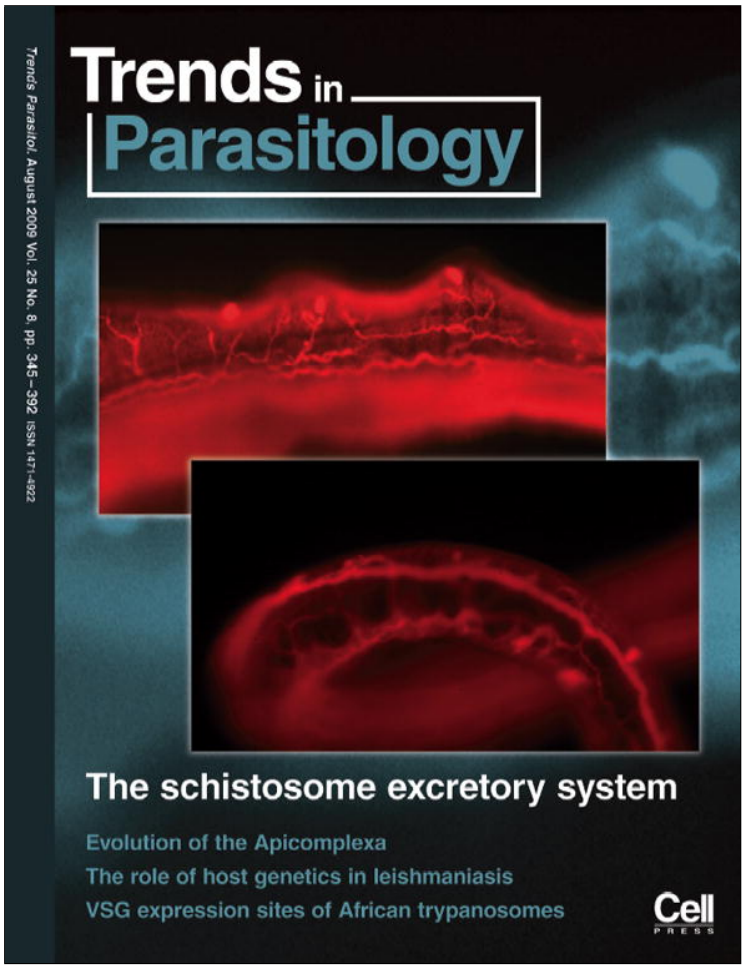

Review

Cell PRESS

Evolutionary Parasitology

Evolution of the Apicomplexa: where are we now?

David A. Morrison

Section for Parasitology, Department of Biomedical Sciences and Veterinary Public Health, Swedish University of Agricultural Sciences, 751 89 Uppsala, Sweden

The Apicomplexa is the only large taxonomic group whose members are entirely parasitic and is, therefore, presumably of major interest to parasitologists. We might, for example, expect that we know a great deal about the biology of the group by now and that we have a clear phylogenetic framework within which to organize that knowledge. It might thus come as a surprise to learn that in terms of biodiversity, the Apicomplexa is actually the least-known group of all. Furthermore, the taxonomic framework for the Apicomplexa is rather tenuous in many respects. This situation is unlikely to change in the short term.

The Apicomplexa

The phylum Apicomplexa (sometimes also known as Sporozoa) forms a large and diverse group of unicellular protists with a wide environmental distribution. They are obligate intracellular parasites, the motile invasive stages of which are characterized by the presence of an evolutionarily unique apical complex (secretory organelles underlying the oral structure). Many of the species are pathogens of humans and domesticated animals, although all animal species are believed to play host to at least one Apicomplexan species (and probably several). The phylum is now recognized as being closely related to the dinoflagellates and ciliates, forming the taxonomic group known as the Alveolata [1–3].

The Apicomplexa is traditionally considered to contain four clearly defined groups [4]: the coccidians, the gregarines, the haemosporidians and the piroplasmids. These groups are based largely on phenotypic characteristics, such as their associated host and/or vector, and which particular tissues they inhabit. These groups were originally designed to be utilitarian rather than to reflect evolutionary history [5,6]; thus, the evolutionary relationships among the four groups and their subsequent taxonomic arrangement are presently unclear.

The current classification of the Apicomplexa is admittedly a conservative one [7,8] and does not take into account modern molecular data. Even the appropriate pages of the Tree of Life Web Project (http://tolweb.org/notes/?note_id=4812) have been rather out of date, until recently. There are very few specialist web resources, notably The Coccidia of the World (http://biology.unm.edu/biology/coccidia/home.html) and iCRYPTO: Taxonomy of the Genus *Cryptosporidium* (http://www.vetsci.usyd.edu. au/staff/JanSlapeta/icrypto/index.htm).

Why do we need to study Apicomplexan evolution?

Probably the two best-known quotations in biology are these: 'On the origin of species by means of natural selection (or the preservation of favoured races in the struggle for life)', which is the title of Charles Darwin's first book [9], and 'Nothing in biology makes sense except in the light of evolution', which is the title of a paper by Theodosius Dobzhansky [10]. What these titles have in common is that they express the over-riding importance of evolutionary thought in modern biology.

Darwin's contribution was to recognize that there are two distinct types of biological evolution: transformational evolution, in which a single object changes through time, and variational evolution, in which groups of variable objects change their relative proportions through time. Transformational evolution is common in the physical sciences as well as in biology, but variational evolution has a special place in the biological sciences because isolated changes in variation will ultimately lead to new species. Both types of evolution can best be represented as a tree-like diagram because this can show the phylesis (changes through time) as relative branch lengths and the cladogenesis (speciation) as the relative branching order.

All organisms have a phylogenetic (evolutionary) history, which can be represented by a phylogenetic tree. Charles Darwin drew the first phylogenetic tree in a notebook in 1836–1837, and the only illustration in his most famous book is a more refined version of that tree, underlining the importance of such trees in modern biological thinking. It is important that all biologists use and understand phylogenetic trees. Sadly, there is considerable evidence that 'tree thinking' (as it is called) is sometimes poor, even in biology (Box 1).

Dobzhansky's comment might be better stated as: 'Very little in biology is not made more comprehensible in the light of phylogenetic history'. For this reason, there are now many biological studies based on phylogenies in all sorts of fields, including such diverse subjects as immunology and epidemiology. In addition, many new areas of study that make extensive use of phylogenies, such as evolutionary development (known as evodevo) and evolutionary ecology, have arisen recently. Elucidating phylogenies, therefore, is now an important component of nearly all of biology. Phylogenies form the framework within

Corresponding author: Morrison, D.A. (david.morrison@bvf.slu.se).

