

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

**Studies on the phylogeny and gene structure
of early-branching eukaryotes.**

by

Andrew J. Roger

**Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy**

at

**Dalhousie University
November, 1996**

© Copyright by Andrew J. Roger, 1996



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*

Our file *Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-24782-1

DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

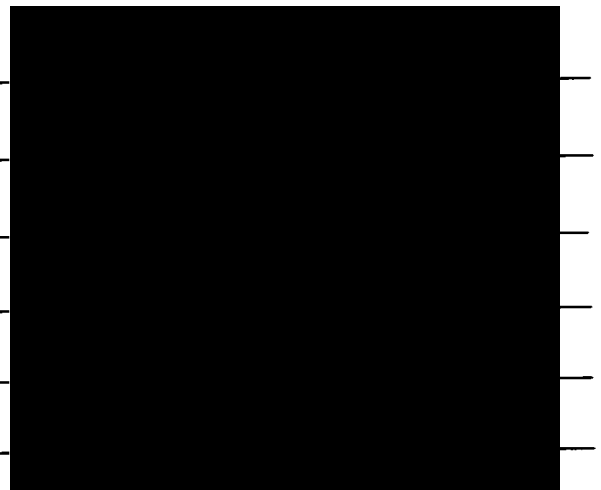
The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “Studies on the phylogeny and gene structure of early-branching eukaryotes”

by Andrew James Roger

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: November 14, 1996

External Examiner
Research Supervisor
Examining Committee



DALHOUSIE UNIVERSITY

DATE: November 14, 1996

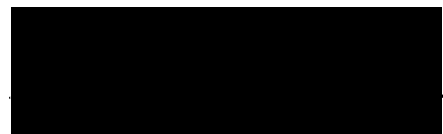
AUTHOR: Andrew James Roger

TITLE: Studies on the phylogeny and gene structure of early-branching
eukaryotes

DEPARTMENT OR SCHOOL: Biochemistry

DEGREE: PhD CONVOCATION: May YEAR: 1997

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.



Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

to Udeni

Table of Contents

Table of Contents	v
Illustrations and Tables	vi
Abstract	vii
Abbreviations and Symbols used	viii
Acknowledgements	ix
Introduction	1
Materials and Methods	14
Chapter 1	32
Chapter 2	72
Chapter 3	112
Chapter 4	126
References	135

Figures and Tables

Figure 1.1	Alignment of EF-1 α homologs characterized in this study	39
Figure 1.2	Southern blots of various genomic DNAs probed with their respective EF-1 α fragments	42
Figure 1.3	An alignment of the partial <i>Nosema locustae</i> β -tubulin gene with homologs from other organisms	44
Figure 1.4	A possibly homologous insertion in EF-1 α genes of Microsporidia, Fungi and Metazoa	46
Figure 1.5	Eukaryotic phylogeny inferred from EF-1 α sequences	47
Figure 1.6	EF-1 α phylogeny with the Microsporidia removed	52
Figure 1.7	Eukaryotic phylogeny inferred from β -tubulin sequences	54
Figure 1.8	A hypothesis of relationships among eukaryotes based on selected molecular and ultrastructural data	71
Figure 2.1	An alignment of TPI sequences obtained in this study with homologs from other organisms	85
Figure 2.2	An alignment of GAPDH sequences obtained in this study with homologs from other organisms	86
Figure 2.3	Positions occupied by introns in TPI homologs	88
Figure 2.4	Positions occupied by introns in GAPDH homologs	89
Figure 2.5	A phylogeny of TPI sequences using the neighbour-joining method	93
Figure 2.6	A hypothetical phylogeny of the TPI gene	94
Figure 2.7	Phylogeny of GAPDH using the neighbour-joining method	96
Figure 2.8	A hypothetical phylogeny of GAPDH	97
Figure 2.9	Estimation of the maximum likelihood ratios of the rate of intron gain to loss	101
Figure 3.1	Phylogenies of cpn60 homologs	117
Figure 3.2	The impact of the sampling of Rickettsiales species on the bootstrap support for two alternative topologies of the cpn60 tree	119
Figure 4.1	An alignment of Fe-SOD sequences obtained in this study with homologs from other organisms	129
Figure 4.2	A tree of Fe-SOD inferred using the parsimony method	130
Table 1.1	Imposing a Ciliate/Alveolate and a Heterolobosea tree on the EF-1 α topology	51
Table 1.2	The position of Microsporidia in the EF-1 α tree	58
Table 1.3	The position of Microsporidia in the β -tubulin tree	60
Table 2.1	A comparison of GAPDH intron positions used in this study with those reported in Kersenach et al. (1994)	91
Table 2.2	A comparison of the minimum number of events required for each introns theory to explain the TPI and GAPDH data	99
Table 2.3	Maximum likelihood reconstructions of ancestral intron sequences	104

Abstract

The Archezoa hypothesis holds that several living protist groups constitute primitive eukaryotic lineages that diverged from the main eukaryotic lineage prior to the endosymbiotic origin of mitochondria. Several aspects of this hypothesis were tested.

Firstly, elongation factor 1 α and β -tubulin genes were developed as phylogenetic markers for early eukaryote evolution by amplifying several homologs from mitochondriate and amitochondriate protists. The phylogenies of these genes suggest that, after a few early branchings, a deep split occurred in eukaryote evolution that resulted in the emergence of two distinct protist/multicell superclusters. Interestingly, the amitochondriate Microsporidia are strongly placed as a sister group to the Fungi in the β -tubulin tree. Statistical tests using both datasets, the shared presence of an insertion in microsporidian, fungal and metazoan EF-1 α genes as well as ultrastructural considerations suggest that this phylogenetic position is the best supported by the data.

To evaluate the "introns-late" claim that spliceosomal introns are a derived feature of eukaryotic genome absent in the earliest protist lineages, triosephosphate isomerase and glyceraldehyde-3-phosphate dehydrogenase genes were obtained from several of the putatively early-branching eukaryotic groups. These, along with other homologs, were assembled into intron datasets and a novel maximum likelihood method was used to evaluate the likelihood of "introns-early" and "introns-late" models. The latter view was shown to confer the greatest probability on the data.

In the final two studies, the Archezoa hypothesis was directly tested. A mitochondrial-like chaperonin 60 gene was cloned and sequenced from the early-branching amitochondriate protist *Trichomonas vaginalis*. Phylogenetic analyses of datasets including this sequence coupled with the hydrogenosomal location for the protein argues strongly for a common ancestry for hydrogenosomes and mitochondria. In the last study, several of the early-branching groups were shown to possess iron superoxide dismutase genes that cluster with proteobacterial homologs. The means by which these genes were acquired by these protists are unclear, but the possibility exists that some or all of them may derive from the mitochondrial endosymbiosis.

Abbreviations and Symbols used

α -tubulin	alpha tubulin
ACR	ancient conserved region
bp	base pair
β -tubulin	beta tubulin
cDNA	complementary DNA
CTAB	cetyltrimethylammonium bromide
EDTA	ethylene-diamine-tetra-acetic acid
EF-1 α	elongation factor 1alpha
Fe-/Mn-SOD	iron or manganese superoxide dismutase
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
kb	kilobase pair
mb	megabase pair
MPR	maximum parsimony reconstruction
PCR	polymerase chain reaction
SDS	sodium dodecyl sulfate
SSU rRNA	small subunit ribosomal RNA
TPI	triosephosphate isomerase
tRNA	transfer RNA

Acknowledgements

There are many people I would like to thank for helping me both scientifically and personally over the last few years.

Tom Cavalier-Smith inspired me to go into the field of molecular evolution and spent many long hours discussing with me his theories of the origin and phylogeny of virtually everything living. These hours taught me how to think about phylogeny and evolution. They also showed me that if one tries hard enough to gather all of the relevant information, usually a testable phylogenetic hypothesis will emerge.

I would like to thank Ford for providing a laboratory and research funding that allowed me to pursue anything I wanted. There is no one I enjoy disagreeing with more than Ford -- this is not because he is wrong a lot (in fact quite the opposite is true) -- rather it is because a disagreement with him is always intellectual and not personal. I hope that I have learned from his example. By cultivating a non-confrontational atmosphere, Ford has made an ideal environment for fruitful thought. In his lab, the ultimate answers to the questions matter more than the methodology applied to them.

Over the years I have enjoyed the interaction with all of the people in the lab and much of it has probably contributed to whatever success I may have had. In the beginning, Leo, Wan and Sandie helped me a lot with technical matters. Cheryl was a supportive ally to have, especially when things went wrong. Olof was a great help (and fun), especially when it came to fighting Canada Post, whipping off witty FAXes to Belgium and making rotten hay plates to feed the acrasids. Most of my lab work was done alongside Dave and Patrick with whom symbiotic relationships developed when it came to running gels, getting plasmids and trying new techniques. I am indebted to both of them for putting up with the stream of consciousness that tends to spill out of my mouth. Thanks

are due to Jim for joining me in the fight against the evils of chimaeric theories. I was very fortunate to start my elongation factor work at the time that Sandie came to Halifax. Her generosity with primers and advice on phylogeny and all the good bug-talk are greatly appreciated. Arlin has been enormously generous with his time throughout my time in the lab. His vigorous skepticism, although not always warranted, has been very useful in helping me articulate my thoughts more rigorously and clearly. I thank Eve for working with me in the midst of my mess...I am sure that her lessons on how to be neat must have had some effect! Annalee was always a cheery face in the morning. Her stories reminded me of what real people's lives were like and what mine could be once I was done! I am grateful to Naomi who is the only person who ever actually believed that I was going to make a *Nosema* library! As it turned out, Naomi, Claire and Mike and I (aka the Dream-team) all made the dream a reality. John was a great email friend and continues to inspire me now in person. His help with compiling introns and careful reading of manuscripts is greatly appreciated.

For the past few years, I have benefitted from wonderful collaborations with people in the US and overseas. Miklós Müller has helped me throughout my project and was very generous with materials and ideas. I feel honoured to have worked with and learned from the "father of the hydrogenosome". Graham Clark has been a good friend and an ideal collaborator. He has taught me much about protists. Likewise, Hervé Philippe has taught me a lot about phylogenetics and our debates over the years have been extremely fruitful in generating new ideas. Joe Felsenstein and Ziheng Yang have both patiently answered my incessant maximum likelihood questions via email. Ziheng has also been a great collaborator. Greg Hinkle has been extremely generous in giving me DNA and playing tree games with me. It was so fun, that I decided to go down there to continue playing them!

My parents have been extremely supportive over the last few years. On various occasions, my father has offered key advice that has helped me weather crises, either real or imagined. I am thankful to my parents for listening to me whine periodically about my hard life as a graduate student!

Lastly, I would like to thank Udeni for her patience and love over the last few years.

Introduction

The establishment of a paradigm.

The enormous gulf in cell structure between prokaryotes and eukaryotes was not appreciated by many biologists until the writings of R. Stanier and C. B. van Niel. In 1962, they attempted to more clearly delimit this boundary in a landmark work "The Concept of a Bacterium" by emphasizing fundamental differences between these cell types such as: (1) the presence in eukaryotes, but absence in prokaryotes, of a nuclear membrane and membrane-bound respiratory and photosynthetic organelles, (2) the presence of a single chromosome found in prokaryotes but not in eukaryotes and the occurrence of mitosis in eukaryotes, lacking in prokaryotes, (3) detailed structural differences between eukaryotic and prokaryotic motility organelles, and (4) the presence of peptidoglycan walls in prokaryotes not found in eukaryotes (Stanier & van Niel, 1962). This recognition of these basic differences in cellular properties, led naturally to the view that both prokaryotes and eukaryotes were evolutionarily distinct entities. However, during the following decade, the relationship between these two kinds of entities remained a frustrating puzzle.

Eukaryote origins: symbiosis or direct filiation?

During this period, two kinds of theories regarding these relationships prevailed, and both assumed that the prokaryotic form was ancestral to the eukaryotes. One position saw the development of the eukaryotic cell directly from prokaryotic ancestors by the modification of existing structures and the origin of some new ones. This hypothesis became known as the "direct filiation" or "autogenous origin" hypothesis for origin of the eukaryote cell and its organelles, and it was very much the default view held by cell biologists in the early 1960s. However in this period, the clear demonstration of DNA enclosed

in chloroplasts (Ris & Plaut, 1962) and mitochondria (Nass & Nass, 1963) set the stage for the re-emergence of a very different theory that held that these organelles were acquired by eukaryotes through the endosymbiosis of free-living bacteria. This latter hypothesis had roots in the writings of a number of biologists around the turn of the century. Two German scientists, Andreas Schimper and Richard Altmann, proposed the symbiotic origin of chloroplasts and mitochondria respectively in the late nineteenth century, and ideas similar to these were later elaborated by Famintzyn and Merechowsky (in Russia), Portier (in France) and Wallin (in the United States) in the early twentieth century (see Sapp, 1994 and references therein). For many years, these symbiotic theories existed on the fringe of biology, and they fell into obscurity during the middle part of the twentieth century. However in 1967, two papers appeared that issued in a new era of support for the endosymbiotic theory of eukaryotic cell origins.

