Parasitol Res (1995) 81:696–699

SHORT COMMUNICATION

John Ellis · David Morrison

Effects of sequence alignment on the phylogeny of *Sarcocystis* deduced from 18S rDNA sequences

Received: 8 February 1995 / Accepted: 3 April 1995

Abstract The family Sarcocystidae contains a wide variety of parasitic protozoa, some of which are important pathogens of livestock and humans. The taxonomic relationships between two of the genera in this family (*Toxoplasma* and *Sarcocystis*) have been debated for a number of years and remain controversial. Recent studies, from comparisons of 18S rDNA-sequence data, have suggested that *Sarcocystis* is paraphyletic, although a hypothesis supporting monophyly of *Sarcocystis* could not be rejected. The present study shows that the phylogenetically informative nucleotide positions within the 18S rDNA are primarily located in the regions that make up the helices in the secondary structure of the 18S rRNA. A phylogenetic analysis of 18S rDNA-sequence data aligned by secondary structure constraints, or a subset of the data corresponding to all nucleotides found in the helices, provide unambiguous evidence supporting monophyly of *Sarcocystis*.

One of the essential steps in any evolutionary study is to establish an hypothesis of homology among the taxa being studied, and in molecular analyses this involves pairwise alignment of the nucleotides or amino acids. 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 maximize some criterion for optimality of the correspondence (Chan et al. 1992). There are many alignment algorithms currently available (Waterman 1989; Doolittle 1990; Chan et al. 1992), which maximize a wide variety of mathematical functions measuring correspondence between the sequences. However, most of these algorithms use heuristic procedures, and hence do not make use of knowledge of the secondary structure of the molecules when such knowledge is available (Chan et al. 1992). Analyses based on alignments that do not incorporate the known biological function of the nucleotide sequences are likely, in general, to be inferior to analyses where the nucleotides are aligned according to their known function within the sequences (Hillis and Dixon 1991). Furthermore, when alignment regions contain many gaps (insertions and deletions), these regions are often excised prior to phylogeny reconstruction on the grounds that they are phylogenetically uninformative, although there are few objective criteria for these judgements (Gatesby et al. 1993).

These considerations suggest that superior phylogenetic analyses should be provided using alignments based on secondary structure, taking into consideration the cladistic informativeness of different regions within the molecule (Wheeler and Honeycutt 1988; Smith 1989; Dixon and Hillis 1993). In the present study we tested this hypothesis for a group of taxa whose phylogeny has previously proved problematic so as to see whether the problems simply derive from inadequate alignments (previous analyses used heuristic techniques only) and/or the inclusion of uninformative data (previous analyses did not exclude any of the nucleotide positions).

The phylogenetic relationships among protozoan parasites of the family Sarcocystidae as inferred from comparisons of 18S ribosomal RNA (rRNA) sequence data remain ambiguous. Generally speaking, it is now accepted that the genera *Toxoplasma* and *Sarcocystis* form a monophyletic group (i.e. are derived from a recent common ancestor; Johnson et al. 1988; Tenter et al. 1992; Ellis et al. 1994a; Holmdahl et al. 1994; Ellis et al. 1995). However, a recent report concluded that the genus *Sarcocystis* itself is paraphyletic, since the analysis of partial 18S rRNA sequences identified two monophyletic subgroups within *Sarcocystis*, one subgroup containing two species having felids as definitive hosts and the other subgroup containing four species having canids as definitive hosts, and these two clades together were monophy-

J. Ellis (✉)
Department of Cell and Molecular Biology,
University of Technology Sydney, Gore Hill,
New South Wales, Australia

D. Morrison
Department of Environmental Biology and Horticulture,
University of Technology Sydney, Gore Hill,
New South Wales, Australia

