs (fig. 3).

The maximum-likelihood cladograms from these alignments showed considerable similarity; however, there were prominent differences among the trees in the placement of particular species, indicating that the gap penalties do influence the phylogenetic inference. The symmetric-difference distances among the 72 trees ranged from 0 to 16, with the most divergent tree being that produced from the alignment with a GOP of 64 and a GEP of 0.031 (symmetric-difference distances to the other trees of 8–16).

The symmetric-difference distances of the 72 trees from the tree produced by the default penalty values ranged from 0 to 8 (fig. 4). The penalty values that produced trees that were identical to the default tree form several noninterconnected "islands" that are widely dispersed (fig. 4), and, consequently, there is no apparent means of predicting the similarity of the cladograms from the gap penalty values. Aligned sequence length was also not a good predictor of tree topology, as (for example) the GOP of 64 produced alignments of almost identical length irrespective of GEP (fig. 3) but, in turn, produced trees that differed by symmetric-difference distances of 0–8 (fig. 4).

The symmetric-difference distances of the 72 trees from the tree produced by the structure alignment ranged from 4 to 14, and so none of the gap penalties resulted in a cladogram that was close to that from the secondary structure considerations. The trees that were most similar to that of the structure alignment were produced from the alignments with GOPs of 4 (average distance 7.3), 8 (7.8), and 16 (7.3), while the most different trees were produced from the alignment with a GOP of 0.5 (average distance 12.2).

Sequence Subsets

In total, six phylogenetic analyses were performed (three cladistic methods for each of two sequence subsets). The helical positions produced four equally parsimonious trees for the weighted-parsimony analysis, while the nonhelical positions produced 14 equally parsimonious trees. In each of these cases, a 50% majority-rule consensus tree was calculated. Most of the variations among the multiple trees were due to local rearrangements of taxa, and so the only consensus tree shown in figure 5 that is not fully resolved is that for the nonhelical positions.

The analyses based on the helical positions were more similar to the equivalent full-structure alignment analyses, with average symmetric-difference distance among the trees of 4.7, compared to the analyses based on the nonhelical positions (average distance of 6.7). Indeed, the maximum-likelihood analyses of the structure data and the helical subset were almost identical (symmetric-difference distance 2). For the analyses of helical positions, the weighted-parsimony and maximum-likelihood analyses produced trees that were identical, while the neighbor-joining analysis was very different (both symmetric-difference distances 16). For the analyses of nonhelical positions, the trees were as different from each other as they were for the structure analyses (average distance of 6.7). Furthermore, several taxa were placed in positions that did not occur in any of the other analyses, notably the placement of *B. equi* and *B. rodhaini* as sisters, which occurred in all three trees, and the placement of *Cryptosporidium* within the dinoflagellates, which occurred in the weighted-parsimony tree.

Discussion

Sequence Alignment

All of the multiple-sequence alignment procedures investigated in this study yielded the same basic struc-