Goksøyr, in a small correspondence to *Nature*, wrote that eukaryotic cells might have evolved by a merger or "coenocytic" association of prokaryotic cells (Goksøyr, 1967). Once this fusion took place, he argued, a multiple genomed entity was the result. As a consequence, enormous selective pressure on this organism to properly segregate its genetic material to daughter cells led first to the evolution of an efficient mitosis, and later to the origin of the nucleus. Then, as a response to the increasing oxygen tension in the early atmosphere, this anaerobe engulfed an aerobic bacterium that eventually became the mitochondrion and a true aerobic eukaryotic cell was born. Later, cells like these diversified by the acquisition of symbiotic cyanobacteria into present-day algae (Goksøyr, 1967).

In the same year, Margulis (then Sagan) elaborated a similar, but much more detailed, theory of the endosymbiotic origin of the eukaryotic cell. In what was later referred to as the Serial Endosymbiotic Theory (SET), she argued that

eukaryotic cells were born out of a stepwise series of endosymbiotic events (Sagan, 1967). The earliest prokaryotes were photosynthetic or heterotrophic anaerobes, she claimed, and due to the nature of photosynthesis the oxygen composition of the atmosphere began to increase. Because of this, heterotrophic anaerobes of the period suffered heavy negative selection. In response, one of them engulfed an aerobic bacterium to become an amitotic aerobic protoeukaryote. The next step was the acquisition, by this amoeboid creature, of a motility organelle which could have been supplied by a symbiotic spirochaete-like organism. In her theory, Margulis suggested that this motility symbiosis led to not only eukaryotic flagella, but also to the development of centrioles and centromeres, as well as connections between them that would ensure proper segregation of the endosymbionts and their genes to daughter cells. Similar to Goksøyr's theory, she proposed that photosynthetic eukaryotes evolved much later by the acquisition of cyanobacteria in several different events of endosymbiosis. Central to her arguments was the view that eumitosis evolved multiple times independently after the acquisition of flagella, and on basis of the different mitotic patterns observed in diverse eukaryotic groups, she proposed a phylogeny of protists.

These papers by Goksøyr and especially Margulis, probably led to the polarization of views regarding the origin of eukaryotic cells that characterized debate on the subject in the 1970s. In his exposition of a direct filiation hypothesis in 1969, Allsopp dismissed the idea of a symbiotic origin for centrioles and flagella, denying that there was any evidence to support the hypothesis at all (Allsopp, 1969). On the other hand, properties shared between bacteria and mitochondria and chloroplasts, he argued, were expected on the autogenous hypothesis of their origin; they were simply retained ancestral features of the prokaryote that gave rise to eukaryotes. He claimed that this

hypothesis was far simpler than an endosymbiotic scenario, and, on the basis of Ockham's razor, it should be preferred. Raff & Mahler also criticized SET theory by advancing two general arguments against it. Firstly, they suggested that fundamental properties of eukaryotic cells argued that they were originally aerobic: the presence of superoxide dismutase in their cytosol and the presence of sterols and unsaturated fatty-acids in their membranes that are produced via oxygen-dependent biochemistry. If protoeukaryotes were aerobic, they argued, what selective advantage would be gained by taking on aerobic endosymbionts? Secondly, they called on evidence that mitochondria lacked the amount of DNA required to produce all of their proteins. Hence, the symbiotic theory required that massive transfer of symbiont DNA to the nucleus must have occurred. A mechanism for integrating this DNA into the nucleus was, for them, difficult to conceive (Raff & Mahler, 1972, 1973).

Nevertheless the popularity of the symbiotic theories increased, and several other defenders began to appear. De Duve argued that even if the eukaryotic ancestor was aerobic, its biochemistry may not have been as efficient as other aerobic bacteria; so the selective advantage of taking one on as an endosymbiont may still have been present (de Duve, 1973). In addition, Raff & Mahler's argument about DNA transfer was weakened somewhat by the fact that several years earlier, Stanier had argued that such transfer was of key importance to the symbiotic theory, as it provided the mechanism by which endosymbiosis could be made permanent (Stanier, 1970). Taylor also found the DNA transfer argument weak, since evidence for gene transfer was already appearing in the literature (Taylor, 1974).

By the early 1970s, the parts of the serial endosymbiosis hypothesis that dealt with the origin of mitochondria and plastids were rapidly becoming the dominant view. In an well balanced review, Taylor discussed the relative merits

of both theories (Taylor, 1974). In particular he discussed at length many cases of endosymbioses where the symbionts retained their autonomy. The early success of the symbiotic theory was largely due the abundant evidence of transitional forms that examples like these afforded. By contrast, few detailed alternative proposals of autogenous origins of these organelles had been elaborated and ones that were, seemed awkward at best (Taylor, 1974). However, in 1975, the debate was refueled by the appearance of a specific alternative autogenous hypothesis for the origin of the eukaryotes conceived by Cavalier-Smith (Cavalier-Smith, 1975). Cavalier-Smith argued that symbiotic theories essentially side-stepped the problem of eukaryote origins since they did not explain how fundamental features like the nucleus, endoplasmic reticulum and the cytoskeleton evolved. His view was reinforced by the fact that most advocates of SET did not believe in Margulis' symbiotic hypothesis for the origin of flagella and centrioles (Stanier, 1970, Taylor, 1974). Cavalier-Smith postulated a series of steps whereby a cyanobacterium with aerobic metabolism could be converted into a primitive eukaryote pre-alga. Following Stanier, he suggested that the selective force that drove the first steps in the origin of eukaryotes was the ability to phagocytose prey. Loss of the cell wall and the origin of exocytosis preadapted the ancestral cyanobacterium to become a phagotroph. The origin of an actomyosin endoskeleton aided in stabilizing this wall-less organism and also functioned in controlling the membrane budding and fusion processes that are characteristic of eukaryotic cells. Cavalier-Smith's theory also provided detailed mechanisms by which internal membranes evolved from thylakoids and how, eventually, these membranes were arranged to form nuclei, mitochondria and chloroplasts. The origin and distribution of DNA between nuclei and the cytoplasmic organelles were also accounted for in this hypothesis. This paper, together with a similar theory outlined by Taylor conceived at the same time as

Cavalier-Smith's proposal but published a year later (Taylor, 1976), revived the credibility of an autogenous view. In his review, Taylor distinguished several views including a fully autogenous eukaryotic origin, a partially xenologous hypothesis and a fully symbiotic origin of the eukaryotic condition. Even though he had proposed a detailed fully autogenous origin hypothesis, Taylor's discussion made it clear that he felt that a partially xenologous hypothesis was the best supported by evidence at the time (Taylor, 1976).

A bit of this one and a bit of that one.

Ultimately, Taylor was right and the partially xenologous hypothesis eventually emerged as the winner. But, unlike the debate in the 1960s and early 70s, it was data and not *a priori* feasibility arguments that won the case for the endosymbiotic origin of mitochondria and plastids, and a lack of comparable data that rendered the symbiotic origin of flagella untenable. These data were in the form of molecular evidence for endosymbiosis that started to emerge in the mid-70s. In 1975, John & Whatley had argued on biochemical grounds that the purple non-sulfur bacterium *Paracoccus denitrificans* strongly resembled mitochondria. Moreover, this organism even possessed membrane invaginations that resembled mitochondrial cristae. Schwartz and Dayhoff later confirmed these observations by producing molecular phylogenies of 5S rRNA and cytochrome c that showed a close affinity between the purple non-sulfur bacteria and mitochondria to the exclusion of other bacterial groups. At the same time they showed that plastids were closely allied with cyanobacteria (Schwartz & Dayhoff, 1978). Around the same time, oligonucleotide cataloguing of small subunit ribosomal RNAs (SSU rRNA) began to yield similar results. Bonen and Doolittle showed that the oligonucleotide catalogues of the red alga *Porphyridium cruentum* and the green alga, *Euglena gracilis* showed significantly more similarity

to cyanobacteria than to other bacteria (Bonen & Doolittle, 1975, 1976). Similarly, catalogues of SSU rRNA of wheat mitochondria revealed profound similarity to eubacterial catalogues, suggesting their phylogenetic affinity to the latter group (Bonen *et al.*, 1977, Cunningham *et al.*, 1977). Later studies of nucleotide modification and secondary structural features confirmed the essentially eubacterial nature of wheat mitochondrial rRNA (see references in Gray, 1983).

These data along with the development of Woese's ideas regarding the ancient divergence of eukaryotes, Eubacteria, and Archaeobacteria indicated that the genes of organelles shared a more recent common ancestry with homologs in eubacteria (Woese, 1977). As full-length 16S rRNA sequences appeared in the 1980s, molecular phylogenies converted the endosymbiotic hypothesis for the origin of mitochondria and plastids into accepted theory (Gray *et al.*, 1984, Yang *et al.*, 1985, and reviewed by Gray, 1992).

The nature of the host and the most primitive eukaryotes.

Advocates of the autogenous theory in the 1970s had settled on an origin of eukaryotes from a photosynthetic, probably cyanobacterial stock. As Taylor explained, this was because the shared features of photosynthetic eukaryotes and cyanobacteria were too complex to have evolved independently along parallel lines (Taylor, 1976). Thus, according to these views, the most primitive eukaryotes were algae. Of these, the rhodophytes were most often suggested as the most primitive eukaryotes since their plastids shared with cyanobacteria characters such as phycobilisomes. Conversely, they were distinguished from higher eukaryotes by possessing a simple cell structure that lacked flagella (Cavalier-Smith, 1975, Taylor, 1976).

However, for the advocates of the endosymbiotic theory, the nature of the eukaryotic ancestor was less clear. As mentioned above, Margulis suggested an

anaerobic prokaryote was this ancestor. It played host first for a mitochondrial symbiont and next for a flagellar symbiont. The combination of the three cell types resulted in the emergence of the first true eukaryotes. In her view, amoebflagellates such as *Tetramitus* were representatives of these earliest eukaryotes because of their aberrant mitosis and a lack of connection between their basal bodies and the nucleus (Sagan, 1967). But Stanier and Taylor were both unhappy with this scenario since they did not favour the flagellar endosymbiosis part of the story (Stanier, 1970, Taylor, 1974). Moreover, Stanier argued strongly for the host already having developed many of the properties of a eukaryotic cell prior to endosymbiosis and preferred a version of the theory whereby plastids were acquired earlier than mitochondria (Stanier, 1970). Stanier and Taylor both suggested dinoflagellates as candidates for the most primitive eukaryotes, since they possessed a strange, possibly primitive mitosis and they appeared to lack histones (Stanier, 1970, Taylor, 1974).

The idea of endosymbiotic origin of plastids and mitochondria led naturally to musings of whether transitional stages in the process were still represented amongst living organisms. In her 1970 book, Margulis briefly mentioned that one protist, the giant amoeba *Pelomyxa palustris*, was known to lack mitochondria and that this might represent such a transitional stage. She cited previous reports that this organism was both amitotic and ameiotic as evidence for her theory that these processes independently evolved in multiple lineages. However, the possibility that this organism represented a transitional stage in the evolution of mitochondria was quickly dismissed because microtubules had been found in *Pelomyxa*, suggesting that the motility symbiosis had already taken place. Since her ordering of events implied that mitochondria evolved prior to the motility symbiosis, she argued that this organism probably lost its mitochondria secondarily (Margulis, 1970). However, the protozoologists

Bovee and Jahn, strongly disagreed. Like others, they were not convinced of a symbiotic origin for flagella, and they strongly believed that *Pelomyxa* was a transitional, primitively amitochondrial organism (Bovee & Jahn, 1973). For the next decade, this view was echoed by many, and figured prominently in John & F. R. Whatley's discussion of mitochondrial origins in 1975 (John & Whatley, 1975). In collaboration with her husband, Jean Whatley became one of the most ardent advocates of the primitive nature of *Pelomyxa*, and made a detailed study of its cell structure and its abundant bacterial symbionts with a view to gaining insight into its primitive nature (Whatley, 1976). Like many others, she was misled by early reports that nuclear division in this organism was by budding, not mitosis and that it lacked flagella (Daniels & Breyer, 1967) and she regarded both conditions as primitive. More recent studies have shown that these early views were incorrect and mitosis and non-motile flagella do in fact exist in this organism (Griffin, 1988).