1471-4922/$ – see front matter doi:10.1016/j.pt.2009.05.010

Box 1. Tree thinking

There seems to be a great deal of confusion among non-specialists as to how an evolutionary tree should be interpreted. As far as relationships among organisms are concerned, evolutionary trees are intended to replace the traditional Aristotelian ladder (also known as the Great Chain of Being), which arranges organisms in a linear sequence, ending with human beings as the ultimate biological form. Instead, we now use a branching diagram in which humans are merely one twig among many, thus supplanting the traditional anthropocentric viewpoint. Unfortunately, many people seem to be imagining a pine tree, with a single central axis leading to the 'most derived' species and many side-branches leading to 'lesser' organisms, rather than picturing a continuously branching bush-like structure, as originally depicted by Charles Darwin. This leads to several persistent and apparently endemic problems [50,51].

First, an evolutionary tree must have a time direction (from ancestors to descendants), which is provided by the root. Darwin himself came to realize this distinction: in his 1836–1837 notebook, he drew an unrooted tree, which became a rooted tree in his 1859 book. An unrooted tree thus cannot be a picture of evolutionary history, although it can be an important step towards obtaining such a picture. It is, therefore, inappropriate to identify 'groups' of species (sometimes called clusters) on an unrooted tree [25,52] because only monophyletic groups (called clades) make any sense in an evolutionary context, these being all of the descendants of a common ancestor.

Second, relationships among clades are equal, in the sense that each clade is the sister to some other clade and vice versa. Thus, clades cannot be 'basal' or 'crown' [53] because each single clade branches from some other single clade, rather than being a side-branch from a main stem. There is no main stem in an evolutionary tree but, instead, there is a series of branches leading to a series of twigs, even if some of the branches have more twigs than others. Furthermore, neither of the sisters in a pair represents the ancestor; instead, they share a common ancestor, which might not look like either of them.

Third, characters change through time, so character states can be either ancestral (the original form) or derived (modified in some descendant). However, clades themselves cannot be either ancestral ('primitive', 'lower') or derived ('advanced', 'higher') because each clade will have a combination of ancestral and derived character states. There is no chain leading from ancestral species to derived species. Instead, each species (or group) is the sister to some other species (or group), with which it shares some characters inherited from their ancestors and from which it differs by some unique characters. Any taxonomic group that is interpreted to be ancestral is paraphyletic (i.e. does not contain all of the descendants from the common ancestor) rather than monophyletic and, thus, has no phylogenetic relevance.

which we can best arrange our knowledge of all aspects of the biological sciences.

In terms of biodiversity, some groups of organisms (such as vertebrates, plants and macrofungi) have been much better studied than others. There are also specific initiatives aimed at expanding this list, such as the Assembling the Tree of Life (AToL) program funded by the US National Science Foundation (http://atol.sdsc.edu/). Unfortunately, the Apicomplexa are not well known phylogenetically, and they are not on any of the lists for special future study. They are unicellular endoparasites that are hard to find and which have few easily studied characteristics (mainly life-cycle patterns, cyst organization and ultrastructure). This combination makes them among the most difficult of organisms to work with. The prospects for elucidating their phylogeny might, thus, not be good.

Most large groups of organisms have subsets of taxa that are well known and other subsets that are poorly known. However, unicellular parasites and pathogens

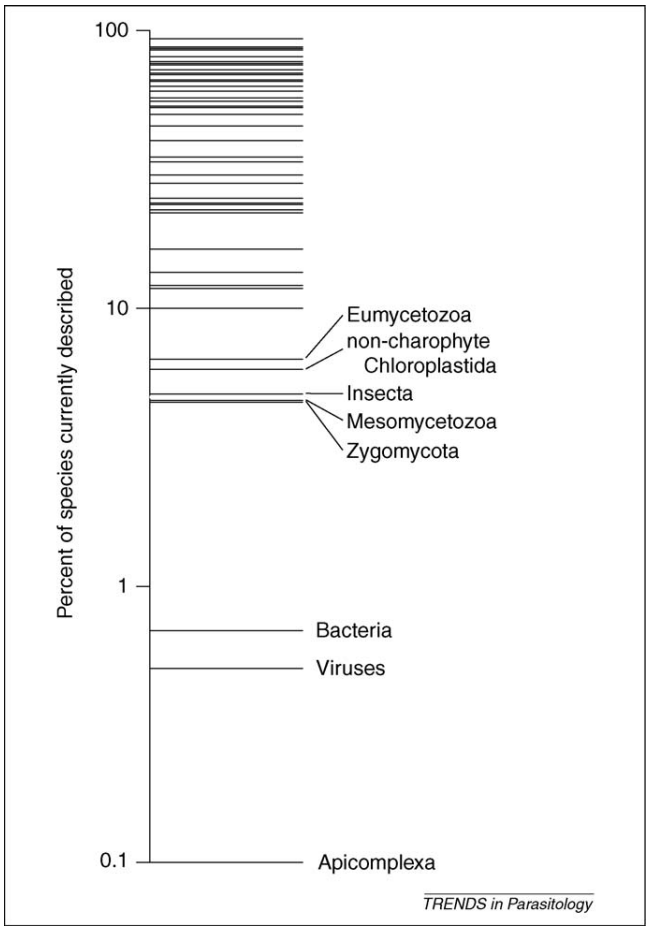

Figure 1. The number of species that are known (i.e. have been described) as a percentage of the estimated total number of living species, for most of the major taxonomic groups. Note that this is a stripe density graph, in which each observation is represented by a horizontal stripe; the size of the stripes is uniform, so the data density is indicated by the apparent clustering of stripes. The eight groups with <10% coverage are labelled. The data are pooled from Ref. [49] (for Prokaryota, Plantae and Metazoa) and from Ref. [11] (for all other groups).

have no well-known subsets, and so they are the worst-known groups in terms of their biodiversity (Figure 1). The Apicomplexa is the largest of these groups and, consequently, they are the worst-known group, with ~6000 named species but perhaps only 0.1% of the total number of species having been named to date (the estimated total number of species is 1.2–10 million) [11]. This compares very unfavourably with all other taxonomic groups; even insects, which are usually listed as the model of a poorly known group, have ~950 000 species known out of an estimated total of 4.5–30 million.