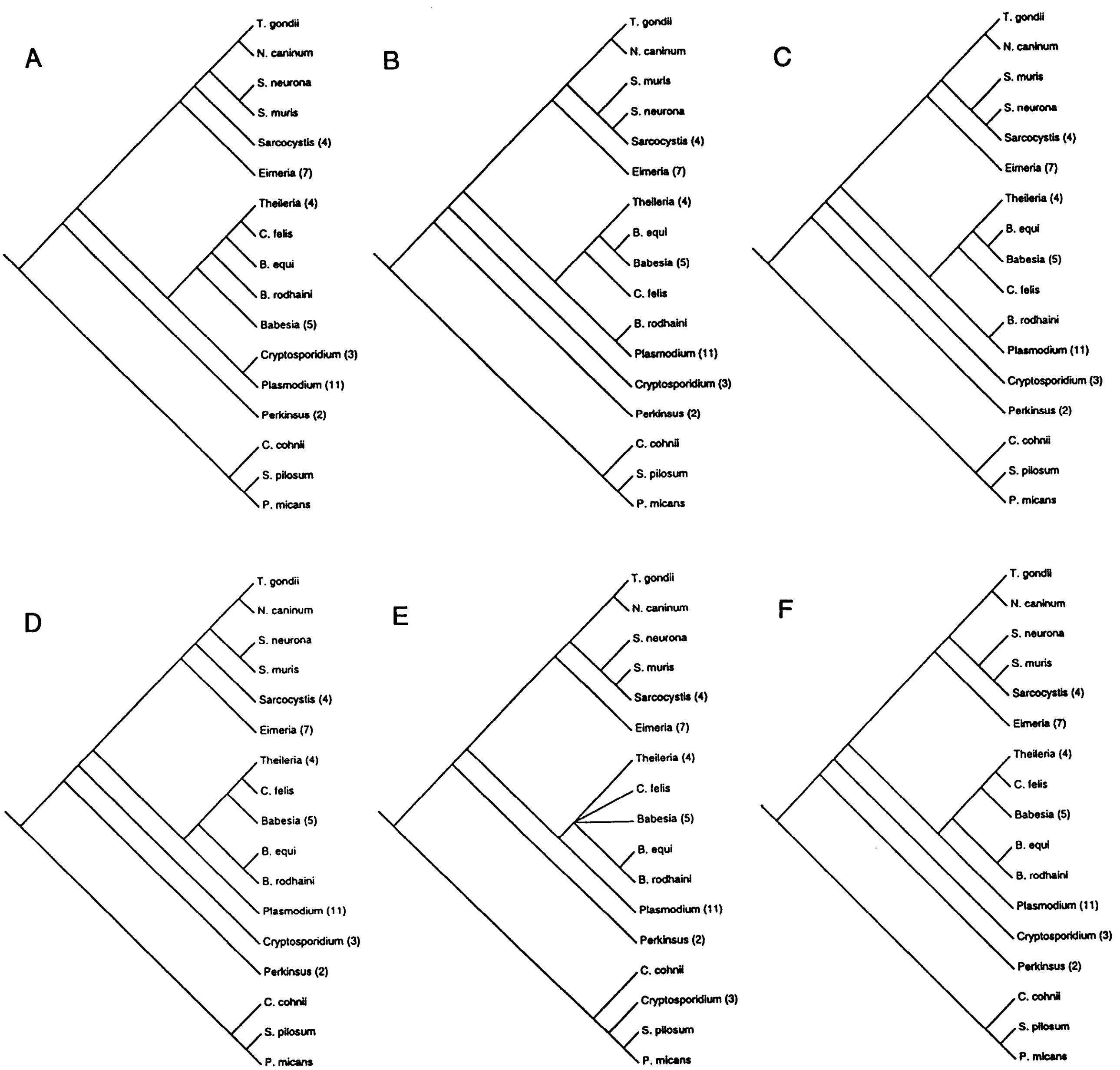

FIG. 5.—Phylogenetic relationships among the Apicomplexa as inferred from different subsets of the SSU rRNA sequence (using the structure alignment) and different cladistic methods. Species names are as in table 1. Only 17 of the 46 taxa analyzed are shown in each cladogram, the numbers in parentheses indicating how many species of each genus form a monophyletic group on that branch. Branch lengths are arbitrary. *A,* Helical positions, neighbor-joining analysis. *B,* Helical positions, weighted-parsimony analysis. *C,* Helical positions, maximum-likelihood analysis. *D,* Nonhelical positions, neighbor-joining analysis. *E,* Nonhelical positions, weighted-parsimony analysis. *F,* Nonhelical positions, maximum-likelihood analysis.

ture for the estimate of the phylogenetic relationship among the members of the Apicomplexa, which presumably represents the underlying phylogenetic signal. This implies that there are phylogenetically informative regions of the SSU rDNA that are relatively robust to different strategies for aligning homologous nucleotides.

Nevertheless, it is clear that the placement of many of the taxa is sensitive to the alignment procedure used; indeed, the variations among the alignments produced trees that were on average more different from each other than did the variations among the tree-building methods used. This is an important conclusion, because although it has long been recognized that differences in both alignment and tree-building methods can influence phylogenetic inference, it is only the latter aspect that has received considerable attention in the literature (Morrison 1996).

The extent to which this conclusion can be generalized to other taxa and genes is unknown. For example, we have only considered rRNA genes, for which an a priori biological model exists; a different situation might exist for protein-coding genes, where a priori models of function may be less clear (Taylor 1987). Nevertheless, it is likely that many other data sets will show the same degree of sensitivity to alignment procedures as does ours; other specific examples where adjusted alignments of SSU rDNA sequences produced different trees are discussed by Eernisse and Kluge (1993) (for vertebrates), Rodrigo, Bergquist, and Bergquist (1994) (for eukaryotes), and Ellis and Morrison (1995) (for apicomplexans). Furthermore, McClure, Vasi, and Fitch (1994) have shown that there is considerable variation in the relative success (at detecting motifs) among several computerized procedures for aligning multiple protein sequences, and protein-coding DNA sequences are often aligned by first translating them into the equivalent protein sequences and then aligning these protein sequences (Hein 1994; Russo, Takezaki, and Nei 1996). Consequently, we suggest that our conclusions from this empirical study probably have a more universal validity.