The rise of the Archezoa hypothesis.

After proposing several changes to the dominant endosymbiotic hypotheses of the time, Cavalier-Smith began to endorse this origin for mitochondria and plastids in the early 1980s. While he accepted the growing body of evidence for symbiotic origins of these organelles, he remained fundamentally opposed to Margulis' hypothesis of a symbiotic origin for the eukaryotic cell, maintaining his earlier view that the evolution of the nucleus, endomembrane system and cytoskeleton (including flagella) occurred purely autogenously (Cavalier-Smith, 1983a).

In 1983, his abandonment of an autogenous origin for mitochondria and plastids led Cavalier-Smith to propose a protist sub-kingdom, the Archezoa, containing the earliest eukaryotes that he argued may primitively lack both

organelles (Cavalier-Smith, 1983a, 1983b). Building on the earlier proposals of transitional forms, Cavalier-Smith suggested that there were four protist phyla that could lack mitochondria and plastids primitively: the Archamoebae (a group of mostly flagellated amoebae, including *Pelomyxa*), the Microsporidia (a group of obligate intracellular parasites), the Metamonada (containing three flagellate groups: diplomonads, retortamonads and oxymonads) and the Parabasalia (containing the trichomonads and hypermastigotes). It is curious, that with the exception of *Pelomyxa*, none of these organisms were previously proposed to represent transitional forms, since their amitochondrial, aplastidic nature had been documented by the early 70s. Yet only three years after Cavalier-Smith's proposal, the first ribosomal RNA sequence appeared for a member of one of these groups, the microsporidian parasite *Vairimorpha necatrix*, revealing that it possessed a fused 5.8S and 28S large subunit ribosomal RNA species (LSU rRNA), a feature it shared with prokaryotes to the exclusion of all other eukaryotic groups (Vossbrinck & Woese, 1986). Later, the sequence of the small subunit rRNA (SSU or 18S rRNA) species of this organism was completed and a phylogenetic tree of existing eukaryotic and prokaryotic sequences placed *V. necatrix* as the deepest in the eukaryotic lineage (Vossbrinck *et al.*, 1987). Soon after, Sogin *et al.* (Sogin *et al.*, 1989) sequenced the ribosomal RNA of the diplomonad parasite, *Giardia lamblia*, and their phylogenetic analysis showed that this organism branched even earlier than *V. necatrix*. A member of a third amitochondrial group, the trichomonads, was represented on the SSU rRNA tree within the same year (Sogin, 1989), and, once again, the organism branched deeply. All of these results were consistent with Cavalier-Smith's original contention that the Microsporidia, diplomonads, and trichomonads at least, were primitively amitochondrial eukaryotes. As ribosomal RNA sequences accumulated, others began to discuss the possibility that these multiple

independent early-branching protist lineages could be primitively amitochondrial (Embley *et al.*, 1994, Schlegel, 1994, Margulis, 1996). Some, like Patterson and Sogin, recognized this possibility by coining an alternative name, the Hypochondria, for a primitively amitochondrial assemblage (Patterson & Sogin, 1992).

Since his original proposal, Cavalier-Smith's own views have changed frequently on the exact composition of the Archezoa. For instance, in 1987 he argued that trichomonads should be excluded since they possessed hydrogenosomes, mysterious energy-generating organelles that he believed most probably evolved from mitochondria (Cavalier-Smith, 1987a). Later, *Entamoeba histolytica* was removed from the Archamoebae because of rRNA evidence that suggested that it had secondarily lost mitochondria (Cavalier-Smith, 1993a). In his most recent writings, Cavalier-Smith has abandoned the Archamoebae altogether in response to new rRNA evidence that none of these amoebae represent deeply diverging organisms (Cavalier-Smith & Chao, 1996, see Chapter 1). By contrast, Patterson's views much more closely resemble the original Archezoa hypothesis. He maintains that all of the amitochondriate groups originally named by Cavalier-Smith primitively lack the organelle (Patterson, 1994).

Testing hypotheses of cell and molecular evolution.

From the foregoing discussions it should be clear that over the past four decades, hypotheses about early cell evolution have vastly outnumbered rigorous tests of their implications. The speculative tone of much of what was written in this period was largely due to the inability to collect data that would bear directly on the historical sequence of events that took place 1-2 billion years ago during the origin of eukaryotes. The technical revolutions in molecular

biology, however, started to shift the emphasis of the field from *a priori* reasoning, to experimental testing in the late 1970s. The subsequent confirmation of the endosymbiotic hypothesis of plastid and mitochondrial origins during the next decade was surely one of the first and most impressive demonstrations of the power of molecular phylogenetics.

The central aim of this thesis is to test some the implications of some of the contemporary theories about early eukaryote cell and genome evolution. Reconstruction of early eukaryotic phylogeny and reconstruction of the early events in evolution of the eukaryote cell are the complementary twin themes of this work.

The Archezoa hypothesis described above makes two related claims: the first is that several amitochondrial protist groups constitute the most deeply branching eukaryotic lineages and the second is the claim that these lineages diverged prior to the endosymbiotic origin of mitochondria. In Chapter 1, I attempt to develop two protein-coding genes, elongation factor 1alpha and β -tubulin as molecular phylogenetic markers to test the first claim. The results of this study suggest that only two of the three deeply branching groups in rRNA trees, diplomonads and trichomonads, are confirmed to be deeply branching in the trees of these proteins. The third, the Microsporidia, appears to belong much higher in the eukaryotic tree as a sister group to Fungi, suggesting the secondary loss of mitochondria in this group. Furthermore, these protein trees, in contrast to rRNA trees, provide evidence that after the early divergences described above, a deep split occurs in eukaryote phylogeny to produce two protist/multicell superclusters. This alternative view of eukaryotic evolution is supported by other molecular and morphological evidence.

Chapter 2 deals with the origin and evolution of the introns in eukaryotic genes, a subject that has engendered much debate among molecular biologists

over the last two decades. Sequences of two genes, glyceraldehyde-3-phosphate dehydrogenase and triosephosphate isomerase, were obtained from a variety of putatively early-branching organisms to reveal whether they contained introns. In addition, I develop an objective method that relies on likelihood models of intron evolution to reconstruct the history of intron evolution in these two gene families and show that the introns in them have evolved relatively recently in eukaryotic lineages.

The study in Chapter 3 is the result of a collaborative effort between this author and Graham Clark to address the second claim of the Archezoa hypothesis in relation to one of the amitochondriate protist groups, the trichomonads. This second claim can be tested: if an organism has diverged after the endosymbiotic origin of mitochondria, relics of the endosymbiosis may persist. We searched for, and found, one such relic: a mitochondrial-like chaperonin 60 gene in the trichomonad *Trichomonas vaginalis*, indicating that mitochondrial endosymbiosis took place before the divergence of this organism.

Finally, in Chapter 4 of this thesis, studies of iron superoxide dismutase genes in deeply branching eukaryotes are briefly described. The results obtained indicate that homologs of this gene were likely acquired by the eukaryotic lineage from Proteobacteria through one or several events of lateral transfer. If this event was the origin of mitochondria, then the endosymbiotic origin of this organelle may have occurred even earlier, prior to the divergence of diplomonads.

Materials and Methods

Organism culture.

Both *Giardia lamblia*, strain WB (ATCC#30957) and diplomonad 50380 (also called *Hexamita sp.*, ATCC#50380) cultures obtained from the ATCC were grown in 15ml glass culture tubes in Keister's modified TY-I-S33 medium (Keister, 1983) supplemented with 250 µg/ml streptomycin and 165 µg/ml penicillin. *Giardia lamblia* trophozoites were incubated at 37°C while *Hexamita sp.* were incubated at 15°C. After significant growth was observed a small portion of the culture was used for subculturing into fresh media while the remainder was harvested by centrifugation and frozen for DNA extraction.

A mixed culture of *Trepomonas agilis* and mixed bacteria (agnotobiotic) was obtained from T. Cavalier-Smith's laboratory. The cultures were maintained at 25°C in 50 ml plastic culture dishes in 30 ml of TGYM-9 media (recipe taken from the 1991 ATCC catalogue). After significant flagellate growth was observed under the light microscope the cultures were centrifuged at low speed to preferentially pellet the large eukaryotic cells. Several rounds of low speed spins were performed with resuspension in PBS buffer. The washed cell pellet was then frozen.

A culture of *Acrasis rosea* (strain T-235) was obtained from F. Spiegel (University of Arkansas). *Acrasis* cells were maintained on 1.5% agar plates supplemented with 0.01% malt extract and 0.01% yeast extract with the yeast *Rhodotorula mucilaginosa* as a food source. Liquid media, consisting of 0.05% yeast extract, 0.25% proteose peptone, 0.1% glucose buffered with 3 mM K₂HPO₄/16 mM KH₂PO₄, was inoculated with a small block of solid media containing the yeast and amoebae and flasks were shaken slowly for 4-5 days at room temperature until significant *Acrasis* growth was observed under the light

microscope. Cells were harvested by centrifugation and frozen for DNA extraction.

DNA sources and extraction procedures.

Diplomonads. Cell pellets were resuspended into lysis buffer consisting of 0.5% SDS, 300 µg/ml proteinase K, 0.1 M NaCl and 1 mM EDTA. This mixture was incubated at 50°C for 1 hour to allow digestion and lysis to occur. The mixture was then gently extracted (to avoid extensive shearing) once with an equal volume of tris-buffered phenol (pH 8.0), a second time with phenol/chloroform/iso-amyl alcohol (in a 25:24:1 ratio) and a third time with chloroform/iso-amyl alcohol (24:1 ratio). The supernatant was mixed with 2 volumes of ethanol and centrifuged at 12,000g for 20 minutes. The nucleic acid pellet was washed once with 70% ethanol, dried, then resuspended in sterile deionized water.

***Acrasis rosea*.** The mixture of *Acrasis* and yeast cells was resuspended in lysis solution containing 200 mg/ml proteinase K and 1% SDS and was incubated at 50°C for 1 hr. This treatment selectively disrupts the acrasid amoebae leaving the hard chitinous *Rhodotorula* cells intact. The yeast cells were subsequently removed by centrifugation and DNA was purified from the supernatant by extraction with phenol, phenol/chloroform/isoamyl alcohol and chloroform as described above. The nucleic acid in the supernatant was then precipitated with 2 volumes of ethanol and 0.1 volumes of 3M sodium acetate (pH 5.0), centrifuged and washed with 70% ethanol. The DNA prepared by this method was extracted further to remove carbohydrates as described below.

***Nosema locustae*.** Spores from this organism were obtained from several sources. One millilitre of frozen washed spores was obtained from the ATCC (ATCC#30860). Another 10ml (1×10^{11} spores) of washed spores were obtained

from L. Mearril at the biocontrol company M&R Durango in Colorado. DNA was extracted from *Nosema* spores by grinding them with a mortar and pestle with liquid nitrogen. Several millilitres of a buffer containing 3% SDS, 50 mM EDTA and 50 mM tris-HCl pH 7.5 were added during grinding. Periodically, a small sample of the mixture was examined under the light microscope to determine the percent of spores that had been disrupted. After approximately 50-70% of spores appeared disrupted, the entire mixture was extracted 3 times with phenol, phenol/chloroform/iso-amyl alcohol and chloroform respectively and DNA was precipitated as described above. DNA prepared from *Nosema* using this technique typically did not digest with restriction enzymes. This was probably due to the presence of co-precipitating carbohydrates, which were removed as described below.

Purification of genomic DNA from carbohydrates.

This procedure was developed by Clark (C. G. Clark, pers. comm.) for purification of DNA from carbohydrate-rich protist cells. It involves taking the DNA prepared as described above and mixing it to a final concentration of 0.7 M NaCl and 1% CTAB. The mixture is incubated at 65°C for 30 minutes then extracted twice with an equal volume of chloroform. The carbohydrate/CTAB complex forms an insoluble layer between organic and aqueous layers during extraction. The supernatant was carefully removed and DNA was precipitated with 2 volumes of ethanol, 0.1 volume sodium acetate, pH 5.0. The pellet obtained from centrifugation was washed twice with 70% ethanol and resuspended in sterile deionized water.

Genomic DNAs received as gifts.