This is unfortunate because the scientific study of biology requires a stable taxonomy based on a robust phylogeny. The groups named in the taxonomy should represent real evolutionary groups (called clades) so that these groups can be used for what is now termed 'comparative biology'. The evolutionary groups identify the pertinent experimental comparisons that are needed for the quantitative study of biodiversity patterns and evolutionary processes.

Phylogeny and taxonomy

In recent years, phylogenetic evidence provided by molecular data has led many biologists to question traditional

taxonomic schemes and propose dramatic changes to the classifications (see below). Widely adopted higher-level taxonomic schemes are neither forced upon scientists as a group nor ratified by formal consensus. Instead, these classifications are usually sanctioned because a majority of the people who have considered the evidence in their favour have found it convincing [12]. It is, therefore, the acquisition of convincing evidence that is the key to a stable taxonomic scheme.

If taxonomy is to reflect phylogeny (as it should), then it is phylogenetic evidence that is the crucial evidence. That is, the key to a robust and well-accepted classification is a robust and well-accepted phylogeny, in which the evolutionary relationships are not subject to change as new evidence accumulates, and which, therefore, can be accepted by non-experts. Too often, traditional taxonomic schemes have been utilitarian rather than phylogenetic, which often means that they are very anthropocentric. A well-known example in the Apicomplexa is the recognition of the primate-host species of *Eimeria* (coccidians) as the separate genus *Cyclospora* [13]. This makes the genus *Eimeria* paraphyletic (i.e. not a clade), which destroys its phylogenetic relevance. This sort of issue can lead to serious conflicts between the scientific requirements of classification and the more pragmatic requirements of end users [14,15].

Experience with recent successful classificatory changes in the best-known taxonomic groups shows that there are five possible impediments to producing a robust and well-accepted phylogeny [6,16], involving the possible failure of taxon sampling, character sampling, phylogenetic analyses, character interpretation and directed data collection. Unfortunately, at least four of these possibilities indicate that the prospects for elucidating the phylogeny and taxonomy of the Apicomplexa are not good, even in the long term. I will take each of these in turn.

Taxon sampling

The sample of taxa used to construct the phylogenetic tree needs to be adequate to provide a convincing case for particular phylogenetic relationships and, thus, taxonomic boundaries. Showing that a problem exists is easy with a small sample size, but revealing the solution takes much more effort. Unfortunately, taxon sampling for the Apicomplexa is woeful for most molecular sequences, and phylogenetic trees based on molecular data have been produced for only a small subset of the known species.

There are currently more than 300 recognized genera of Apicomplexa, in at least 60 families [7], so we need to choose exemplar taxa for phylogenetic analysis (e.g. at least one species from each genus). Unfortunately, sampling to date has been almost entirely opportunistic [17], as it almost always is in parasitology. Opportunities for sampling arise principally from studies of medical diseases (e.g. malaria, cryptosporidiosis and toxoplasmosis) and of veterinary diseases (e.g. coccidiosis, neosporosis and babesiosis). Phylogenetic relevance has not been the criterion for choice, which leads to small and biased samples.

A biased sample usually leads to biased estimates of phylogenetic relationship. For example, the early analyses based on the gene for the small-subunit ribosomal RNA, or rRNA (18S rRNA), of the Apicomplexa indicated clearly that the genus *Cryptosporidium* does not belong with the other coccidians (where it had traditionally been placed), but the authors often then concluded that it was the sister group to the rest of the Apicomplexa. However, these analyses did not include any samples from the gregarine group (i.e. they sampled only coccidians, haemopsoridians and piroplasms), and most of the recent analyses indicate that *Cryptospordium* is actually the sister to the gregarines [18,19]. Relationships cannot be detected if the related groups have not been sampled. Unfortunately, this trend continues to this day; many recent phylogenetic analyses of 18S rRNA do not include haemosporidians, even when the other three groups are included.

The DDBJ–EMBL–GenBank database currently contains nucleotide sequence data for only a few subsets of the Apicomplexa [16], with a major bias towards just five genera: *Babesia*, *Cryptosporidium*, *Plasmodium*, *Theileria* and *Toxoplasma*, which account for 98% of the nucleotide sequences. In addition, these genera, along with *Eimeria*, *Neospora* and *Sarcocystis*, account for 79% of the expressed sequence tag (EST) sequences. Although there are now samples of the 18S rRNA gene from 35 genera (see the phylogenetic analysis below), a lot more widespread sampling needs to occur before any reliable phylogenies are likely to emerge.

Character sampling

If taxonomy is to reflect phylogeny, we need to reconstruct the phylogenetic tree of the species involved. A tree from a single molecular sequence represents only the phylogeny of that one gene, which is not necessarily the phylogeny of the species. A species phylogeny needs to be based on either phenotypic characters or multiple genotypic characters, or even on a combination of the two. Unfortunately, phenotypic characters are rather limited in the Apicomplexa, and the use of multiple genotypic characters to date has been very restricted.

The phenotype characters traditionally used for the Apicomplexa mainly involve life-cycle patterns and ultrastructure [7]. It might be rather difficult to determine homologies among such characters, so that related character states are being compared, and the data are also regrettably incomplete for most species. Apicomplexan phylogenies based solely on phenotypic characters have been rare [20,21], and they have not been particularly robust. Interestingly, there have been no phylogenetic analyses based on combined phenotypic and genotypic data, although these are becoming increasingly common for other taxonomic groups (notably insects).

In the Apicomplexa, genotype characters have been restricted mainly to nucleotide sequences. Here, only concordance between the phylogenies derived from several molecular sequences will be accepted as evidence for the species phylogeny. However, Apicomplexan phylogenies have usually been based solely on the nucleotide sequence of the 18S rRNA gene [16]. Many of the other genes sequenced for the Apicomplexa are those for host recognition or for dealing with the host immune system (sequenced as part of projects producing new drugs or vaccines), which are often unique to each taxonomic group

Box 2. Multi-gene phylogenies

To assess the potential phylogenetic usefulness of the sequence data currently in the DDBJ/EMBL/GenBank database, we can examine the analysis provided by the PhyLoTA Browser (release 1.01, which is based on GenBank release 159) [54]. This shows that the Apicomplexa has 25 539 sequences, relating to 1175 organism 'names'. However, these sequences are associated with only ~300 named species. There are 56 clusters of sequences that could be used to produce a phylogenetic tree for some subset of these species. However, no cluster contains more than 75% of these species, and ~55% of the species occur in only one cluster, as shown in the frequency histogram (Figure I). Not unexpectedly, the most populous cluster is for the (nuclear) 18S rRNA gene, with the (mitochondrial) cytochrome *b* gene next (25% of the species); no other cluster has more than 10% of the species. In addition, although the 18S rRNA gene has been sampled from all of the four main groups, the data for the cytochrome *b* gene are restricted almost entirely to the haemosporidians.