We thus emphasize that considerable attention needs to be paid to alignment problems in phylogenetic studies, as they are at least as important for phylogenetic studies as are the other well-known problems associated with sequence length (e.g., Cummings, Otto, and Wakeley 1995; Russo, Takezaki, and Nei 1996) and tree-inference methods (e.g., Huelsenbeck 1995; Russo, Takezaki, and Nei 1996). Further comparative studies need to be undertaken to refine our understanding of the ramifications of different alignment procedures. In particular, we need to understand more about the effects of different parameter values on the alignments. Some comparisons have been reported of the relative effects of varying gap costs and substitution costs for proteins (e.g., Henneke 1989; Taylor 1996) and nucleotides (e.g., Wheeler 1995), but the relative effects of varying gap opening and gap extension penalties have previously been reported only for proteins (e.g., Tyson 1992).

We expect that the putative secondary-structure model that we employed is likely to have produced a multiple-sequence alignment that is closer to the true alignment (in the sense of having aligned homologous nucleotides) than is the output from any of the computerized algorithms (Kjer 1995; Hickson et al. 1996; Taylor 1996), since the higher-order structures inferred from comparative analyses are now quite refined, and the method provides a powerful way of identifying functionally important elements in a molecular structure (Olsen and Woese 1993; Gutell, Larson, and Woese 1994; Gutell 1996). This thus represents our preferred set of hypotheses concerning the homology of the SSU rDNA sequences. We emphasize, however, that all multiple alignments are only hypotheses, and are thus open to further testing as new information or other structure models (e.g., Gutell, Larson, and Woese 1994; Kjer 1995; Gutell 1996; Hickson et al. 1996) become available.

The main limitation of the computerized algorithms is that they attempt to maximize sequence similarity rather than sequence homology, the difference being that similarity can be the product of several evolutionary processes, including homology, reversal, convergence, and parallelism. Thus, these algorithms can only succeed to the extent that similarity is the result of homology rather than these alternatives in the particular set of sequences being aligned. Since there is no objective way of assessing this situation, each of the programs implements a different stratagem for producing the multiple alignment. The algorithms employed in the ClustalW and PileUp programs use sequence similarity as the global optimality criterion for the alignments, and so cannot guarantee optimum sequence homology. Furthermore, these programs have to use heuristic procedures, and so they cannot even guarantee achievement of the global optimum of similarity. In a similar manner, the algorithm in the SAM program uses a maximum-likelihood model to maximize sequence similarity, and so cannot guarantee optimum sequence homology. On the other hand, the algorithms in the MALIGN and TreeAlign programs use parsimony of the resulting phylogenetic tree (i.e., the fewer nucleotide changes there are on the tree the better) as their global optimality criterion, thus uniting alignment and tree-building. They thus attempt to minimize homoplasies on the phylogenetic tree, these homoplasies being inferred to be the result of reversals, convergences, and parallelisms; and they thus try to take into account nonhomologous similarities. However, they only use approximate or heuristic methods, and so cannot guarantee achievement of the global optimum (unless an exhaustive search is implemented in MALIGN, which is only practical for extremely small data sets).

It is not straightforward for us to make a direct comparison of the various computerized alignment procedures that we used, because we made no attempt to optimize the performance of most of these methods (e.g., by varying the available parameters). However, there are several general points that can be made about the output from these programs, at least as far as our particular data set is concerned.

First, the ClustalW and PileUp programs produced alignments that are very similar to each other, as would be expected from the similarity of their algorithms, and the trees resulting from these alignments are also quite similar. The TreeAlign and MALIGN programs, although similar to each other in intent, produced alignments and trees that are quite different from each other. In many ways, the alignment and trees from the SAM program represent a compromise among those produced by the other programs.

Second, compared to the structure alignment, the aligned sequence lengths from the PileUp and ClustalW programs are very much shorter (93% and 94%, respectively), while those from the TreeAlign and MALIGN programs are much longer (105% and 131%, respectively). In the latter two cases, this is because there are lengthy parts of the alignment where the nucleotide positions are unique (i.e., there is a contiguous group of nucleotides in one sequence aligned against a gap in all of the other sequences), while in the former two cases these parts have been compressed so that the sequences overlap, along with many of the similarly organized parts of the structure alignment (i.e., some parts of the sequence that the secondary-structure model indicates should be unique also overlap). The SAM program produced an alignment that was very similar in length to that of the structure alignment.