Heteroloboseans. Genomic DNA from axenic cultures of two different strains of *Naegleria andersoni* were obtained as gifts. Dr. S. Kilvington (Bath Public Health Service, Bath) generously provided DNA from *N. andersoni* (strain PPFMB-6) that was used in the isolation of a partial glyceraldehyde-3-phosphate dehydrogenase homolog from this organism. Dr. J. de Jonckheere (Instituut voor Hygiene en Epidemiologie, Brussels) provided DNA from the A-2 strain of this species which was used in the isolation of an elongation factor 1-alpha (EF-1 α) homolog. Dr. G. Hinkle (Marine Biological Laboratory, Woods Hole, MA) provided genomic DNA from a culture of *Tetramitus rostratus* (ATCC# 30216) grown with *Klebsiella pneumoniae* as the food organism.

Trichomonads. Dr. M Müller (Rockefeller University, New York) kindly provided genomic DNA from the trichomonads, *Trichomonas vaginalis* (strain NIH-C1, ATCC# 30001) and *Tritrichomonas foetus* (ATCC#30924) and Dr. P. Johnson (UCLA, Los Angeles) generously provided a cDNA library for *T. vaginalis* used in the isolation of EF-1 α and chaperonin 60 cDNAs.

Diplomonads. Dr. P Keeling provided DNA extracted from the axenic culture of diplomonad 50380 described above.

Microsporidia and Entamoebids. Dr. G. Clark generously provided DNA from the microsporidian *Encephalitozoon hellem* (strain CDC:0291:V213) grown in cell culture and the entamoebid, *Entamoeba histolytica* (strain HM-1:IMSS) grown in axenic culture.

α -Proteobacteria. DNA from *Agrobacterium tumefaciens* (strain 19553) and *Rhizobium etli* (strain CFN42) used in the isolation of iron-manganese superoxide dismutase homologs was provided by Dr. P. Keeling.

PCR primer design.

Regions upon which degenerate PCR primers were based were selected to meet the following criteria. Blocks of amino acid residues in the protein that showed conservation over a great variety of phylogenetically distinct groups and located at the extreme N- and C-termini of the protein were used as the basis for primer sequence design. These regions were selected to avoid the amino acids leucine, serine and arginine, all of which possess 6 codons and cause the primer to be more degenerate. Generally, the four codons at the 3' end of a primer were made entirely degenerate with respect to the genetic code with the exception of the last (closest to the 3' end) codon in N-terminal primers, which was truncated to two bases to avoid degeneracy. Codons greater than four codons away from the 3' end of the primer were selected to balance the G+C composition of the primer.

Elongation factor 1 α . Primers for EF-1 α were based on those described by Baldauf and Palmer (Baldauf & Palmer, 1993) and were directed against the following regions. The 5' primer was based on the sequence VIGHVD (primer 1F) and was used in conjunction with several 3' primers directed against the regions GDNV (primer 7R), DMRQT (primer 8R) and QTVAVGI (primer 10R). These had the following sequences 1F: 5'-CGAGGATCCGTTATTGGNCA YGTNGA-3'; 7R: 5'-ACGTTGGATCCAACRTTRTCNCC-3'; 8R: 5'-GGTCGCGACAGTYTGNCKCATRTC-3' and 10R: 5'-GTCTTAAGGCTGTCGNTGN CARAC-3'. Primers 8R and 10R were designed by S. Baldauf and given to the author as a gift.

β -tubulin. Primers for both β -tubulin and α -tubulin were designed by the author and only those for β -tubulin were used in this study. The N-terminal β -tubulin primer, BtubA was based on the amino sequence GQCGNQ while the C-terminal primer, BtubB was directed against the sequence MDEMEFT and had the

sequences: 5'-TCCTGCAGGNCARTGYGGNAAYCA-3' and 5'-TCCTCGAGTRAAYTCCATYTCRRTCCAT-3' respectively.

Glyceraldehyde-3-phosphate dehydrogenase. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) N- and C-terminal primers, GAPN and GAPC were provided as a gift by M.W. Smith and R.F. Doolittle. These primers were directed against the conserved amino acid sequences NGFGRI and WYDNE found at the N- and C-termini of most GAPDH genes and had the sequences: 5'-GAGAGAGCTCRAYGGNTTYGGNMGNAT-3' and 5'-GAGAGAGCTCWYTCRTRTRTCRTACCA-3' respectively.

Triosephosphate isomerase. Primers TF-1 and TF-2 were directed against two different variants of an N-terminal motif in the triosephosphate isomerase (TPI) protein. Primer TF-1 directed at VGGNWK had the sequence 5'-ACGTCTCGAGTTCGGTGGNAAYTGGAA-3', TF-2 was directed at VGGNFK and had the sequence: 5'-ACGTCTCGAGTTCGGTGGNAAYTTYAA-3' while TR-1 was directed against VGGASL, and had the sequence: 5'-ATCTCTAGAAGTGATGCNCCNCCNAC-3'.

Chaperonin 60. The isolation of a chaperonin 60 homolog from *T. vaginalis* was carried out in collaboration with Dr. Graham Clark. Clark used the following PCR primers to isolate a partial cpn60 homolog from this organism. Primer HSP5.4 was based on the amino acid sequence PKVTKDGVTV and had the sequence 5'-CCAAAARTTACWAAAGATGGAGTTACWGTT-3' while primer TvHSP3.1, based on GGVAVIKV had the sequence 5'-CCRACCTTGATRACAGCRACRCCRCC-3'.

Iron-Manganese superoxide dismutase. An alignment of Fe-Mn SOD amino acid sequences revealed two good candidate regions against which primers were designed. The 5' primer SODF-1 was based on LEYHHD/GKHH, and had the sequence 5'-CTCGAATACCACCAYGRYAAGCAYCA-3' and the 3' primer

SODR-1 was designed against DVWEHAYYI and had the sequence 5'-GATGTAGTAAGCRTGYTCCCARACRTC-3'. These primers, although based on conserved sequences in Fe-SOD genes, also amplified Mn homologs in some of the DNAs used.

DNA amplification and cloning of amplified fragments.

Amplification of DNA was carried out using standard methods. Typically, 10-100ng of genomic DNA was used as a template in 100 μ l reactions with 1 μ M of each primer, 0.5 units of Taq polymerase, and buffered with 10mM tris-HCl pH 8.3, 1.5mM MgCl₂, 50mM KCl and 0.001% gelatin. These reactions were covered with 100 μ l of sterile mineral oil prior to temperature cycling. Cycling consisted of 92° C for 1 minute (denaturation), 45-55° C for 1 minute (annealing) and 72° C for 1-2 minutes (elongation) repeated 35 times with a final elongation step of 72° C for 10 minutes. Usually 20-50% of a single reaction containing amplification products was run on an agarose gel (ranging from 0.8-1.2% agarose depending on the predicted size of the product) and DNA was extracted from gel slices containing DNA of the appropriate molecular weight using the Prep-a-gene kit (BIO-RAD). Amplified fragments were ligated into the pCRII T-tailed vector using the manufacturer's protocol (Invitrogen). Ligations were precipitated and washed with ethanol (as described above) and resuspended in sterile deionized water for transformation into *Escherichia coli*.

Library construction.

A *Nosema locustae* genomic DNA library was constructed in collaboration with N. Fast and C. Richardson. The lambda ZAP Express vector, purchased from the Stratagene company, was pre-cut with BamHI and dephosphorylated with calf alkaline phosphatase. Our strategy in constructing the library was to

partially digest *Nosema* DNA with the four-cutter *Sau3AI* and size select fragments in the 6-10 kilobase (kb) range to ligate into the lambda ZAP Express arms. Total genomic *Nosema locustae* DNA was isolated and purified as described above and several test digests with *Sau3AI* were performed to find a digestion time that optimized the amount of cut DNA in the 6-10 kb range. For the library, approximately 6 µg of *Nosema* DNA was cut with 10 units of *Sau3AI* at 37° C for 5 minutes under standard buffer conditions. The digest was run on an agarose gel and fragments between 6-10 kb in size were cut out of the gel. This DNA was isolated from the gel slice using the Prep-a-gene kit (BIO-RAD). After test ligations, packaging and titring were completed, it was decided that all of the remaining DNA isolated from the gel slice would be used in a single ligation with the lambda ZAP express arms. Following the manufacturer's protocol a ligation was set up with the insert DNA described above, 1.0 µg of the lambda ZAP Express arms, 1.0 mM ATP, 0.5 µl of T4 DNA ligase (New England Biolabs) buffered with 50 mM tris-HCl, 7 mM MgCl₂, 1 mM dithiothreitol in a total volume of 5.0 µl. The ligation was incubated overnight at 15° C and then for 1 day at 4° C. The ligation was then packaged into lambda phage heads with Stratagene's Gigapack II packaging extract following the manufacturer's protocol exactly. After packaging, titring of the library revealed that the primary library contained approximately 8.3×10^4 recombinants. The number of recombinants lacking inserts (identified by blue plaques on titring plates containing X-GAL and IPTG) was less than 1%. Assuming a genome size of 6 megabases (mb) for *Nosema locustae* (D. J. Streett, pers. comm.) and an average insert size of 7 kb in the library, this constitutes approximately 97x genome coverage. The entire library was then amplified (using manufacturer's protocols) to a final titre of 3.0×10^8 plaque forming units (pfu) in a total volume of 30 ml.

Library screening.

For both the *T. vaginalis* lambda ZAP II cDNA library and the *N. locustae* lambda ZAP Express genomic DNA library, essentially the same screening procedure was employed. This protocol is described in detail in the manufacturer's (Stratagene) protocols. A brief explanation follows.

Mid-log-phase cultures of *E. coli* strain XL-1Blue were grown in LB supplemented with 10 mM MgSO₄ and 0.2% maltose. These were pelleted and resuspended in 0.5 volumes of 10 mM MgSO₄. Cells were infected by mixing diluted phage with 300-600 µl of cells and incubated at 37° C for 15 minutes. Several millilitres (the exact volume depended on the size of the plates) of melted NZY top-agarose at 50° C were then mixed with the infected cells and plated on NZY agar plates. The plates were incubated overnight at 37° C. The next day, the plates were overlaid with nylon membranes (Colony/Plaque Screen membranes, DuPont). The membranes were then treated with 0.5 N NaOH to denature the phage, neutralized with 1 M tris-HCl pH 7.5 and air-dried to fix the DNA on. To ensure complete fixation, the dry membranes were also irradiated with UV light in a Stratalinker (Stratagene) under standard conditions. Hybridization with the radioactive probe was carried out under exactly the same conditions as for Southern blotting described below.

To obtain hybridizing clones from the *T. vaginalis* cDNA library (for both EF-1 α and cpn60 genes), approximately 10,000 plaques were screened on the assumption that the genes were highly expressed and would be well represented in the cDNA population. To obtain a *N. locustae* full-length EF-1 α homolog, approximately 40,000 plaques from the genomic library were screened, corresponding to roughly 47x genome coverage. Isolating single plaques required secondary screening. Plasmids containing the inserts of interest were obtained from lambda ZAP clones using the *in vivo* excision method (Stratagene).

Excision from the lambda ZAP II vector required the use of the R408 helper phage while excising plasmids from lambda ZAP Express employed the Exassist helper phage and the XL0LR strain of *E. coli*. In both cases, the manufacturer's *in vivo* excision protocols were followed exactly.

Southern transfer and hybridization.

Approximately 1-3 μg of restricted genomic DNA (from the organism of interest) was run on an agarose gel for Southern transfer. Before blotting, the gel was treated with 0.25 N HCl to partially nick and depurinate the DNA followed by 0.4 N NaOH/0.6 M NaCl to cleave the depurinated sites and finally neutralized in 1.5 M NaCl/0.5 M tris-HCl pH 7.5. The gel was then placed face-down onto wicks immersed in 10x SSC and a nylon membrane (GenescreenPlus, DuPont) cut to the appropriate size was placed on top of the gel. Upon this, several layers of Whatman paper (3MM) were placed, followed by a thick layer of dry paper towels. Southern transfer was allowed to proceed overnight. After transfer was complete, the membrane was removed from the gel, denatured in 0.4 N NaOH, neutralized in 0.2 M tris-HCl pH 7.5/1x SSC and dried. The DNA was fixed onto the membrane by UV crosslinking.