A multi-gene phylogeny, therefore, is unlikely to be produced from these current data. Supertree methods [25] provide one possible approach to the phylogenetic analysis of disparate data such as these, but the taxonomic sampling is currently far too fragmentary for a worthwhile analysis of the Apicomplexa to emerge.

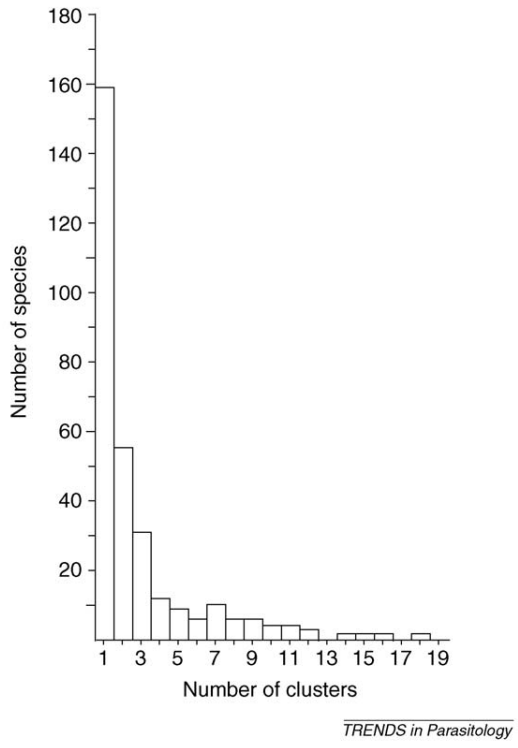

Figure I. Frequency histogram of the cluster sizes formed by the Apicomplexan sequence data in GenBank. Each cluster refers to a single gene. So, for example, there are 159 species that occur in only one cluster (i.e. they have only one gene sequence in the database), and there is one species that occurs in 18 clusters (i.e. has 18 gene sequences).

[22,23] or are subject to heavy selection and are, thus, not necessarily useful for phylogeny.

It seems unlikely that a multigene phylogeny will be produced from the currently available data (Box 2). Nevertheless, multigene phylogenies are starting to appear for individual genera (notably, *Cryptosporidium* [24,25] and *Plasmodium* [26–28], and also for a few other groups [29]). There is now also a large amount of EST data available [30], but the taxon sampling is still too limited for meaningful phylogenetic analysis (*cf*. Ref. [31]).

Box 3. Genome phylogenies

Sequences of complete nuclear genomes have contributed much to comparative genomics, which assumes that the phylogeny is known and can be used as the basis for determining pertinent comparisons among species. However, these genomes might never prove to be useful for phylogeny reconstruction itself. The only situation in which they are likely to be useful is where the original gene samples were biased, because the genomes will then correct the sampling error. However, if the genes previously examined were a representative sample of the genome, then the complete genomes will only confirm what was already known in terms of both confident and problematic relationships. This has already been illustrated for the limited number of available Apicomplexan genomes [55]. For multigene phylogenies, taxon sampling is, thus, likely to be more important than character sampling.

Within the Apicomplexa, organellar genomes are also unlikely to be phylogenetically useful. For example, their mitochondrial genome is the smallest one known [56,57], and the apicoplast (the plastid-like organelle) does not occur universally in all of the taxonomic groups [58].

We, therefore, need to be realistic about what we can expect from the phylogenetic analysis of sequence data, especially genomic sequences. Of particular importance will be our ability to locate representative genes that are appropriate to the evolutionary timescale being examined, rather than merely the quantity of the data *per se*.

An even greater concern is the quality of the data. There are four main potential problems with uncurated public access sequence databases: misidentified and mislabelled taxa, wrongly annotated genes (location, orientation and boundaries), poor sequence quality, and inconsistencies between the database data and the published data. Anyone experienced with the use of these databases will have repeatedly encountered all four of these problems in practice. The unknown (and perhaps unknowable) quality of the data raises the serious question of whether these databases are of any use for high-quality phylogenetic analyses. Indeed, it is obvious that an increasing number of authors are becoming very selective about which, if any, database sequences will be included in their phylogenies.

This does not mean that we must despair, of course. Misannotation, for example, often can be dealt with by careful comparative analysis. Poor sequencing can be corrected if the DNA is still available, although there are no generally accessible depositories for preserving DNA. Taxon misidentification and mislabelling can be straightforward to deal with by having voucher specimens, cultures or microscope slides for every sequence. Sadly, protozoology has a poor history of depositing vouchers in any of the many available depositories, which has been obvious for at least half a century [59], and this has continued into the age of molecular biology. It is, therefore, important to emphasize that without an extant voucher of the sequence source, the identification can never be verified, and the sequence might as well be labelled as 'unidentified'.

Very few nuclear genomes have yet been sequenced within the Apicomplexa. ApiDB, the Apicomplexan bioinformatics resource centre [32], currently lists 14 genomes, but not all of these can be considered complete. Indeed, the US National Center for Biotechnology Information lists only *Plasmodium falciparum* as complete, with 23 other assemblies (11 of which are for *P. falciparum* strains) and eight genomes in progress. These 32 genomes refer to only 17 species, and so genomes have not yet made any important contribution to Apicomplexan phylogenetic analyses (Box 3).

In this regard, it is important to consider the quality of the data contained in the public access databases, particularly the DDBJ/EMBL/GenBank database. Most of these sequence databases are not curated, so the sequences submitted are entirely the responsibility of the submitter.

Box 4. Phylogenetic analysis

Phylogenetic analysis of molecular sequences consists of three distinct procedures: sequence alignment, character coding and tree building. All three of these need to be fully described for a phylogenetic analysis to be repeatable. It is possible to perform all three procedures simply by choosing some popular computer program and then using the default parameter values of that program. For example, one strategy popular in the literature is to choose Clustal for alignment, to ignore coding, and then to choose PAUP* for tree building. Unfortunately, this is a very naïve approach because it does not consider the possible unsuitability of the analyses for the specific dataset at hand, which might lead to results that are artefacts [60].