It is these two opposing tendencies (i.e., compression vs. tension of sequence length) that account for many of the differences among the alignments. Their relative strengths are, presumably, a result of the different gap and substitution weights set as the defaults in the programs (i.e., by default some of the programs weight gaps sufficiently heavily that substitutions are preferred over gaps, while others do the reverse). Our analyses show that varying these weights can potentially produce alignments that are at least as different from each other as those produced by the different algorithms. Furthermore, for the ClustalW program there was no combination of gap weights that produced a tree that was the same as that from the structure alignment, in spite of the fact that many of the alignments were similar in length to that of the structure alignment. Clearly, published studies need to be specific about the program parameter values that were used to create the alignments, rather than simply mentioning which program was used, if their alignments are to be repeatable.

This issue is clouded by the fact that in phylogenetic studies manual "improvements" are often carried out on the computer-produced alignment in order to further maximize the apparent similarities among the aligned sequences. For the TreeAlign and MALIGN alignments that we carried out, these post hoc modifications would presumably have resulted in much shorter multiple alignments, overlapping many of the unique sequence regions and thus making the lengths much closer to that of the structure alignment. We did not attempt to do this, however, because there is no objective criterion for carrying out the procedure, it relying instead on the subjective judgment of each individual phylogeneticist. Furthermore, it is unclear whether these improvements actually increase sequence homology, as opposed to merely increasing sequence similarity.

The third general point that can be made is that aligned sequence length is not necessarily a good predictor of how similar the resulting phylogenetic trees will be. In particular, the MALIGN alignment produced trees that were generally more similar to those from the structure alignment than did the alignments from the other programs, in spite of being the one that was most dissimilar in length. Furthermore, the PileUp and ClustalW alignments are very similar to each other, but the resulting trees are quite different when compared to the trees from the structure alignment (and, indeed, the ClustalW program, which appears to be the most popular alignment program in the literature, produced the alignment that was least similar to the structure alignment for our data set). Moreover, we produced many alignments of approximately the same length when varying the gap penalties for ClustalW, but these often resulted in very different trees. It is thus clear that simple similarity of alignment is not the same as similarity of phylogenetic inference. This is an important conclusion, because there appears to be a tacit assumption in the literature that similar alignments should produce similar trees; this assumption is refuted by our analyses. It is for this reason that we chose to compare the different alignment procedures by testing the robustness of the resulting phylogenetic trees (cf. Feng and Doolittle 1996) rather than by simply comparing the alignments directly.

On a more positive note, we emphasize the underlying similarity between the trees produced by the various computerized algorithms and by the structure model. There are thus well-defined regions in the rDNA sequences which are relatively robust to variation in the alignment procedures, which form a backbone for the more variable regions. In the absence of an a priori model, the computerized algorithms can therefore be effectively used as heuristic procedures to produce a good first approximation for the final multiple alignment (cf. Higgins, Thompson, and Gibson 1996). The variable regions can then be dealt with as a separate issue. For example, it is common in the literature (when using computerized algorithms) for those regions where the sequence alignment was problematic to be deleted.

These problematic regions are usually associated with gaps in most of the sequences (gaps in only a few of the sequences are usually not problematic to align), as it is the relative placement of the gaps that creates the problems (Higgins, Thompson, and Gibson 1996). However, few objective methods have been proposed for subdividing nucleotide sequences into regions where alignment is problematic and where it is unambiguous. Rodrigo, Bergquist, and Bergquist (1994) describe a repeatable manual method based on finding invariant positions on each side of the gaps; Fernandes, Nelson, and Beverley (1993) describe a repeatable computerized method based on pairwise percent similarity; and Gatesy, DeSalle, and Wheeler (1993) suggest that regions where the alignment is sensitive to the gap weights in the computerized algorithm are the most ambiguous.