For hybridization, the membrane was wet briefly with 10x SSC, rolled in a nylon mesh, and placed in 10-15 ml of a pre-hybridization fluid containing 1 M NaCl, 1% SDS and 1 g dextran sulfate in a hybridization bottle. Pre-hybridization was carried out for 1-5 hours at 65° C in a Hybaid rotary hybridization oven. During this time, the hybridization probe was labelled with $\alpha\text{-}^{32}\text{P}\text{-dATP}$ using either the random-priming strategy employed in the High-Prime kit (Boehringer-Mannheim) or (for oligos and small DNA fragments) polynucleotidylterminal transferase (Promega). This latter method introduces several radioactive As on to a 3'-projecting end of DNA. After pre-hybridization was completed, 500 μg of

sheared herring sperm DNA was added to the probe solution and the resulting mixture was incubated at 95° C for 5 minutes to denature the DNA. The probe was then added to the bottle containing the pre-hybridized membrane and solution. Hybridization was then carried out at 65° C overnight. After hybridization, the membrane was washed several times with 100 ml 2x SSC at room temperature for 5 minutes, several times with 100 ml of 1% SDS, 2x SSC at 65° C for 30-60 minutes and finally once with 500 ml of 0.1x SSC at room temperature for 30 minutes. Following washing, the membranes were placed in cassettes with X-ray film (Kodak) at -70° C. Film was exposed for varying lengths of times depending on the strength of the hybridization signal.

Transformation of *E. coli*.

DNA was transformed into electrocompetent *E. coli* (strain DH5 α F' or in some cases XL-1Blue) cells by electroporation in 0.1cm cuvettes (BIO-RAD) using a BIO-RAD electroporator set at 1.8kV and 200 Ω resistance. Shocked cells were resuspended in super-broth, grown at 37° C for 30 minutes and then plated on an appropriate medium (see below).

Manipulation of clones in *E. coli*.

Clones containing the pCRII, pBS (Bluescript) or pBK-CMV vectors were grown and maintained at 37° C on 1.5% agar plates or liquid culture containing Luria-Bertani (LB) medium supplemented with 100 μ g/ml ampicillin (for pCRII and pBS) or 50 μ g/ml kanamycin (for pBK-CMV). If blue-white selection was employed, plates were overlaid with 50 μ l of 2% X-GAL in dimethylformamide and 5 μ l of 100 μ M isopropylthiogalactoside (IPTG).

DNA sequencing.

Both manual radioactive sequencing and automated sequencing methods were employed to obtain the sequence of cloned PCR products or isolated genomic or cDNA clones. The full-length sequences of genes were obtained using either a primer-walking strategy or a subcloning strategy. Most sequences were determined on two strands. However, regions sequenced only on a single strand were confirmed by several independent sequences.

Sequence entry and contig assembly.

Sequence management utilized software modules from the Lasergene package (DNASTAR). The program Editseq was used for sequence entry and editing and individual sequences were assembled into contigs using the SeqMan module with default settings. The nature of the sequences was confirmed by BLAST searches against the GenBank database. Nucleotide sequences were translated into amino acid sequences using the DNASTRIDER 1.1 program.

Alignment and Phylogenetic Analysis.

Amino acid sequences of gene families considered in these studies were taken from the most recent release of the GenBank database (release 91). Multiple alignments of these amino acid sequences were constructed with the CLUSTALW (Thompson *et al.*, 1994) program using default parameter settings. Regions in alignments that did not appear optimal were subsequently improved by eye by using a text editor. Once the alignment could not be further improved, it was imported into the PAUP 3.1.1 program (Swofford, 1993). This program allows regions of uncertain alignment to be marked by defining character sets and excluded from the phylogenetic analysis.

In general, phylogenetic analyses were performed on aligned protein datasets using three classes of methods: parsimony, distance and maximum likelihood.

For parsimony analysis, the PAUP 3.1.1 program (Swofford, 1993) was exclusively employed. Regions of the alignment where some sequences were incomplete or contained an alignment gap were coded as missing data. A heuristic search method was employed to find the maximum parsimony tree whereby 10-50 random sequence addition replicates were performed with tree-bisection-reconnection (TBR) branch-swapping. These random addition replicates were employed to avoid getting trapped in regions of local minima in tree space. Bootstrap analysis (Felsenstein, 1985) used 300-500 resamplings of the data. The maximum parsimony tree of each resampling of the dataset was found by a simple sequence addition heuristic search. In all parsimony analyses each character as well as all character state transitions were accorded equal weight.

Distance analyses were performed using programs from the PHYLIP 3.57c package (Felsenstein, 1993). Regions of missing sequence data as well as alignment gaps were scored as missing data. Distance matrices were inferred from the alignment using the PROTDIST program employing the Dayhoff PAM250 matrix as the model of sequence change. This method estimates the maximum likelihood distance between pairs of sequences. Trees were obtained by analyzing distance matrices with the neighbor-joining (NJ) method as implemented in the NEIGHBOR program. Bootstrap resamplings were generated by the SEQBOOT program and analysed as above. The bootstrap majority-rule consensus tree was constructed using the CONSENSE program. Typically 300-500 bootstrap replicates were performed depending on the size of the dataset.

To perform maximum likelihood estimation, the versions 2.2 and 2.3 of the PROTML program from the MOLPHY package were used (Adachi & Hasegawa,

1992). In all analyses, the maximum likelihood amino acid transition probability matrix employed was the Jones, Taylor and Thornton model adjusted to accommodate amino acid frequencies observed in the dataset (the JTT-F model) since it has been shown to perform well for several real datasets (Hashimoto & Hasegawa, 1996). Since the datasets in this study typically included many sequences, an exhaustive search strategy (i.e. evaluating the likelihood of all possible topologies) was not feasible due to the computationally intensive nature of the likelihood calculations. Thus several heuristic strategies were developed to estimate a maximum likelihood tree for many sequences. The first strategy, used in Chapter 3, was to search exhaustively for the maximum likelihood tree containing several constrained subtrees. The semi-constrained portions of the trees were developed from subtrees found in the distance and parsimony trees. A second strategy uses the heuristic searching procedure called quick-add OTU search followed by improvement of the best topology by local rearrangements (the latter option is only available in PROTML version 2.3). This strategy was used to find maximum likelihood trees in the analysis of EF-1 α homologs in Chapter 1. Both of these strategies likely do not find the globally optimal topology but rather find a local maximum likelihood topology. The second strategy may be superior to the first under conditions where parsimony and distance methods fail to reconstruct the correct phylogeny, such as extreme inequalities in the rates of evolution in different lineages. Bootstrap values for nodes in question were estimated using the resampling estimated log likelihood (RELL) procedure (Hasegawa & Kishino, 1994) implemented in the PROTML program with 10,000 replications (Adachi & Hasegawa, 1992). The method developed by Kishino and Hasegawa (Kishino & Hasegawa, 1989) was used to evaluate the standard error of the difference in ln likelihood between alternative topologies. This allows one to test whether tree topologies with higher likelihood

are significantly preferred over other lower likelihood alternatives. For these studies, differences of log likelihood greater than 1.96 standard errors (corresponding to a 95% confidence interval) were considered significant (Kishino & Hasegawa, 1989).

Intron maximum likelihood analysis.

The evaluation of the most parsimonious reconstructions of introns given a tree of the genes was carried out using PAUP 3.1.1 (Swofford, 1993).

In order to analyze intron position data using a maximum likelihood (ML) method, version 4.0 of the PHYLIP program DNAML was employed. The use of this version of the program was critical since it incorporates a method for maximum likelihood reconstruction of ancestral sequences at internal nodes in a phylogenetic tree. The ML model of DNA sequence change employed by Felsenstein in DNAML has recently been described in detail (Felsenstein & Churchill, 1996). The model is quite complex, allowing for separate rates for transitions and transversions as well as allowing one to adjust the transition probabilities to account for different equilibrium frequencies of nucleotides in the dataset.

In order to evaluate intron data with DNAML 4.0, the data were coded in the following manner. Every pair of neighbouring nucleotides in a protein-coding gene can potentially be interrupted by an intron. One can thus represent the intron sequence of a gene as a series of positions that are either interrupted by an intron or not. For instance, a gene containing 5 introns and 300 nucleotide positions would have 294 uninterrupted nucleotide pairs and 5 interrupted pairs. This information can be represented by a sequence of 299 positions (each corresponding to a nucleotide pair) each of which is occupied by either a symbol

for uninterrupted pairs or a second symbol for interrupted pairs. This sequence will be henceforth referred to as the intron position sequence.

Two intron models were considered for this analysis. The first model allowed that introns could be both gained and lost, but the rates of intron loss and gain need not be equal. The second model, a special case of the first, allows for intron loss, while intron gain is made very very improbable (effectively impossible). The first model corresponds to both an introns-late model and an introns-early model where a process of intron gain is allowed to operate. The second model corresponds to a strict introns early model where introns are never inserted (referred to as the "hard" introns early model). For a more detailed discussion of these models and their relevance to the intron debate, see the introduction to Chapter 2. Both of these models actually correspond to a special case of the DNAML 4.0 nucleotide model where only two kinds of bases, pyrimidines (Y) and purines (R), are considered (J. Felsenstein, pers. comm.). The intron position sequence described above can thus be written as a series of Y's (for uninterrupted nucleotide pairs) and R's (for interrupted nucleotide pairs). In such a model, there are two kinds of transitions possible: intron gain, represented by Y \rightarrow R changes or intron loss, represented by R \rightarrow Y changes. The relative rate of gain to loss can be adjusted using the user-defined base frequencies option. For instance, if one wishes the model to have a rate of gain which is 9x the rate of loss, then one sets the equilibrium frequencies of R and Y in the model to be 0.9 and 0.1 respectively (this is due to the fact that if this ratio of rates is 9:1, then the frequencies of R's and Y's in the dataset will converge on these proportions at equilibrium).

Testing the alternative intron models.

The intron positions in various homologs of TPI and GAPDH for which full-length genomic sequences were available were compiled and the intron position sequences for each taxon were thus determined. The alignment of these sequences was inferred by superimposing it on the amino acid sequence alignment. For GAPDH and TPI proteins both gene trees, inferred from protein distance analysis (using the PROTDIST program employing the Dayhoff model followed by neighbor-joining analysis using the NEIGHBOR¹ program (see citations above), and user-defined trees were used in the tests of alternative intron models. The user-defined trees were constructed by altering the distance tree topology to more closely match other phylogenies of the species (inferred from small subunit rRNA, EF-1 α and tubulin comparisons) taking into account specific gene-tree anomalies.

To distinguish between the gain and loss model versus the loss-only model described above, DNAML 4.0 was used to calculate the ln likelihood (lnL) of observing the intron sequences given the tree and the model of evolution. A curve was then constructed by plotting lnL versus a series of different user-defined ratios of the rates of gain to loss ranging from extremely low to high values. The peak of this curve is considered to be a maximum likelihood estimate of this ratio and a 95% confidence interval on this estimate is bounded by the ratios having a ln likelihood within $1/2 \chi^2$, with 1 degree of freedom of the maximum ln likelihood (Edwards, 1972, J. Felsenstein, pers. comm.). If the loss-only model were correct, then the maximum likelihood estimate of the ratio of gain to loss would not be significantly different from 0.

In order to distinguish between an introns early view and an introns late view where both intron gain and loss are incorporated into the model, the maximum likelihood reconstruction of ancestral sequences option of DNAML 4.0

was utilized. An introns early view would have it that the common ancestor of prokaryotes and eukaryotes possessed at least a single intron in their GAPDH and TPI genes (see the introduction to Chapter 2). Since neither of the trees of these genes can be rooted to define the node of this common ancestor, it was necessary to examine the two nodes between which the root is likely to fall. By evaluating the reconstructed ancestral sequences at these nodes, it was possible to evaluate whether a node between them contained introns. If either of these neighbouring nodes possessed introns in the reconstructed ancestral sequences, then the introns-early view was considered supported. If neither of them possessed introns, then this view was rejected and the introns-late view was supported. Since the reconstructed ancestral sequences are probability estimates, it was also possible to evaluate the confidence of each position in the ancestral sequence reconstruction.

Chapter 1

INTRODUCTION

In the last decade, phylogenetic analysis of small and large subunit ribosomal RNA sequences from diverse eukaryotic groups has revolutionized our understanding of eukaryotic phylogeny, allowing us to define major monophyletic eukaryotic groups as well as suggesting probable branching orders among these groups. Currently, the two prokaryotic groups, the Archaeobacteria and Eubacteria, root the SSU rRNA tree of eukaryotes such that three amitochondrial protist groups form the deepest branches of the eukaryotic lineage: the diplomonads, the microsporidia and the trichomonads (Sogin, 1991, Leipe *et al.*, 1993, Cavalier-Smith, 1993a, Embley *et al.*, 1994, Cavalier-Smith & Chao, 1996). These trees are most parsimoniously viewed as consistent with Cavalier-Smith's original hypothesis, proposed on the basis of ultrastructure, that these lineages diverged prior to the endosymbiotic origin of mitochondria and thus form a paraphyletic group he calls the Archezoa.