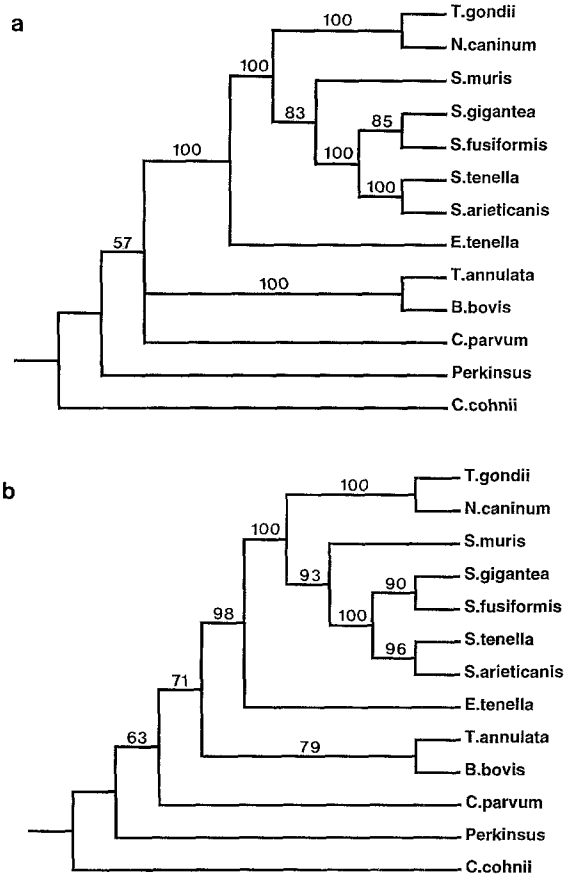

Fig. 1a, b Results of parsimony analysis. **a** 50% majority-rule consensus tree obtained from a bootstrap analysis performed on the entire sequence data set. **b** 50% majority-rule consensus tree resulting from a bootstrap analysis performed on the nucleotides located in helices. The values at the nodes show the percentage of times that each node was supported out of 100 bootstrap replicates. The branch lengths are arbitrary. The sequence data analyzed were obtained from the following taxa: *Toxoplasma gondii, Neospora caninum, Sarcocystis muris, S. gigantea, S. fusiformis, S. tenella, S. arieticanis, Eimeria tenella, Theileria annulata, Babesia bovis, Cryptosporidium parvum, Perkinsus* spp. and *Crypthecodinium cohnii*

letic only if *T. gondii* was included (Tenter et al. 1992). Similarly, Ellis et al. (1994b) used a relatively small data set derived from the five prime end of the 18S rRNA gene (rDNA) and found evidence for paraphyly of *Sarcocystis*. Both of these studies therefore also provided further support for a correlation between parasite phylogeny and that of their definitive hosts (Barta 1989). Subsequently, however, using a much larger data set derived from the entire 18S rDNA that were aligned by the program Clustal V, Ellis et al. (1995) concluded that an hypothesis promoting monophyly of *Sarcocystis* could not be rejected.

Despite the apparent inability of the data set to resolve completely the controversy that remains over the phylogeny of *Sarcocystis*, the most obvious conclusion that can be drawn from these studies is that there are probably two subsets of data within the data set analyzed by Ellis et al. (1995), one supporting monophyly of *Sarcocystis* and the other supporting paraphyly. Herein we show that the phylogenetically informative nucleotide positions are primarily located within regions that correspond to the helices that make up much of the secondary structure of the 18S rRNA. A phylogenetic analysis of either the entire data set aligned according to secondary structure constraints, or a subset of the data corresponding to all nucleotides found in the helices, provides unambiguous evidence supporting monophyly of *Sarcocystis*.

The DNA sequences analyzed in this study were those previously described by Ellis et al. (1995), and the 18S rDNA sequence of *S. fusiformis* was also included (Holmdahl et al. 1994). The multiple sequence alignment was that described by Van de Peer et al. (1994), which defines the complete secondary structure of 18S rRNA from the genera *Toxoplasma* and *Sarcocystis* as well as the other taxa analyzed in the present study, this alignment being kindly supplied by Peter De Rijk (Departement Biochemie, Universiteit Antwerpen). The alignment contains all the sequences in the data set except that of *Eimeria tenella*, aligned according to the two constraints of homology and secondary structure, and all of the currently recognised helices and single-stranded regions within the 18S rRNA were identified (Van de Peer et al. 1994). The sequence derived from the 18S rDNA of *E. tenella* was added to the alignment using the DCSE sequence editor (De Rijk and De Wachter 1993), and all of the helices and loops were similarly identified with this editor. The alignment used in this study is available from the authors.