The problem with the deletion of the regions of ambiguous alignment around gaps is that phylogenetically informative characters are being ignored when these gaps represent indels. So, a more useful approach is to determine which regions are phylogenetically informative, rather than which are ambiguously aligned (Olsen and Woese 1993). For rDNA sequences, the a priori model indicates that there could be different phylogenetic information in the helical and nonhelical regions, and our data confirm this, with most of the information being in the helical positions for our data set. A similar conclusion has been reached by Smith (1989) and Ellis and Morrison (1995) for 18S rRNA, and by Dixon and Hillis (1993) for 28S rRNA genes, although Wheeler and Honeycutt (1988) arrived at the opposite conclusion for 5S and 5.8S rRNA. It is thus clear that nucleotide positions in helical and nonhelical regions should be given different weights in phylogenetic analyses of rDNA sequences, although complete exclusion of either region is not recommended (Dixon and Hillis 1993).

It should also be noted that if ambiguous alignment is associated with gaps, then for our data set this is strongly correlated with a subdivision into helical and nonhelical sets, as 64% of the positions in the nonhelical regions have a gap in at least one of the sequences, while this is true of only 19% of the positions in the helical regions. It has often been suggested in the literature that indels preferentially occur in nonconserved regions between helices, both for proteins and for nucleotide sequences. Consequently, computerized alignment algorithms need to take this into account by having position-specific gap penalties rather than having an "average" value that is applied throughout the sequence. Attempts have been made to automate this idea for protein sequences (e.g., Barton and Sternberg 1987; Henneke 1989; Bell, Coggins, and Milner-White 1993; Higgins, Thompson, and Gibson 1996), but it does not yet appear to have been implemented for nucleotide sequences.

Phylogeny of the Apicomplexa

Although our intention has not been to produce a definitive estimate of the phylogeny of the Apicomplexa based on SSU rDNA, there are several general observations that can be made which are sensitive to neither alignment nor tree-building method.

First, the phylum Apicomplexa is monophyletic if *Perkinsus* is excluded. Several of the analyses place *Perkinsus* within the phylum Dinozoa clade, while the structure alignment places it as the sister to either the Apicomplexa or the Dinozoa, depending on where the tree is rooted. The placement of *Perkinsus* can only finally be resolved by including in the analysis more dinoflagellate taxa, as well as including the sister to the Apicomplexa + Dinozoa clade, which is probably the phylum Ciliophora (e.g., Levine 1988; Barta, Jenkins, and Danforth 1991; Gajadhar et al. 1991; Schlegel 1991; Wolters 1991; Sadler et al. 1992; Cavalier-Smith 1993; Goggin and Barker 1993; Rodrigo, Bergquist, and Bergquist 1994; Escalante and Ayala 1995; Siddall, Stokes, and Burreson 1995). The analysis of Escalante and Ayala (1995), based on a ClustalV alignment with post hoc adjustments, suggests that the root of our trees should actually be on the branch between *Perkinsus* and the Apicomplexa. That *Perkinsus* should be placed with the Dinozoa is also suggested by the analyses of Goggin and Barker (1993) and Siddall, Stokes, and Burreson (1995), based on Clustal alignments. Sleigh (1989) summarizes the phenotypic evidence in favor of excluding *Perkinsus* from the Apicomplexa.

Second, the class Coccidia is monophyletic if *Cryptosporidium* is excluded, as also suggested by Barta, Jenkins, and Danforth (1991), based on an alignment by eye. *Eimeria* is the sister to the *Sarcocystis* + *Toxoplasma* + *Neospora* clade, in agreement with the taxonomic schemes of Levine (1985, 1988) and Vivier and Desportes (1990). However, the monophyly of *Sarcocystis* is sensitive to the tree-building method.

Third, *Cryptosporidium* is the sister to the rest of the Apicomplexa in most of our analyses, as suggested by Escalante and Ayala (1995), although some of the alignments do place it as the sister to the class Hematozoea, as suggested by Barta, Jenkins, and Danforth (1991). Either of these placements of *Cryptosporidium* conflicts with both the phenotypically based phylogeny (Barta 1989) and the recent taxonomic schemes (Levine 1985, 1988; Vivier and Desportes 1990; Corliss 1994), and thus deserves further molecular studies.

Fourth, the class Hematozoea is monophyletic. This confirms the taxonomic schemes of Vivier and Desportes (1990), Cavalier-Smith (1993), and Corliss (1994), rather than that of Levine (1985, 1988), although the phylogenetic tree of Levine (1985) does place the piroplasms and haemosporidians as sister taxa. Cox (1991, 1994) treats the coccidia, piroplasms and haemosporidians as equal taxonomic groups. Within the Hematozoea, the order Piroplasmida is monophyletic, but the genus *Babesia* is not monophyletic, as is also suggested by Ellis et al. (1992) and Mackenstedt et al. (1994), based on ClustalV alignments.