Another group suggested to be primitively amitochondrial, the Archamoebae, do not appear to be deeply-branching in ribosomal RNA trees. Ribosomal RNA sequences from members of this group such as *Entamoeba histolytica*, *Phreatamoeba balamuthi* and most recently *Pelomyxa* sp. all appear to diverge later in rRNA trees as independent amitochondriate lineages arising from within the mitochondrial eukaryotes (Sogin, 1991, Hinkle *et al.*, 1994, Morin & Mignot, 1995).

In addition to placing these amitochondrial protist groups, rRNA phylogenetics has yielded candidates for the earliest diverging mitochondrion-containing groups. The Euglenozoa (trypanosomatids, bodonids and euglenoids) and the Heterolobosea (schizopyrenid amoeboid flagellates and acrasid slime moulds), are early-branching lineages in SSU rRNA trees, with the former group

usually occupying the deepest position amongst mitochondriate eukaryotes (Sogin, 1991, Hinkle & Sogin, 1993, for an exception see Cavalier-Smith, 1993a).

As ribosomal RNA data have accumulated, the outline of early eukaryotic phylogeny inferred from these data has been challenged on several fronts. Cavalier-Smith, Patterson and O'Kelly have all developed separate phylogenetic schemes for the early branching order of the eukaryotes by considering both ribosomal RNA trees and ultrastructural data of the relevant groups (Cavalier-Smith, 1993a, O'Kelly, 1993, Patterson, 1994). Each of these schemes differs from the others and from most published ribosomal RNA trees.

A second challenge comes from the fact that different analyses of rRNA with different taxonomic samplings appear to yield different results for the branching order of the deepest groups. For instance, Leipe *et al.* have shown that the relative branching order of diplomonads, trichomonads and microsporidia depends strongly on what prokaryotic taxa are used as outgroups in the analysis (Leipe *et al.*, 1993). They suggested that the extremely high G+C composition of deeply-branching sequences such as *Giardia lamblia* and low G+C composition of others like *Vairimorpha necatrix* coupled with similar biases amongst outgroup taxa may be responsible for this effect. Recently, an analysis by Galtier and Gouy has confirmed that base composition is biasing the early branching order of eukaryotes, and they have proposed a phylogenetic method for dealing with the problem (Galtier & Gouy, 1995).

The existence of alternative phylogenetic hypotheses based on ultrastructure as well as potential lack of resolution and biases influencing the branching order of the SSU rRNA makes it clear that other molecules must be developed as independent estimators of early eukaryote phylogeny. Progress on this front is now being made as several protein-coding genes including elongation factors (Hashimoto & Hasegawa, 1996), DNA-dependent RNA

polymerases (Klenk *et al.*, 1995), V-type ATPases (Gogarten *et al.*, 1996), glyceraldehyde-3-phosphate dehydrogenase (Martin *et al.*, 1993, Rozario *et al.*, 1996, and Roger *et al.*, 1996), α - and β - tubulin (Baldauf & Palmer, 1993, Edlind *et al.*, 1996, Li *et al.*, 1996), actins (Drouin *et al.*, 1995) and hsp70 (Gupta *et al.*, 1994) have recently been utilized to develop global phylogenies of eukaryotes.

Hasegawa and his colleagues have argued that trees of protein-coding genes such as these may be less sensitive to base composition artifacts (Hasegawa & Hashimoto, 1993). As evidence for this, they have shown that homologs of elongation factor 1alpha (EF-1 α) and 2 from organisms with a highly biased nucleotide composition, such as *Giardia lamblia*, do not display a significantly biased amino acid composition (reviewed in Hashimoto & Hasegawa, 1996).

In this study, I have chosen to further these efforts by sequencing homologs of EF-1 α and β -tubulin from representatives of some of the putatively early branching protist taxa. My reasons for choosing these genes are severalfold.

Elongation factor 1 α and β -tubulin families have very diverse phylogenetic representation especially for deeply-branching eukaryotic groups (Hashimoto and Hasegawa, 1996, Edlind *et al.*, 1996). Since adequate species representation has been shown to be important in obtaining reasonable phylogenetic estimates (Lecointre *et al.*, 1993), these datasets seem particularly well suited to the task of reconstructing the early branching order of eukaryotes.

Both EF-1 α and β -tubulin perform essential functions in the cell: EF-1 α delivers aminoacyl-tRNA to the A-site of the ribosome during protein synthesis while β -tubulin is a primary component of microtubules and microtubule-based organelles. As a result, both are very highly conserved proteins and homologs from phylogenetically distant taxa are easily aligned. In addition, each of these

proteins interacts with multiple other cellular factors suggesting that neither is likely to be subject to lateral transfer.

Finally, previous studies of EF-1 α have established that the overall branching order for this gene is similar to SSU rRNA. However, trees based on tubulins show a radically different phylogeny (Edlind *et al.*, 1996 and Li *et al.*, 1996). For example, the microsporidian *Glugea plecoglossi* is placed significantly as the deepest branch of the eukaryotes in the EF-1 α tree (Kamaishi *et al.*, 1996). By contrast, trees of both α - and β - tubulin trees show the microsporidia branching from within the fungi, with significant bootstrap support. This and other significant conflicts between tubulin, elongation factor and SSU rRNA phylogenies are puzzling and demand an explanation.

In this chapter I will attempt to reconcile the differences between EF-1 α and tubulin trees and discuss the kinds of artifacts that may lead to such conflicts in molecular phylogenies. Furthermore, I will try to develop a consensus view of early eukaryotic phylogeny that takes into account both molecular and ultrastructural data.

In order to improve the taxonomic representation in the EF-1 α dataset, I chose to sequence homologs of EF-1 α from members of each of the putatively early-branching groups including two trichomonads, *Trichomonas vaginalis* and *Tritrichomonas foetus*; one free-living diplomonad, *Trepomonas agilis*; one microsporidian, *Nosema locustae* and three heteroloboseans, *Naegleria andersoni*, *Tetramitus rostratus* and *Acrasis rosea*. For β -tubulin, I obtained a single sequence from the microsporidian *Nosema locustae*.

RESULTS

Elongation factor 1 α sequences.

With the exception of *T. vaginalis* and *N. locustae*, libraries were not available for the organisms studied and nearly full-length EF-1 α clones were obtained exclusively using PCR methods. However, due to the fact that only some primer sets worked with some organisms, different strategies were employed for some organisms.

For *Naegleria andersoni*, which failed to amplify an EF-1 α homolog with the 1F/8R primer pair (primer 10R was not used as it was designed after this work was completed), the 1F/7R pair was used and the resulting PCR product (corresponding to approximately the N-terminal two thirds of the gene) was cloned and sequenced. Based on the 3' terminal sequence of this product, an exact-match primer was synthesized and used in conjunction with 8R to isolate the 3' end of the gene. Due to the presence of multiple copies of EF-1 α in this organism (see below), a number of clones corresponding to the 3' end of this gene were sequenced to identify one that matched the 1F/7R product exactly in the region of overlap.

For *Acrasis rosea*, *Tetramitus rostratus* and *Tritrichomonas foetus*, the primer set 1F/10R was used to amplify a fragment containing most of the phylogenetically informative coding sequence of EF-1 α . For *T. rostratus*, two PCR clones were sequenced and differed at 19 synonymous positions and 1 non-synonymous position.

The 1F/7R primer pair was used to isolate a fragment of EF-1 α from the diplomonad *Trepomonas agilis*. Attempts to obtain the 3' end of this gene from this organism using all available primer sets failed.

To obtain the *T. vaginalis* EF-1 α , a fragment of the gene was amplified, cloned and sequenced with the primer pair 1F/7R and used as a homologous

probe to isolate four cDNAs from a lambda ZAP II cDNA library. The excised plasmids of these cDNAs were restricted and the clone containing the largest cDNA was completely sequenced. The sequence of this cDNA contained no start codon and the inferred amino acid sequence of the N-terminus aligned with homologs from other organisms several amino acids downstream from their N-termini. This suggested that the cDNA was truncated downstream of the 5'-UTR as well as the start codon, consistent with reports of other similarly truncated cDNAs isolated from this library (Hrdy & Müller, 1995). The presence of a UAA stop codon indicated that the 3' end of the coding region of the EF-1 α cDNA was intact.

An EF-1 α fragment from *N. locustae* was obtained using the primer set 1F/7R and was cloned and fully sequenced. Repeated attempts to amplify the 3' end of the gene with a variety of primer sets failed. A lambda ZAP Express total genomic DNA library was developed for *N. locustae* partly in order to obtain a full-length copy of an EF-1 α gene from this organism. The 1F/7R fragment was used as a homologous probe to screen this library. Eight hybridizing phage clones were isolated after secondary screening and plasmids corresponding to each were obtained by *in vivo* excision. Sequencing primers, developed from the sequence of the 1F/7R fragment, were used to sequence two different genomic DNA clones to obtain the 5' and 3' terminal regions of the gene. The full length coding sequence was obtained and consisted of 478 codons. Interestingly 12-24 bp upstream of the ATG start codon, there is an element consisting of TTTAAAATTTTTTTTTT. Since there are no other genomic sequences published from microsporidia where upstream and downstream sequence have been determined, it is not known whether this tract represents a conserved sequence element involved in gene regulation. Sequences immediately upstream and downstream of the coding region were compared to sequences in GenBank using

the BLAST algorithm and no significant hits were detected, tentatively suggesting that this elongation factor gene is not organized in an operon.

From an alignment of all of the obtained elongation factor amino acid sequences (Fig. 1.1), it appeared that none of these EF-1 α homologs contained introns.

***Trepomonas agilis* uses an alternative genetic code.**

Translation of the *T. agilis* PCR product with the universal genetic code revealed that the sequence was interrupted by 8 stop codons: all of them were TAA. Recently, the GAPDH gene of this organism was determined (Rozario *et al.*, 1996) and a single TAA codon was identified and tentatively attributed to the use of a variant genetic code in this organism. This explanation is very likely since two other closely related diplomonads, diplomonad 50330 (*Hexamita sp.* of Keeling & Doolittle, 1996, or "*Spironucleus*" sp. of Rozario *et al.*, 1996) and *Hexamita inflata*, appear to possess a variant of the universal genetic code where TAR codons specify the amino acid glutamine (Q) (Keeling & Doolittle, 1996). The sequence of this partial EF-1 α gene strongly supports the view that *T. agilis* also contains this variant genetic code. For the purposes of phylogenetic analyses, all TAA codons in this sequence were translated as glutamine. It is currently unknown whether this organism also uses TAG codons to code for this amino acid as in the other diplomonads.

Multiple copies of EF-1 α in *T. vaginalis* and *N. andersoni*.

For *T. vaginalis*, *N. andersoni* and *N. locustae*, sufficient genomic DNA was available to make Southern blots to evaluate the copy number of the EF-1 α gene in these organisms. To cut the genomic DNAs, restriction enzymes were chosen such that they did not cut the homologs that were sequenced. Figure 1.2 shows

Figure 1.1 Alignment of EF-1 α homologs characterized in this study. This alignment is taken from a full alignment of 79 EF-1 α genes of eukaryotes and archaeobacteria. Species name abbreviations are *T.va*: *Trichomonas vaginalis*, *T.fo*: *Tritrichomonas foetus*; *T.ag*: *Trepomonas agilis*; *N.an*: *Naegleria andersoni*; *A.ro*: *Acrasis rosea*; *T.ro I*: *Tetramitus rostratus I*; *T.ro II*: *Tetramitus rostratus II*; and *N.lo*; *Nosema locustae*. Asterisks (*) indicate stop codons.