Three data sets were then analyzed phylogenetically using the branch and bound option (Hendy and Penny 1982) in PAUP 3.11 (Swofford 1990) using *Crypthecodinium cohnii* and *Perkinsus* spp. as the outgroup (Ellis et al. 1994b, 1995). The data sets were: (1) an alignment containing all nucleotides derived from the 18S rRNA gene of all taxa that starts 3 nucleotides upstream of helix 1 and finishes 9 nucleotides downstream of helix 50 (using the nomenclature described in Neefs et al. 1993); (2) an alignment containing all nucleotides located in helices (including unbase-paired nucleotides) 4 through 50′; and (3) an alignment containing all of the remaining nucleotides located in all other regions between helices 4 through 50′ (referred to hereafter as the single-stranded regions) of the 18S rRNA structure. The latter two alignments were created from the total alignment using the DCSE editor.

The phylogenetic analysis of the total sequence alignment (containing 2050 characters) found 2 near-most parsimonious trees of 1048 and 1049 steps. The next shortest tree was a further 2 steps longer than these. Both of the shortest trees supported monophyly of *Sarcocystis* and monophyly of *Sarcocystis* plus *Toxoplasma* and *Neospora*. The two trees differed only in the placement of the piroplasms (*Babesia bovis* and *Theileria annulata*) in relation to the outgroup. A bootstrap analysis showed that 100% of the bootstrap replicates supported monophyly of the Sarcocystidae and 83% of the replicates supported monophyly of *Sarcocystis* (Fig. 1a). Similarly, the

Table 1 Summary of the extent of pairwise nucleotide differences among the taxa. Nucleotide differences between the total sequence data sets are given above the diagonal; nucleotide differences between the helical regions are shown below the diagonal

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
1. *Toxoplasma gondii*	–	5	66	114	115	112	118	124	178	187	188	190	247
2. *Neospora caninum*	3	–	69	115	118	112	118	124	177	183	190	188	247
3. *Sarcocystis muris*	38	39	–	101	10	89	94	125	178	184	198	194	237
4. *S. gigantea*	77	78	61	–	86	62	91	158	195	183	195	219	244
5. *S. tenella*	79	81	64	52	–	77	56	153	194	181	188	218	234
6. *S. fusiformis*	74	74	53	33	48	–	79	146	196	184	185	210	237
7. *S. arieticanis*	76	75	53	52	35	46	–	150	192	192	199	207	238
8. *Eimeria tenella*	83	84	84	112	108	104	98	–	163	165	196	166	216
9. *Theileria annulata*	126	126	127	137	134	136	128	107	–	122	178	182	189
10. *Babesia bovis*	139	137	140	140	134	141	144	114	98	–	174	169	188
11. *Cryptosporidium parvum*	145	146	148	150	145	143	150	144	129	127	–	175	184
12. *Perkinsus* spp.	144	144	145	165	165	159	151	119	128	127	129	–	173
13. *Crypthecodinium cohnii*	188	189	188	184	172	179	176	158	146	142	132	128	–

analysis using nucleotide positions found only in the helical regions (data set of 1301 characters) found 2 near-most parsimonious trees of 719 and 720 steps. The next shortest tree was a further 3 steps longer than these. Both of the shortest trees supported monophyly of *Sarcocystis*, and they differed only in the placement of *Cryptosporidium parvum* and *Perkinsus* spp. A bootstrap analysis showed that 100% of the bootstrap replicates supported monophyly of the Sarcocystidae and 93% supported monophyly of *Sarcocystis* (Fig. 1b). Finally, the analysis of the nucleotides found in the single-stranded regions of the 18S rRNA structure revealed many near-most parsimonious trees, including 5 trees of 359 steps; 2 of these trees supported monophyly of *Sarcocystis* and 3 supported paraphyly. A further 75 trees were found within 3 steps of these trees. A bootstrap analysis of these data would be uninformative, because there is apparently very little unequivocal phylogenetic information in this part of the data set.