Finally, we point out that many of the literature disagreements concerning the phylogeny of the apicomplexans are probably based on differences in sequence alignment strategies rather than on differences in data (or even in tree-building methods). The phylogenetic placement of the taxa in our data set was sensitive to the alignment used; and in many cases the alternative relationships among the taxa that have been suggested in the literature are no more different than are the relationships suggested by trees derived from the various alignments that we used.

Acknowledgments

We thank R. De Wachter for providing the rDNA structure alignments, P. Sharp and N. Saitou for helpful discussions and comments concerning the work, and A. Watts for help compiling the computer programs.

LITERATURE CITED

ALLARD, M. W., and M. M. MIYAMOTO. 1992. Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. Mol. Biol. Evol. **9**:778–786.

BARTA, J. R. 1989. Phylogenetic analysis of the class Sporozoea (phylum Apicomplexa Levine, 1970): evidence for the independent evolution of heteroxenous life cycles. J. Parasitol. **75**:195–206.

BARTA, J. R., M. C. JENKINS, and H. D. DANFORTH. 1991. Evolutionary relationships of avian *Eimeria* species among other apicomplexan protozoa: monophyly of the Apicomplexa is supported. Mol. Biol. Evol. **8**:345–355.

BARTON, G. J., and M. J. E. STERNBERG. 1987. Evaluation and improvements in the automatic alignment of protein sequences. Protein Eng. **1**:89–94.

BELL, L. H., J. R. COGGINS, and E. J. MILNER-WHITE. 1993. Mix'n'Match: an improved multiple sequence alignment procedure for distantly related proteins using secondary structure predictions, designed to be independent of the choice of gap penalty and scoring matrix. Protein Eng. **6**: 683–690.

CAVALIER-SMITH, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**:953–994.

CHAN, S. C., A. K. C. WONG, and D. K. Y. CHIU. 1992. A survey of multiple sequence comparison methods. Bull. Math. Biol. **54**:563–598.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD et al. (39 coauthors). 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. **80**:528–580.

Corliss, J. O. 1994. An interim utilitarian ("user-friendly") hierarchical classification and characterization of the protists. Acta Protozool. **33**:1–51.

Cox, F. E. G. 1991. Systematics of parasitic Protozoa. Pp. 55–80 *in* J. P. Kreier and J. R. Baker, eds. Parasitic Protozoa. Vol. 1, 2nd edition. Academic Press, San Diego, Calif.

———. 1994. The evolutionary expansion of the Sporozoa. Int. J. Parasitol. **24**:1301–1316.

Cracraft, J., and K. Helm-Bychowski. 1991. Parsimony and phylogenetic inference using DNA sequences: some methodological strategies. Pp. 184–220 *in* M. M. Miyamoto and J. Cracraft, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

Cummings, M. P., S. P. Otto, and J. Wakeley. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. **12**:814–822.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics **7**:367–394.

De Rijk, P., and R. De Wachter. 1993. DCSE, an interactive tool for sequence alignment and secondary structure research. Comput. Appl. Biosci. **9**:735–740.

Devereux, J., P. Haeberli, and O. Smithers. 1985. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:216–223.

Dixon, M. T., and D. M. Hillis. 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. **10**:256–267.

Doolittle, R. F., ed. 1990. Molecular evolution: computer analysis of protein and nucleic acid sequences. Methods Enzymol. **183**:303–502.

Eernisse, D. J., and A. G. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Mol. Biol. Evol. **10**:1170–1195.

Ellis, J., C. Hefford, P. R. Baverstock, B. P. Dalrymple, and A. M. Johnson. 1992. Ribosomal DNA sequence comparison of *Babesia* and *Theileria.* Mol. Biochem. Parasitol. **54**:87–96.

Ellis, J., and D. Morrison. 1995. Effects of sequence alignment on the phylogeny of *Sarcocystis* deduced from 18S rDNA sequences. Parasitol. Res. **81**:696–699.

Escalante, A. A., and F. J. Ayala. 1995. Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proc. Natl. Acad. Sci. USA **92**:5793–5797.

Feng, D.-F., and R. F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. **25**:351–360.

———. 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Methods Enzymol. **266**:368–382.

Fernandes, A. P., K. Nelson, and S. M. Beverley. 1993. Evolution of nuclear ribosomal RNAs in kinetoplastid protozoa: perspectives on the age and origins of parasitism. Proc. Natl. Acad. Sci. USA **90**:11608–11612.