Alignment is the process of establishing the possible homology relationships among the sequence residues [61,62]. Sequence similarity is often used as strong evidence for potential homology, and this is the basis of all automated alignment procedures. Unfortunately, sequence similarity decreases rapidly as taxa become more distant, so that processes causing sequence length variation (such as duplication, translocation, deletion, insertion and inversion) become more probable. Under these circumstances, similarity cannot be treated as homology. This exacerbates the problems of poor taxon sampling, particularly for rRNA genes. It also exacerbates the problems caused by distant outgroups, which can be very difficult to align with the ingroup.

Character coding [63,64] is often overlooked in sequence analyses. Those parts of the sequence alignment involving length variation (where there are 'gaps') are sometimes considered to be uncertainly aligned, and most computer programs treat gaps as missing data. Furthermore, some regions in sequence alignments are considered to be ambiguously aligned across the dataset, even if subsets of sequences have been aligned with certainty, and these regions are often excluded from the tree-building analysis. In both cases, phylogenetic information is lost. This issue can be dealt with by coding the length-variable regions as a set of independent characters, which are then included in the tree-building analysis.

Tree building simply displays the information obtained from the sequence alignment and coding steps as a branching diagram [25]. That is, conceptually all it should do is change the tabular data into a picture of the data, all of the hard work having been done in the previous two steps. Unfortunately, different tree-building methods often add to the diagram 'extra' artefactual information that does not reflect evolutionary history. For example, compositional biases are a recurring problem (e.g. A+T bias or codon bias), as are juxtaposed long and short branches (resulting in what is known as 'long-branch attraction'). These exacerbate the problems caused by poor taxon sampling and distant outgroups. These issues often can be dealt with by deleting length-variable regions and autapomorphies from the alignment or by choosing appropriate evolutionary models for the analysis.

It is becoming increasingly obvious that there are serious limitations to the use of these unchecked data for phylogenetic purposes (Box 3).

Phylogentic analyses

Phylogenetic analysis of molecular sequences usually consists of three distinct procedures: sequence alignment, character coding and tree building. Sadly, there are many known artefacts potentially associated with each of these procedures, and they need to be considered seriously in all analyses (Box 4). These can be illustrated using a phylogenetic analysis of the Apicomplexa based on the 892 full-length sequences for the 18S rRNA gene currently in the DDBJ/EMBL/GenBank database (Figure 2).

Such an analysis of the Apicomplexa has not been attempted for many years because the computational task can increase exponentially with the addition of each new sequence. Thus, all of the recent phylogenetic analyses have involved only subsets of the species, whereas the analysis presented here simultaneously includes all of the groups. The tree thus represents a summary of the current state of the phylogeny of the Apicomplexa, although the limited taxon sampling and use of only a single gene makes it somewhat tentative. Interestingly, most of the relationships shown in the tree reflect those discussed in the recent literature [13,33–38]. Three of the four traditionally recognized groups form clades, although the adeleids are not closely related in the tree to the rest of the coccidians.

For the sequence-alignment step, reduced sequence similarity made it very difficult to align the gregarine sequences against each other and, also, to align them against the rest of the Apicomplexa. Here, information from the rRNA secondary structure was necessary to reliably establish homologies. This was particularly true for the gregarine sequences, which do not differ from the other groups in length but differ very much in sequence composition.

For the character-coding step, there are seven variable-length stems (so-called regions of expansion and contraction) in the 18S rRNA molecule that show clear evidence of relationships among subsets of the sequences but were not alignable between subsets across the Apicomplexa as a whole. These regions, therefore, were ignored in the analysis. This was particularly true for the haemosporidian sequences (*Plasmodium* and *Hepatocystis*), which differ greatly from the other groups in sequence length in three of these stems.

For the tree-building step, it is likely that the association of *Aggregata*, *Plasmodium* and *Rhytidocystis* with the gregarines (Figure 2), making the gregarines non-monophyletic, is the result both of their long branches and of their strong nucleotide (A+T) bias. Nucleotide bias has long been a serious problem for studies of the Apicomplexa, where the variation of 30–90% A+T among species clearly violates the assumptions of almost all current mathematical models of evolution, which assume homogeneous composition (called 'stationarity') [25]. Possible artefacts associated with long-branch attraction have also been long recognized in the Apicomplexa, and the analysis here certainly places all of the longest branches together.

The choice of evolutionary model thus seems to be crucial for phylogenetic analysis of the Apicomplexa. Furthermore, the position in the phylogenetic tree of some sequences varies depending on the tree-building method used. For example, the maximum-likelihood analysis placed *Rhytidocystis polygordiae* and *Selenidium vivax* with the outgroup species, which is based on doubtful evidence in the alignment. However, none of the analyses placed the haemosporidians with the piroplasms, which is their traditionally expected location. Either this location is wrong or there are no current models that can handle all of the Apicomplexa in a single analysis.

Few phylogenetic analyses of Apicomplexans have explicitly tried to address these analysis issues [16] (often preferring instead to delete 'problematic' taxa), but it