T.va VDAGKSTTTGHLIYKCGGLDKRKLAAIEKEAEQLGKSSFKYAFVMSLKAERERGITIDI
T.fo AGKSTTTGHLIYKCGGIDKRKIAQIEKEAEQLGKGSFKYAFVMSLKAERERGITIDI
T.ag NGKSTLTGHLIYKCGGIDARTLEEYEKANELGKGSFKYAWVLDQLKDERERGITINI
N.an AGKSTTTGHLIYKCGGIDKRVEIEKFEKEAAEMGKSSFKYAWVLDKLAERERGITIDI
A.ro SGKSTTTGHLIYKCGGIDKRVEIEKFEKEAADMKGQSFKYAWVMDKLAERERGITIDI
T.ro I SGKSTTTGHLIYKCGGIDKRVEIEKFEKEAAEIGKGSFKYAWVMDKLAEKERGITIDI
T.ro II SGKSTTTGHLIYKCGGIDKRVEIEKFEKEAAEIGKGSFKYAWVMDKLAEKERGITIDI
N.lo MEGKKPNLNVCIIGHVDSGKSTTMGNLAYQLGVFDQRQLTKLKAADSHGKGTFAAYFFDNATAERKRGITIDI

150

T.va SLWKFEGQKFSFTIIDAPGHRDFIKNMITGT'SQADAAILVIDSTLGGFEAGIAEQGQ'TREHALLAFTLGIKQVIV
T.fo SLWKFESPKYFFTIIDAPGHRDFIKNMITGT'SQADAAILVIDSTRGGFEAGIAEQGQ'TREHALLAFTLGIKQIII
T.ag ALWKFETKRFIVTIIIDAPGHRDFIKNMITGT'SQADVAIILVIVASGVGEFEAGISNEGQ'TREHATLANFTLGIKTMIL
N.an ALWKFESKQYVFTIIIDAPGHRDFIKNMITGT'SQADVAIILVVDSTNGGFEAGFKDQGQ'TREHALLAYTLGIKQMIIV
A.ro ALRKFETSKTMFTIIIDAPGHRDFIKNMITGT'SQADAAVLVIDSTTGGFEAGISKDGQ'TREHALLAFTLGIKQMIIV
T.ro I SLWKFQSAKYDFTIIIDAPGHRDFIKNMITGT'SQADVAIILMIDSTTGGFEAGISKDGQ'TREHALLAQTILGVKQMIIV
T.ro II SLWKFQSAKYDFTIIIDAPGHRDFVKNMITGT'SQADVAIILMIDSTTGGFEAGISKDGQ'TREHALLAQTILGVKQMIIV
N.lo TLKEFKLKKFNANIIDCPGHKDFIKNITVIGAAQADVAVALVPSD--FAAATSPKATLKDHIIMSGVMGIKRLII

225

T.va AVNKMDDKTVNYNKARFDEITAEMTRILTGIGYK-----PE-MFRF-VPI SGWAGDNMTEK-SPNMPWYNG---
T.fo AINKVDDSTVNYSQERFNEIKGEMTRVLTNIGFK-----PE-QYKF-VPI SGFKGDNMTEK-SANLGGWNG---
T.ag AINKMDDPQVNYSQARVEEIKTEMQKTLKAIGFK-----HWEEFNY-IPTSGWTGDNIMEK-SPKMPWYNG---
N.an CMNKFDSTSVSYKEDRYNEIKSEVGRYLKGLGFNVDAADKPN-LVQF-VPI SGWTGDNMIEK-TDKMPWYK---
A.ro CTNKMDDKSVQYKEDRYKEIQKEVADYLLKVGYN-----PK-NVPF-VPI SGWAGDNMLEK-STNMPWYK---
T.ro I CLNKFDEKTVNFSQARYDEIHKEVAAYLKKVGYN-----PE-KVPF-IPLSGFQGYNMTERATDKMPWYK---
T.ro II CLNKFDEKTVNFSQARYDEIHKEVAAYLKKVGYN-----PD-KVPF-IPLSGFQGYNMTERATDKMPWYK---
N.lo CVNKMDFEPPEKQKEKFEWIKKEMLFISQRLHPD-----KDIIPISGLKGINIADH-GEKFEWFEWQK

300

T.va -----PYLLEALDSLQPPKRPFDKPLRPLQDVYKINGIGTVPVGRVESGMTKPGMIVNFPSTVIT--
T.fo -----GTLLETLDLTLQPPKRPFDRLRPLPIQDVYKISGIGTVPVGRVESGIMKPGQVVFAPSGIN--
T.ag -----PCLIEAIDGLKAPKRPNDKPLRPLPIQDVYKINGVGTVPAGRVESGELIPNMNVFAPQTQT--
N.an -----PCLLDALDNLVEPVRPTDKPLRPLPFQDVYKIGGIGTVPVGRVETGKLPKGMIIHFAPGAAD
A.ro -----PTLLEALDALEPPKRPTEKPLRPLPIQDVYKIGGIGTVPVGRVETGVMKREKGTVLFAPVDVS--
T.ro I -----PCLLEALDAIVPPKRPPTDKPLRPLQDVYKIGGIGTVPVGRVETGLLKPGMNVTFAPGNKT--
T.ro II -----PCLLEALDAIVPPKRPPTDRPLRPLQDVYKIGGIGTVPVGRVETGLLKPGMNVTFAPGNKT--
N.lo KDANNNLIGEKVFTLEGALNVCDLPERPIGKPLRMPITDIHTITIGITITYTGRTGTVIRPGMSISIQPANVF--

375

T.va AEVKSIEMHHESLPEALPGDNIGFNVKNVSTADVKRGYVVGDTKRDPPE---CASFTAQMIISNHPGKIHAGYQ
T.fo TDVKSIEMHHTQLPEALPGDNVGFNVK-IPVSDIKRGHVLGEQARDPPE---CINFATAQMIISNHPGKIHAGYQ
T.ag AEVKSVEMHHEQLAKAGP
N.an TEVKSIVEMHHTSVPEAGPGDNVGFNVKGLSTTDIKRGYVASDAKSDPSRE---AVSFNAQVIIMNHPEIRAGYT
A.ro TEVKSIEMHHEQLAEAILGDNVGFNINAVKDIRRGNVC SNIKNDPARG---CENFEAQVILNHPGEIQNGYA
T.ro I TEVKSIVEMHHEQLAEAEPGDNVGFNVKNLSVKDIRRGYVCSDTKRDPKQ---CESFEAQVIMNHPGQISNGYS
T.ro II TEVKSIVEMHHEQLAEAEPGDNVGFNVKNLSVKDIRRGYVCSDSKRDPKQ---CESFEAQVIMNHPGQISNGYS
N.lo GEVKSQIHRDQKEVICGENIGLALKSGAKGNLTQIKKGNVISTKTSPCVIQPACKARVIVVEHPKGIKTGYC

450

T.va PVFDCHTAHIACKFDKLIQRIDRRHGKATEN-PEYIQKDDAAIVEVVPKPLVVEFSQEQYPPPLGRFAIRDMKQT
T.fo PVFDCHTAHIACKFAKLIQRIDRRHGKVTET-PEWIKKDDAAVVVVEPSKPLVVETFOEYAAALGRFAVRDMK
T.ag
N.an PVLDCHTSHIACKFDKLIQIDRRRTGKSQEAN-PEKIKKGDASIVQVPTPKPMCVGEFTEYPPPLGRFAVR
A.ro PVLDCHTAHIACKFQELITKIDKRTGQEMEKL-PKSLKSGQAGIVNLVVPKPMCVVEEYQYPPPLGRFAVRDMR
T.ro I PVLDCHTAHIACKFETIKSLIDKRSQAVKEEN-PKGLKRGDSGIVKLVPMKPMCVESYTEYPPPLGRFAVRDMR
T.ro II PVLDCHTAHIACKFQELIKSLIDKRSQAVKEEN-PKGLKRGDSGIVKLVPMKPMCVESYTEYPPPLGRFAVRDMR
N.lo PVMDLGSHHVPAKIAKFINK--KGPDKPEVTEFDSIQNKDINALCVIVPQKPIVMEVLKDFPSSLRFPALRDGGKI

494

T.va VAVGVIRSVNKKPNPIK*
T.fo
T.ag
N.an
A.ro
T.ro I
T.ro II
N.lo VAIGSIVEVLTKEQCEKLGADVSIITQGKAAAASKPAETS SKSKK*

Figure 1.1

Southern blots of total *T. vaginalis*, *N. andersoni* and *N. locustae* DNA using the cloned EF-1 α genes from each organism respectively as hybridization probes. Although it is difficult to determine the copy number precisely by this method, both *N. andersoni* and *T. vaginalis* appear to have multiple copies of the gene (4-5 hybridizing bands in the former and 4 bands in the latter (Fig. 1.2A & C)) while *N. locustae* likely possesses only a single copy (Fig. 1.2B). The *N. andersoni* blot shows that the lanes containing DNA cut with HhaI and HinPI have hybridizing bands of less than 1 kb. Since the length of EF-1 α genes typically exceeds 1200 kb, it appears that these bands derive from EF-1 α copies that contain cutting sites for these enzymes. The alignment of the EF-1 α homolog sequenced from *N. andersoni* revealed 2 insertions relative to other homologs: a 6 amino acid insertion at positions 185-190 and a 2 amino acid insertion at positions 299-300 in the alignment (Fig. 1.1). This was somewhat surprising since the amino acid sequences of EF-1 α from independent lineages of eukaryotes are generally highly conserved in length and do not frequently display unique insertions. In addition, the presence of unique insertions in a homolog of EF-1 α from *Porphyra* was shown to be correlated with developmentally regulated expression (Liu *et al.*, 1996). Together, these facts suggested that perhaps the sequenced *N. andersoni* EF-1 α fragment could represent a non-constitutively expressed paralogous variant of the gene. In order to evaluate this possibility, an oligo was designed against the 6 codon insertion and flanking regions (this oligo had the sequence: 5'-GCTGATAAGCCAAACTTGGTCCAATTC-3'). This was used as a probe with the *N. andersoni* Southern blot. It appears that at least 4 of the possible 5 copies of EF-1 α contain the insertion (Fig. 1.2D). From these data, the presence of a single remaining EF-1 α homolog that does not possess the insertion cannot be ruled out. Characterization of all of the copies of this gene in this organism will be necessary to test whether this is true.

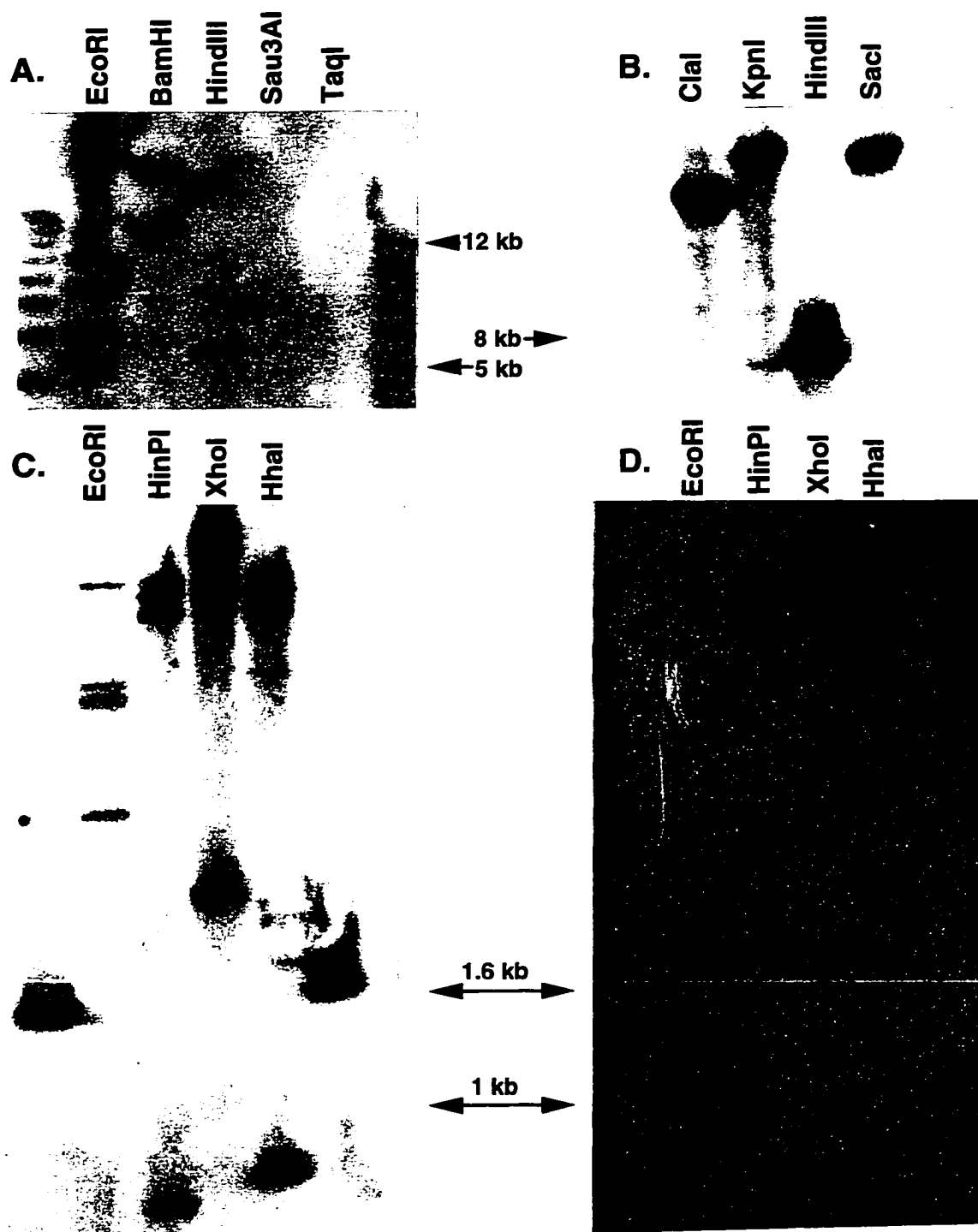


Figure 1.2 Southern blots of various genomic DNAs probed with their respective EF-1 α fragments. (A) Digested *Trichomonas vaginalis* DNA probed with EF-1 α cDNA. Digestion with Sau3AI and Taq I produced small fragments (not shown). (B) Digested *Nosema locustae* DNA probed with homologous EF-1 α fragment. (C) Digested *Naegleria andersoni* DNA probed with its EF-1 α fragment. (D) Blot (C) reprobed with 27-mer oligonucleotide complementary to the 6 amino acid insertion and flanking regions in the obtained EF-1 α homolog.