The pairwise number of nucleotide substitution differences detected among the taxa for two of the data sets analyzed are shown in Table 1. Ellis et al. (1995) observed that there were fewer nucleotide differences between the 18S rRNA genes of *S. muris* and *T. gondii* than there were between *S. muris* and the other *Sarcocystis* species analyzed, and this finding was used to explain partially the hypothesized paraphyly of *Sarcocystis*. Our analyses confirm that there are fewer nucleotide substitution differences between *S. muris* and both *T. gondii* and *N. caninum* than there are between *S. muris* and most of the rest of the *Sarcocystis* species, both for the total sequence and for the helical regions (Table 1). However, it is clear from our analyses that many of the extra nucleotide substitutions are not phylogenetically informative.

In conclusion, the results obtained from the phylogenetic analyses presented herein provide substantial evidence in support of the hypothesis that the members of *Sarcocystis* are all derived from a recent common ancestor. The failure of previous phylogenetic analyses to resolve this issue appears to be the result of several factors. Firstly, the sequence alignments previously used were based on heuristic alignment methods rather than on the known secondary structure of the 18S rRNA molecule. Secondly, our analyses show that the majority of the phylogenetically informative sites are located in the helices (irrespective of whether the nucleotides were base-paired or not). The helices make up a relatively large proportion (nearly two-thirds) of the 18S rRNA secondary structure, and removing the single-stranded parts of the sequences from the analysis apparently removes a great deal of "noise" that is obscuring the underlying phylogenetic signal.

Acknowledgements We thank Peter De Rijk and Yves Van de Peer (Universiteit Antwerpen, Belgium) for their help and assistance; we also thank Dr. Astrid M. Tenter (Hannover, Germany) and Prof. Alan M. Johnson (UTS, Australia) for their encouragement and support.

References

Barta JR (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

Chan SC, Wong AKC, Chiu DKY (1992) A survey of multiple sequence comparison methods. Bull Math Biol 54:563–598

De Rijk P, De Wachter R (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. CABIOS 9:735–740

Dixon MT, Hillis DM (1993) Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. Mol Biol Evol 10:256–267

Doolittle RF (ed) (1990) Molecular evolution: computer analysis of protein and nucleic acid sequences. Methods Enzymol 183:303–502

Ellis J, Morrison DA, Johnson AM (1994a) The molecular phylogeny of sporozoan parasites. Today's Life Sci 6:30–34

Ellis J, Luton K, Baverstock PR, Brindley PJ, Nimmo KA, Johnson AM (1994b) The phylogeny of *Neospora caninum*. Mol Biochem Parasitol 64:303–311

Ellis J, Luton K, Baverstock PR, Whitworth G, Tenter AM, Johnson AM (1995) Phylogenetic relationships between *Toxoplasma* and *Sarcocystis* deduced from a comparison of 18S rDNA sequences. Parasitology (in press)

Gatesby J, Desalle R, Wheeler W (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. Mol Phyl Evol 2:152–157

Hendy MD, Penny D (1982) Branch and bound algorithms to determine minimal evolutionary trees. Math Biosci 59:277–290

Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453

Holmdahl JM, Mattson JG, Uggla A, Johansson K-E (1994) The phylogeny of *Neospora caninum* and *Toxoplasma gondii* based on ribosomal RNA sequences. FEMS Micro Lett 119:187–192

Johnson AM, Illana S, Hakendorf P, Baverstock PR (1988) Phylogenetic relationships of the apicomplexan protist *Sarcocystis* as determined by small subunit ribosomal RNA sequence comparison. J Parasitol 74:847–860

Neefs J-M, Van de Peer Y, De Rijk P, Chapeele S, De Wachter R (1993) Compilation of small ribosomal subunit RNA structures. Nucleic Acids Res 21:3025–3049

Smith AB (1989) RNA sequence data in phylogenetic reconstruction: testing the limits of its resolution. Cladistics 5:321–344

Swofford DL (1990) PAUP: Phylogenetic Analysis Using Parsimony, version 3.1.1. Illinois Natural History Survey, Champaign, Illinois

Tenter AM, Baverstock PR, Johnson AM (1992) Phylogenetic relationships of *Sarcocystis* species from sheep, goats, cattle and mice based on ribosomal RNA sequences. Int J Parasitol 22:503–513

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

Waterman MS (1989) Mathematical methods for DNA sequences. CRC, Boca Raton, Florida

Wheeler WC, Honeycutt RL (1988) Paired sequence difference in ribosomal RNAs: evolution and phylogenetic implications. Mol Biol Evol 5:90–96