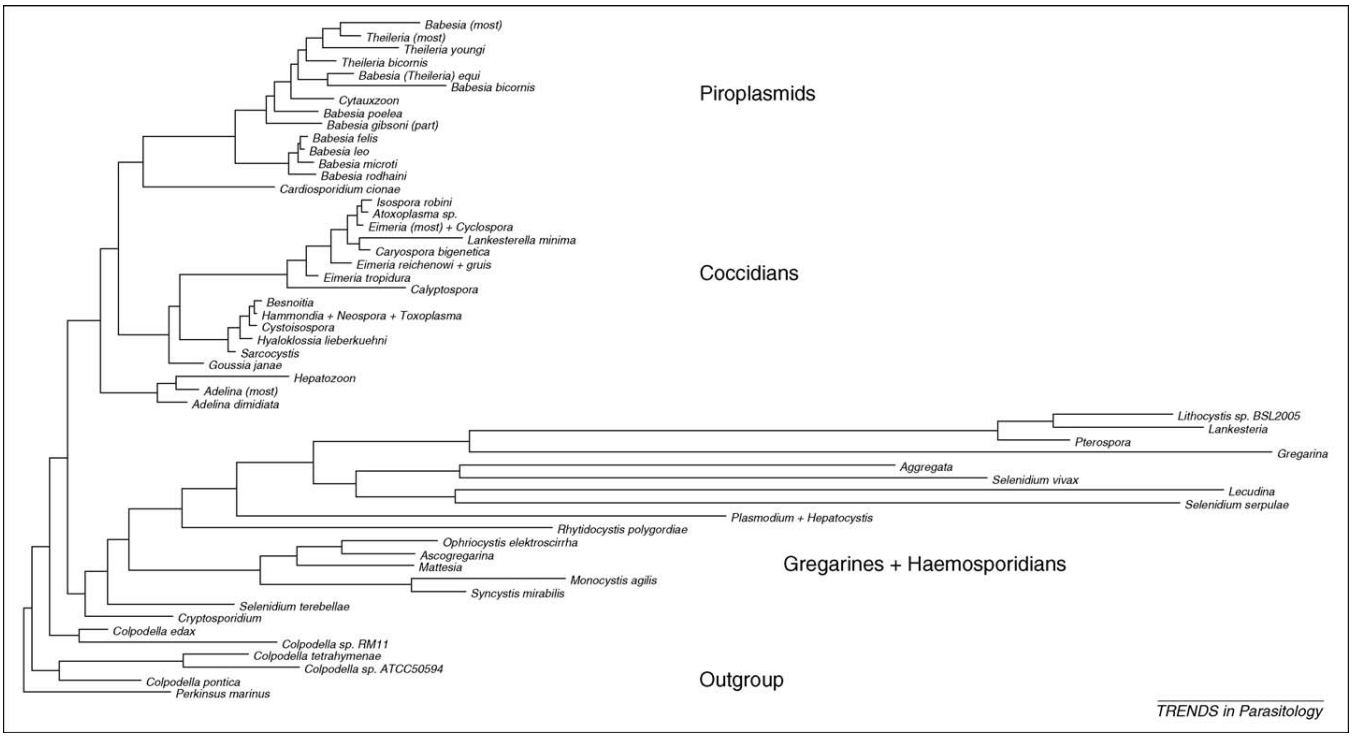

Figure 2. Phylogenetic tree of the Apicomplexa based on full-length sequences of the small subunit rRNA (18S rRNA) gene. The branch lengths are proportional to the inferred amount of evolutionary change (phylesis). *Perkinsus* + *Colpodella* form the outgroup to root the Apicomplexan part of the tree, and the four traditionally recognized Apicomplexan groups are labelled. The tree has been edited to the level of genus, based on the complete tree of those 900 sequences that are currently available in the DDBJ/EMBL/GenBank database and have been identified at least to genus. The sequences were aligned manually based on the secondary structure of the 18S rRNA molecule. The bayesian tree was constructed based on a two-partition analysis, with separate models for the unpaired and paired (stem) alignment positions. The analysis, and its relationship to the three steps of phylogenetic analysis, is described in more detail in the Supplementary Data.

should not be difficult to do so. For example, several of the recent phylogenies of *Plasmodium* seem to have been based on thoughtful analyses [39–42], but this is rarely the case for *Cryptosporidium*. For *Cryptosporidium*, most of the phylogenies involve a 'black-box' analysis pipeline that might be contributing little to our understanding of the genus [25].

Character interpretation

New phylogenies often contradict previous ones, and it needs to be possible to reinterpret the previous homologies in the light of the contradictory new evidence so they are consistent with the new phylogeny [43]. These days, this usually involves reinterpreting (traditional) phenotypic characters to resolve their conflict with (modern) genotypic characters. This might involve the realization that some of the characters are functionally or developmentally correlated rather than being independently informative phylogenetically (e.g. various features of the apical complex). Alternatively, the phenotypic data might have been oversimplified, by combining several independent characters into one complex character (e.g. sporocyst size in *Sarcocystis* [44]), or the dataset might have been incomplete or inaccurate (e.g. the role of stieda bodies in distinguishing *Cystoisospora* from *Isospora* [45]).

Failure of any such reinterpretation, or disagreement about the reinterpretation, greatly weakens the evidence provided by gene sequences. It is, therefore, fortunate that reinterpretation has often proved to be successful when it has been attempted in the Apicomplexa [16,44–46]. Perhaps the best-known example involves *Cryptosporidium*, which has traditionally been placed with the coccidians [7]. Recent phylogenies contradict this placement (Figure 2), which helps explain why anti-coccidial agents are ineffective on members of this genus.

Such reinterpretation involves 20–20 hindsight, of course. What is perhaps more important is to be able to turn these reinterpretations into foresight, so that we can predict which pieces of our current 'knowledge' are likely to also be incorrect. This is what makes a phylogenetic framework so useful for arranging our knowledge of biodiversity. It has been shown repeatedly to be the best predictive framework because organisms share features by inheriting them from their common ancestor. It is these shared features that are involved in the predictions. Well-known examples include the successful prediction of previously unknown hosts of Apicomplexans [13,47]. Phylogenetic trees can also be used to identify sequence samples of unknown origin.

Directed data collection

There needs to be a widespread base of people actively collecting a purposive sample of phylogenetically relevant multigene data. Without this base, both the taxon and the character sampling will be inadequate, in the sense that data will not be available for the crucial exemplar taxa. This leads to uncertainties about species relationships and taxon boundaries, and concordance of multiple gene sequences cannot be demonstrated.

The current collection of pertinent data for the Apicomplexa can best be described as haphazard, which is unlikely to be of much practical value for phylogenetic analysis. In

most cases, with the notable exception of the 18S rRNA gene, the data have not been collected with a phylogenetic analysis in mind. This means that the crucial taxa have not necessarily been the focus of data collection, which they must be if phylogenetic and taxonomic problems are to be resolved.

Informal collaborative groups of researchers have dealt successfully with this issue for the flowering plants, notably the Angiosperm Phylogeny Group (http://www.mobot.org/MOBOT/Research/APweb/) and the Grass Phylogeny Working Group (http://www.umsl.edu/services/kellogg/gpwg/), both of which have produced robust phylogenies and reclassifications based on multiple gene sequences from a large set of exemplar taxa. More recently, formal collaborative AToL-funded groups have had a considerable impact, notably for Assembling the Fungal Tree of Life (http://aftol.org/) and the Angiosperm Tree of Life (http://www.flmnh.ufl.edu/angiospermATOL/).

I know of only one explicit suggestion that this type of activity needs to occur within the Apicomplexa, for the coccidians [6]. Unfortunately, trying to get funding for the study of unicellular endoparasites of non-medical and non-veterinary importance, even if they are of phylogenetic importance for the study of biodiversity, might not be too easy.

Concluding remarks and future perspectives

Every ten years or so, there is a brief survey published of the current state of play regarding the systematics of the Apicomplexa [5,48], and so we are due for another one now. It is regrettable that, in many respects, there is little difference between these various commentaries and that we have not made as much progress as the parasitological importance of the group might call for.