Fitch, W., and T. Smith. 1983. Optimal sequence alignments. Proc. Natl. Acad. Sci. USA **80**:1382–1386.

Gajadhar, A. A., W. C. Marquardt, R. Hall, J. Gunderson, E. V. Ariztia-Carmona, and M. L. Sogin. 1991. Ribosomal RNA sequences of *Sarcocystis muris, Theileria annulata* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol. Biochem. Parasitol. **45**:147–154.

Gatesy, J., R. DeSalle, and W. Wheeler. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol. Phylogenet. Evol. **2**:152–157.

Goggin, C. L., and S. C. Barker. 1993. Phylogenetic position of the genus *Perkinsus* (Protista, Apicomplexa) based on small subunit ribosomal RNA. Mol. Biochem. Parasitol. **60**: 65–70.

Gutell, R. R. 1996. Comparative sequence analysis and the structure of 16S and 23S rRNA. Pp. 111–128 *in* R. A. Zimmermann and A. E. Dahlberg, eds. Ribosomal RNA. CRC Press, Boca Raton, Fla.

Gutell, R. R., N. Larsen, and C. R. Woese. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol. Rev. **58**:10–26.

Hein, J. 1989*a*. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol. **6**:649–668.

———. 1989*b*. A tree reconstruction method that is economical in the number of pairwise comparisons used. Mol. Biol. Evol. **6**:669–684.

———. 1990. Unified approach to alignment and phylogenies. Methods Enzymol. **183**:626–645.

———. 1994. An algorithm combining DNA and protein alignment. J. Theor. Biol. **167**:169–174.

Henneke, C. M. 1989. A multiple sequence alignment algorithm for homologous proteins using secondary structure information and optionally keying alignments to functionally important sites. Comput. Appl. Biosci. **5**:141–150.

Hickson, R. E., C. Simon, A. Cooper, G. S. Spicer, J. Sullivan, and D. Penny. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. Mol. Biol. Evol. **13**:150–169.

Higgins, D. G., J. D. Thompson, and T. J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. **266**:383–402.

Hillis, D. M., and M. T. Dixon. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**:411–453.

Hirosawa, M., Y. Totoki, M. Hoshida, and M. Ishikawa. 1995. Comprehensive study of iterative algorithms of multiple sequence alignment. Comput. Appl. Biosci. **11**:13–18.

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

Hughey, R., and A. Krogh. 1996. Hidden Markov models for sequence analysis: extensions and analysis of the basic method. Comput. Appl. Biosci. **12**:95–107.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. **4**:314–330.

Krogh, A., M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. 1994. Hidden Markov models in computational biology: applications to protein modeling. J. Mol. Biol. **235**: 1501–1531.

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

Levine, N. D. 1985. Phylum II. Apicomplexa Levine, 1970. Pp. 322–374 *in* J. J. Lee, S. H. Hutner, and E. C. Bovee, eds. An illustrated guide to the Protozoa. Society of Protozoologists, Lawrence, Kans.

———. 1988. The protozoan phylum Apicomplexa. CRC Press, Boca Raton, Fla.

Mackenstedt, U., K. Luton, P. R. Baverstock, and A. M. Johnson. 1994. Phylogenetic relationships of *Babesia divergens* as determined from comparison of small subunit ribosomal RNA gene sequences. Mol. Biochem. Parasitol. **68**:161–165.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: analysis of phylogeny and character evolution. Sinauer, Sunderland, Mass.

McClure, M. A., T. K. Vasi, and W. M. Fitch. 1994. Comparative analysis of multiple protein-sequence alignment methods. Mol. Biol. Evol. **11**:571–592.

Mindell, D. P. 1991. Aligning DNA sequences: homology and phylogenetic weighting. Pp. 73–89 *in* M. M. Miyamoto and J. Cracraft, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

Miyamoto, M. M., and J. Cracraft. 1991. Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. Pp. 3–17 *in* M. M. Miyamoto and J. Cracraft, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

Morrison, D. A. 1996. Phylogenetic tree-building. Int. J. Parasitol. **26**:589–617.

Muse, S. V. 1995. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. Genetics **139**: 1429–1439.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48**:444–453.

Olsen, G. J. 1988. Phylogenetic analysis using ribosomal RNA. Methods Enzymol. **164**:793–812.

Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

Olsen, G. J., and C. R. Woese. 1993. Ribosomal RNA: a key to phylogeny. FASEB J. **7**:113–123.