Convincing evidence for a well-resolved phylogeny of the Apicomplexa and, thus, a stable classification, has yet to appear. The Apicomplexa is the most poorly known large taxonomic group, in terms of biodiversity, and taxon and character sampling seem to be the most serious impediments to elucidating its phylogeny. This situation is unlikely to change in the short term without directed data collection, and this situation is unlikely to change in the long term without the realization that Apicomplexans constitute one of the largest components of world biodiversity.

It is a moot point whether lagging behind everyone else is a good thing. On one hand, being late can mean that we benefit from the insights gained by others, and perhaps avoid some pitfalls; on the other hand, less progress simply means that we have so much further to go.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.pt.2009.05.010.

References

1 Yoon, H.S. *et al.* (2008) Broadly sampled multigene trees of eukaryotes. *BMC Evol. Biol.* 8, 14

2 Moore, R.B. *et al.* (2008) A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451, 959–963

3 Gould, S.B. *et al.* (2008) Alveolins, a new family of cortical proteins that define the protist infrakingdom Alveolata. *Mol. Biol. Evol.* 25, 1219–1230

4 Adl, S.M. *et al.* (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.* 52, 399–451

5 Ellis, J.T. *et al.* (1998) The phylum Apicomplexa: an update on the molecular phylogeny. In *Evolutionary Relationships Among Protozoa* (Coombs, G.H. *et al.*, eds), pp. 255–274, Kluwer

6 Tenter, A.M. *et al.* (2002) The conceptual basis for a new classification of the coccidia. *Int. J. Parasitol.* 32, 595–616

7 Perkins, F.O. *et al.* (2000) Phylum Apicomplexa Levine, 1970. In *An Illustrated Guide to the Protozoa* (Vol.1) (Lee, J.J. *et al.*, eds), In pp. 190–369, Society of Protozoologists

8 Kaya, G. (2001) An overview of classification of the phylum Apicomplexa. *Kafkas Üniv. Vet. Fak. Derg.* 7, 223–228

9 Darwin, C. (1859). On the Origin of Species by Means of Natural Selection (or the Preservation of Favoured Races in the Struggle for Life. John Murray

10 Dobzhansky, T. (1973) Nothing in biology makes sense except in the light of evolution. *Am. Biol. Teach.* 35, 125–129

11 Adl, S.M. *et al.* (2007) Diversity, nomenclature, and taxonomy of protists. *Syst. Biol.* 56, 684–689

12 Entwisle, T.J. and Weston, P.H. (2005) Majority rules, when systematists disagree. *Aust. Syst. Bot.* 18, 1–6

13 Morrison, D.A. *et al.* (2004) The current status of the small subunit rRNA phylogeny of the Coccidia (Sporozoa). *Int. J. Parasitol.* 34, 501–514

14 Slapeta, J.R. *et al.* (2002) Dog shedding oocysts of *Neospora caninum*: PCR diagnosis and molecular phylogenetic approach. *Vet. Parasitol.* 109, 157–167

15 Pérez-Tris, J. *et al.* (2005) What are malaria parasites? *Trends Parasitol.* 21, 209–211

16 Morrison, D.A. (2008) Prospects for elucidating the phylogeny of the Apicomplexa. *Parasite* 15, 191–196

17 Barta, J.R. (2001) Molecular approaches for inferring evolutionary relationships among protistan parasites. *Vet. Parasitol.* 101, 175–186

18 Barta, J.R. and Thompson, R.C.A. (2006) What is *Cryptosporidium*? Reappraising its biology and phylogenetic affinities. *Trends Parasitol.* 22, 463–468

19 Leander, B.S. (2007) Marine gregarines: evolutionary prelude to the apicomplexan radiation? *Trends Parasitol.* 24, 60–67

20 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

21 Jakes, K. *et al.* (2003) Phylogenetic relationships of *Hepatozoon* (*Haemogregarina*) *boigae, Hepatozoon* sp., *Haemogregarina clelandi* and *Haemoproteus chelodina* from Australian reptiles to other Apicomplexa based on cladistic analyses of ultrastructural and life-cycle characters.. *Parasitology* 126, 555–559

22 Templeton, T.J. (2007) Whole-genome natural histories of Apicomplexan surface proteins. *Trends Parasitol.* 23, 205–212

23 Kuo, C.H. and Kissinger, J.C. (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC Evol. Biol.* 8, 108

24 Xiao, L. *et al.* (2002) Host adaptation and host-parasite co-evolution in *Cryptosporidium*: implications for taxonomy and public health. *Int. J. Parasitol.* 32, 1773–1785

25 Morrison, D.A. (2006) Phylogenetic analyses of parasites in the new millennium. *Adv. Parasitol.* 63, 1–124

26 Rathore, D. *et al.* (2001) A phylogenetic comparison of gene trees constructed from plastid, mitochondrial and genomic DNA of *Plasmodium* species. *Mol. Biochem. Parasitol.* 114, 89–94

27 Perkins, S.L. *et al.* (2007) The phylogeny of rodent malaria parasites, simultaneous analysis across three genomes. *Infect. Genet. Evol.* 7, 74–83

28 Martinsen, E.S. *et al.* (2008) A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. *Mol. Phylogenet. Evol.* 47, 261–273

29 Leander, B.S. *et al.* (2003) Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and β-tubulin. *Int. J. Syst. Evol. Microbiol.* 53, 345–354

30 Li, L. *et al.* (2004) ApiEST-DB, analyzing clustered EST data of the Apicomplexan parasites. *Nucleic Acids Res.* 32, D326–D328

31 de la Torre, J.E. *et al.* (2006) ESTimating plant phylogeny: lessons from partitioning. *BMC Evol. Biol.* 6, 48
32 Aurrecoechea, C. *et al.* (2006) ApiDB, integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res.* 35, D427–D430
33 Matsubayashi, M. *et al.* (2005) Molecular characterization of crane Coccidia, *Eimeria gruis* and *E. reichenowi*, found in feces of migratory cranes. *Parasitol. Res.* 97, 80–83
34 Kopecná, J. *et al.* (2006) Phylogenetic analysis of coccidian parasites from invertebrates: search for missing links. *Protist* 157, 173–183
35 Allsopp, M.T.E.P. and Allsopp, B.A. (2006) Molecular sequence evidence for the reclassification of some Babesia species. *Ann. N. Y. Acad. Sci.* 1081, 509–517
36 Ciancio, A. *et al.* (2008) Redescription of *Cardiosporidium cionae* (Van Gaver and Stephan, 1907) (Apicomplexa: Piroplasmida), a plasmodial parasite of ascidian haemocytes. *Eur. J. Protistol.* 44, 181–196
37 Rueckert, S. and Leander, B.S. (2008) Morphology and phylogenetic position of two novel marine gregarines (Apicomplexa, Eugregarinorida) from the intestines of North-eastern Pacific ascidians. *Zool. Scr* 37, 637–645
38 Jirků, M. *et al.* (2009) Goussia Labbé, 1896 (Apicomplexa, Eimeriorina) in Amphibia: diversity, biology, molecular phylogeny and comments on the status of the genus. *Protist* 160, 123–136
39 Leclerc, M.C. *et al.* (2004) Evolutionary relationships between 15 *Plasmodium* species from New and Old World primates (including humans): a 18S rDNA cladistic analysis. *Parasitology* 129, 1–8
40 Hagner, S.C. *et al.* (2007) Bayesian analysis of new and old malaria parasite DNA sequence data demonstrates the need for more phylogenetic signal to clarify the descent of *Plasmodium falciparum*. *Parasitol. Res.* 101, 493–503
41 Dávalos, L.M. and Perkins, S.L. (2008) Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. *Genomics* 91, 433–442
42 Nishimoto, Y. *et al.* (2008) Evolution and phylogeny of the heterogeneous cytosolic SSU rRNA genes in the genus *Plasmodium*. *Mol. Phylogenet. Evol.* 47, 45–53
43 Taylor, F.J.R. (1999) Ultrastructure as a control for protistan molecular phylogeny. *Am. Nat.* 154, S125–S136
44 Holmdahl, O.J.M. *et al.* (1999) Evolution of ruminant *Sarcocystis* (Sporozoa) parasites based on small subunit rDNA sequences. *Mol. Phylogenet. Evol.* 11, 27–37
45 Barta, J.R. *et al.* (2005) The genus *Atoxoplasma* (Garnham 1950) as a junior objective synonym of the genus *Isospora* (Schneider 1881) species infecting birds and resurrection of *Cystoisospora* (Frenkel 1977) as the correct genus for *Isospora* species infecting mammals. *J. Parasitol.* 91, 726–727
46 Kvicerova, J. *et al.* (2008) Phylogenetic relationships among *Eimeria* spp. (Apicomplexa, Eimeriidae) infecting rabbits: evolutionary significance of biological and morphological features. *Parasitology* 135, 443–452
47 Dahlgren, S.S. *et al.* (2008) Phylogenetic relationships between *Sarcocystis* species from reindeer and other Sarcocystidae deduced from ssu rRNA gene sequences. *Vet. Parasitol.* 151, 27–35
48 Levine, N.D. (1988) Progress in taxonomy of the Apicomplexan protozoa. *J. Protozool.* 35, 518–520
49 Chapman, A.D. (2005) *Numbers of Living Species in Australia and the World*. Report for the Australian Department of the Environment and Heritage, Canberra
50 Gregory, T.R. (2008) Understanding evolutionary trees. *Evol. Educ. Outreach* 1, 121–137
51 Baum, D.A. *et al.* (2005) The tree-thinking challenge. *Science* 310, 979–980
52 Wilkinson, M. *et al.* (2007) Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends Ecol. Evol.* 22, 114–115
53 Krell, F-T. and Cranston, P.S. (2004) Which side of a tree is more basal? *Syst. Entomol.* 29, 279–281
54 Sanderson, M.J. *et al.* (2008) The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. *Syst. Biol.* 57, 335–346
55 Kuo, C.H. *et al.* (2008) The apicomplexan whole genome phylogeny: an analysis of incongruence among gene trees. *Mol. Biol. Evol.* 25, 2689–2698
56 Mather, M.W. and Vaidya, A.B. (2008) Mitochondria in malaria and related parasites: ancient, diverse and streamlined. *J. Bioenerg. Biomembr.* 40, 425–433
57 Seeber, F. *et al.* (2008) Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. *Trends Parasitol.* 24, 468–478
58 Waller, R.F. and McFadden, G.I. (2005) The apicoplast, a review of the derived plastid of Apicomplexan parasites. *Curr. Issues Mol. Biol.* 7, 57–80
59 Corliss, J.O. (1962) Taxonomic-nomenclatural practices in protozoology and the new international code of zoological nomenclature. *J. Protozool.* 9, 307–324
60 Roger, A.J. and Hug, L.A. (2006) The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 361, 1039–1054
61 Morrison, D.A. (2006) Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.* 19, 479–539
62 Morrison, D.A. A framework for phylogenetic sequence alignment. *Plant Syst. Evol.* (in press)
63 Müller, K. (2006) Incorporating information from length-mutational events into phylogenetic analysis. *Mol. Phylogenet. Evol.* 38, 667–676
64 Ochoterena, H. Homology in coding and non-coding DNA sequences: a parsimony perspective. *Plant Syst. Evol.* (in press)

Celebrating Darwin: Evolution of Hosts, Microbes and Parasites

To commemorate the 200th anniversary of Charles Darwin's birthday (12th February, 1809), *Trends in Parasitology* is featuring several articles with evolutionary themes in the course of 2009, along with *Trends in Microbiology* and *Cell Host & Microbe*.

In this issue of *Trends in Parasitology* (August 2009), David Morrison has a Review entitled '**Evolution of the Apicomplexa: where are we now?**' on the phylogeny of apicomplexan parasites.

In the August issue of *Trends in Microbiology*, Maureen O'Malley has an Opinion article entitled '**What did Darwin say about microbes, and how did microbiology respond?**'. This article presents a historical viewpoint of Darwin's writings on microbes, his interaction with microbiologists and how this shaped microbiological research.

Forthcoming articles:

Genetic and genomic analyses of host-pathogen interactions in malaria by Silayuv E. Bongfen, Aurelie Laroque, Joanne Berghout and Philippe Gros
Trends in Parasitology, September 2009.

Infrequent marine-freshwater transitions in the microbial world by Ramiro Logares, Jon Bråte, Stefan Bertilsson, Jessica L. Clasen, Kamran Shalchian-Tabrizi and Karin Rengefors
Trends in Microbiology, September 2009

All the articles in the series are collected in the following webpage:
http://www.cell.com/trends/parasitology/Darwin