Rinsma-Melchert, I. 1993. The expected number of matches in optimal global sequence alignments. N. Z. J. Bot. **31**: 219–230.

Rodrigo, A.G., P. R. Bergquist, and P. L. Bergquist. 1994. Inadequate support for an evolutionary link between the metazoa and the fungi. Syst. Biol. **43**:578–584.

Russo, C. A. M., N. Takezaki, and M. Nei. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. **13**:525–536.

Sadler, L. A., K. L. McNally, N. S. Govind, C. F. Brunk, and R. K. Trench. 1992. The nucleotide sequence of the small subunit ribosomal RNA gene from *Symbiodinium pilosum*, a symbiotic dinoflagellate. Curr. Genet. **21**:409–416.

Schlegel, M. 1991. Protist evolution and phylogeny as discerned from small subunit ribosomal RNA sequence comparisons. Eur. J. Protistol. **27**:207–219.

Schulze-Kremer, S. 1996. Molecular bioinformatics: algorithms and applications. Walter de Gruyter, Berlin.

Siddall, M. E., N. A. Stokes, and E. M. Burreson. 1995. Molecular phylogenetic evidence that the phylum Haplosporidia has an alveolate ancestry. Mol. Biol. Evol. **12**:573–581.

Sleigh, M. A. 1989. Protozoa and other protists. Cambridge University Press, Cambridge.

Smith, A. B. 1989. RNA sequence data in phylogenetic reconstruction: testing the limits of its resolution. Cladistics **5**: 321–344.

———. 1994. Rooting molecular trees: problems and strategies. Biol. J. Linn. Soc. **51**:279–292.

Stevens, P. F. 1984. Homology and phylogeny: morphology and systematics. Syst. Bot. **9**:395–409.

Swofford, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Smithsonian Institution, Washington, D.C.

Taylor, W. R. 1987. Protein structure prediction. Pp. 285–322 *in* M. J. Bishop and C. J. Rawlings, eds. Nucleic acid and protein sequence analysis. IRL Press, Oxford.

———. 1996. Multiple protein sequence alignment: algorithms and gap insertion. Methods Enzymol. **266**:343–367.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. 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**:4673–4680.

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

Tyson, H. 1992. Relationships between amino acid sequences determined through optimum alignments, clustering, and specific distance patterns: application to a group of scorpion toxins. Genome **35**:360–371.

Van de Peer, Y., and R. De Wachter. 1993. TREECON: a software package for the construction and drawing of evolutionary trees. Comput. Appl. Biosci. **9**:177–182.

Van de Peer, Y., J. M. Neefs, and R. De Wachter. 1990. Small ribosomal subunit RNA sequences, evolutionary relationships among different life forms, and mitochondrial origins. J. Mol. Evol. **30**:463–476.

Van de Peer, Y., I. Van den Broeck, P. De Rijk, and R. De Wachter. 1994. Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res. **22**:3488–3494.

Vawter, L., and W. M. Brown. 1993. Rates and patterns of base change in the small subunit ribosomal RNA gene. Genetics **134**:597–608.

Vingron, M., and M. S. Waterman. 1994. Sequence alignment and penalty choice: review of concepts, case studies and implications. J. Mol. Biol. **235**:1–12.

Vivier, E., and I. Desportes. 1990. Phylum Apicomplexa. Pp. 549–573 *in* L. Margulis, J. O. Corliss, M. Melkonian, and D. J. Chapman, eds. Handbook of Protoctista. Jones & Bartlett, Boston, Mass.

Waterman, M. S. 1989. Sequence alignments. Pp. 53–92 *in* M. S. Waterman, ed. Mathematical methods for DNA sequences. CRC Press, Boca Raton, Fla.

Wheeler, W. C. 1995. Sequence alignment, parameter sensitivity, and phylogenetic analysis of molecular data. Syst. Biol. **44**:321–331.

Wheeler, W. C., and D. S. Gladstein. 1994. MALIGN: a multiple sequence alignment program. J. Hered. **85**:417–418.

Wheeler, W. C., and R. L. Honeycutt. 1988. Paired sequence difference in ribosomal RNAs: evolution and phylogenetic implications. Mol. Biol. Evol. **5**:90–96.

Williams, D. M. 1993. A note on molecular homology: multiple patterns from single datasets. Cladistics **9**:233–245.

Wolters, J. 1991. The troublesome parasites—molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-ciliate clade. BioSystems **25**:75–83.

Naruya Saitou, reviewing editor

Accepted December 31, 1996