β -tubulin from *Nosema locustae*.

Amplification using primers BtubA and BtubB using *N. locustae* DNA as a template yielded a single fragment of approximately 1200 bp which was cloned and sequenced. This fragment corresponds to approximately 90% of the coding region of β -tubulin and aligns easily with homologs of this gene from other organisms (Fig. 1.3). This sequence also appears to lack introns.

Alignment features.

The 8 EF-1 α sequences described above were entered into an alignment with 71 other homologs (containing both developmentally regulated and constitutively expressed homologs). Most of these aligned easily with other homologs and few gaps were introduced into the alignment. The exception to this was the *N. andersoni* EF-1 α , which contained two insertions (as described above), and the *N. locustae* sequence which shared several small insertions in common with the *G. plecoglossi* sequence (and unique to these microsporidia), and a 15 amino acid insertion occurring at the same place as the 12 amino acid insertion in metazoa and fungi and an 11 amino acid insertion in *G. plecoglossi* (positions 223-237 in Fig. 1.1). The 12 amino acid insertion shared by Metazoa and Fungi was recently suggested to be a molecular synapomorphy uniting these groups (Baldauf & Palmer, 1993). The corresponding insertions of the microsporidia occur in exactly the same position, bear some resemblance to those of animals and fungi and may be homologous (Fig. 1.4). In addition to possessing several insertions, the *N. locustae* EF-1 α sequence appeared extremely divergent from all other eukaryotic and archaebacterial homologs (Fig. 1.1). This divergence occurs over the entire length of the molecule including the tRNA-binding regions such as alignment positions 285, 308-310 and 380-388, as well as the GTP-binding region spanning positions 91-94 in Fig. 1.1 (Liu *et al.*, 1996).

75

N.lo VGSKFWEVI SEEHGINNEGHFVGHSSNQLERINVVYNEASSSKYV PRAVLIDLEPGTMD
E.he MREIHLQGTGCGNQVGCQFWETISGEHGIDQTKYVGTSDNQLERVNVVYNEASSSKYV PRAVLIDLEPGTMD
N.cr MREIVHLQGTGCGNQIGAAFQWTISGEHGLDASGVYNGTSELQLERMNVYFNEASGNKYV PRAVLVDLEPGTMD
G.la MREIVHIQAGQCGNQIGAKFWEVISEDHGVDPSPGEYRGDSELQIERINVYFNEAAGGRYV PRAILLVDLEPGTMD
T.va MVREIVHIQAGQCGNQIGAKFWEVISEDHGDIDPTGSYHRSDQLQLERINVVYNEATGAKYV PRAILLVDLEPGTSE
E.va MREIVHVQAGQCGNQIGSKFWEVISEDHGDIDPTGSYHGSDQLQLERINCYFNEATGGRYV PRAILLMDLEPGTMD
150

N.lo SVRAGPLGRLFRPDNFIFGQSGAGNNWAKGHYTEGAELIDSVLDVVRKEAESDCLQGFQFTHSLGGGTGAGMGT
E.he AVRQGPFGDLFRPDNFVFGQSGAGNNWAKGHYTEGAELIDSVMDVVRKEAESDCLQGFQFITHSLGGGTGAGMGT
N.cr AVRAGPFGQLFRPDNFVFGQSGAGNNWAKGHYTEGAELVDQVLDVVRREAEGCDCLQGFQFITHSLGGGTGAGMGT
G.la SVRAGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDAVLDVVRKRSEACDCLQGFQFICHSLGGGTGAGMGT
T.va SVRAGQFGQLFRPDNFVFGQSGAGNNWAKGYTTEGQELCESILDVIRKEAESDCLQGFQFVHSLGGGTGAGLGT
E.va SVRAGPFGQLFRPDNFVFGQSGAGNNWAKGHYTEGAELIDSVLDVVRKEAESCDALQGFQFVHSLGGGTGAGMGT
225

N.lo LLLSKIREEYPDRMMCTFSVVPSPKVS DTVVEPYNATLSIHQLVENADETFCIDNEALYDICFRTLKLSPTGYGE
E.he LLLSKITREDFDRMICTFSVVPSPKVS DTVVEPYNATLSIHQLVENADETFCIDNEALYDICFRTLKMSNPGYGD
N.cr LLLSKIREEFDRMMATYSVVPSPKVS DTVVEPYNATLSVHQLVENSDETFCIDNEALYDICMRTLKLSNPSYGD
G.la LLLAKIREEYPDRMMCTFSVVPSPKVS DTVVEPYNATLSVHQLVEHADEVFCIDNEALYDICFRTLKLTCPITYGD
T.va LLLNKLREEYPDRILSTYSIVPSPKVS DTVVEPYNCTLSVHQLVESADEVFCIDNEALYDICFRTLKLTPTITYGD
E.va LLLSKVREEYPDRVMSTYSIVPSPKVS DTVVEPYNATLSVHQLVENADQCFTLDNEALYDICFRTLKLTPTITYGD
300

N.lo LNHLVSLVMSGVTTCLRFPGQLNADLRKLA VNMVFPRLHFFIVGFAPLIAQGT SQYRTYSVSELT SQMFDSKNM
E.he LNHLVSLVMSGVTTCLRFPGQLNADLRKLA VNMIPFRLHFFVVGSAPLIAIGTQKFKTYSVSELT SQMFDSKNM
N.cr LNHLVSAVMSGVTVSLRFPGQLNSDLRKLAVNMVFPRLHFFMVGFAPLTSRGAHFRFRAVSVPELT SQMFDPKNM
G.la LNHLVSLVMSGCTSLRFPGQLNADLRKLA VNLIPFRLHFFLVGFAPLTSRGSQYRALTVPELVSQMF DNKNM
T.va LNHLVSMVMSGTTCALRFPGQLNSDLRKLAVNLPFRLHFFIVGFAPLTSRGSQQYRALTVPELTSQLF DNKNM
E.va LNHLVSAACGTTC SLRFPGQLNCDLRKLAVNMVFPRLHFFMIGFAPLTSRGSQQYRALTVPELT SQCFDSKNM
375

N.lo MAASDPRHGRYLTAGLPVFRGKISMKD VDEQMLQVQTRNSAHFVEWIPNNVKTAVCDIPPSGLDMSATFIGNSTS
E.he MTACDPRKGRYLTVAAMFRGKISMKD VDEQMSMVQSKNSTL FVEWIPSNVKTAVCDIAPTGLEMSATFVGNITS
N.cr MAASDFRNRGRLTCSAIFRGRKISMKEVEDQMRNVQNKNSYFVEWIPNNVQTALCSIPPRGLKMSSTFVGNSTA
G.la MAASDPRHGRYLTAAA-MFRGRMSTKEVDEQMLNIQNKNSYFVEWIPNNMKVSVCDIPPRGLKMAATFIGNSTC
T.va MAACDPRRVSYLTS-A-HFRGRMSSKEVDEQMLNIQARNTSYFVEWIPSNVKS AICDIPPRGLKMAATFIGNTTA
E.va MCAADPRHGRYLTCAV-MFRGRMSTKEVDEQMLNVVNKNSYFVEWIPNNVKASICDIPPKGLKMSSTFVGNITA
450

N.lo IQELFKRISDQFSVMFRRKAFLHWYTGE GMDMEFTR
E.he IQELFKRISDQFTVMFRRKAFLHWYTGE GMDMEFTEAESNMNDLLSEYQYQDATVEDAE EFLVN*
N.cr IQELFKRIGEQTAMFRRKAFLHWYTGE GMDMEFTEAESNMNDLVSEYQYQDAGVDEEE EEEEEYEEEA PLEGE*
G.la IQELFKRVGEQFSAMFRRKAFLHWYTGE GMDMEFTEAESNMNDLVSEYQYQEA GVD EGE EEEEEDFGDE*
T.va FRELFTRVDSQFQKMYARRAFIHWYVNE GLETVEFDEARSNM TDLIQEYEMYETAG*
E.va IQEVVKRVAEQFTSMFRRKAFLHWYTGE GMDMEFTEAESNMNDLVSEYQYQDATAEE EGEFDEDEELDDAMG*

Figure 1.3 An alignment of the partial *Nosema locustae* β -tubulin gene with homologs from other organisms. The species name abbreviations are *N.lo*: *Nosema locustae*; *E.he*: *Encephalitozoon hellem*; *N.cr*: *Neurospora crassa*; *G.la*: *Giardia lamblia*; *T.va*: *Trichomonas vaginalis*; and *E.va*: *Ectocarpus variabilis*. The bolded taxon label indicates the sequence obtained in this study. Asterisks (*) indicate stop codons. The sequences in the alignment are selected from a full alignment of 67 β -tubulin homologs.

Extreme divergence from other eukaryotic EF-1 α sequences was also noted for the *G. plecoglossi* homolog (Kamaishi *et al.*, 1996).

The *N. locustae* β -tubulin sequence was aligned with few gaps to 65 other eukaryotic homologs. No especially unique divergence from other eukaryotic homologs was noted (Fig. 1.3).

Phylogenetic analysis.

For phylogenetic analysis of both EF-1 α and β -tubulin, a subset of taxa were selected from the alignments to represent all major organismal groups and exclude those sequences known to be developmentally regulated. To prevent long branch attraction effects, the highly divergent *Entamoeba histolytica* sequence (Edlind *et al.*, 1996) was removed from the β -tubulin dataset. Since the two *Tetramitus* sequences differ only by a single amino acid, the *T. rostratus* II sequence was not included in the final EF-1 α dataset. The final full datasets contained 38 β -tubulin and 40 EF-1 α sequences. Alignment positions were chosen for phylogenetic analysis by criteria described in the Materials and Methods. For distance and parsimony analysis, 430 positions of EF-1 α and 432 positions of β -tubulin were considered. Partial sequences with a large proportion of missing data (e.g. *Spironucleus muris* and *Trepomonas agilis* EF-1 α sequences) as well as gapped positions in the alignment were removed from the datasets for maximum likelihood analysis leaving datasets of 38 taxa and 382 positions for EF-1 α and 38 taxa and 394 positions for β -tubulin.

EF1 α phylogeny.

Neighbour-joining protein-distance, maximum parsimony and maximum likelihood methods were used to infer trees from the EF-1 α dataset, and for each method bootstrap support for branches was evaluated (Figure 1.5A & B).

Artemia	DRLPWYKGNIERKEGKADK---	TLLDALDAI] Metazoa
Xenopus	PNMPWFKGKTRKEGSGSPT---	TLLEALDCI	
Drosophila	TNMPWFKGVEVGRKEGNADK---	TLVDALDAI	
Mucor	TNMPWFKGNFETKASSTK---	TLLEAIDAI	
Candida	TNCPWYKGEETKSKVTK---	TLLEAIDAI] Fungi
Yeast	TNAPWYKGEETKAVVKK---	TLLEAIDAI	
Glugea	DKFEWFKGKPVSGAEDSIF----	TLEGALNSQ] Microsporidia
Nosema	EKFEWFEGQKDKANNLISEKVF	TLEGALNYC	
Trichomonas	PNMPWYNGP-----	YLLEALDSL] Trichomonads
Giardia	DKMPWYEGP-----	CLIDAIDGL] Diplomonads
Methanococcus	ERNPWYKGP-----	TIAEVIDGF] Archaeobacteria
Thermococcus	DKMPWYNGP-----	TLIEALDQM	
Sulfolobus	TKMPWYNGP-----	TLEELLDQL	

Fig 1.4 A possibly homologous insertion in EF-1 α genes of Microsporidia, Fungi and Metazoa. This region corresponds to positions 215-247 in the EF-1 α alignment shown in Fig. 1.1 and is taken from a full alignment of 79 eukaryotic and archaeobacterial homologs. Shaded boxes indicate amino acids located within the insertions that are conserved between microsporidian sequences and either fungal or metazoan sequences.

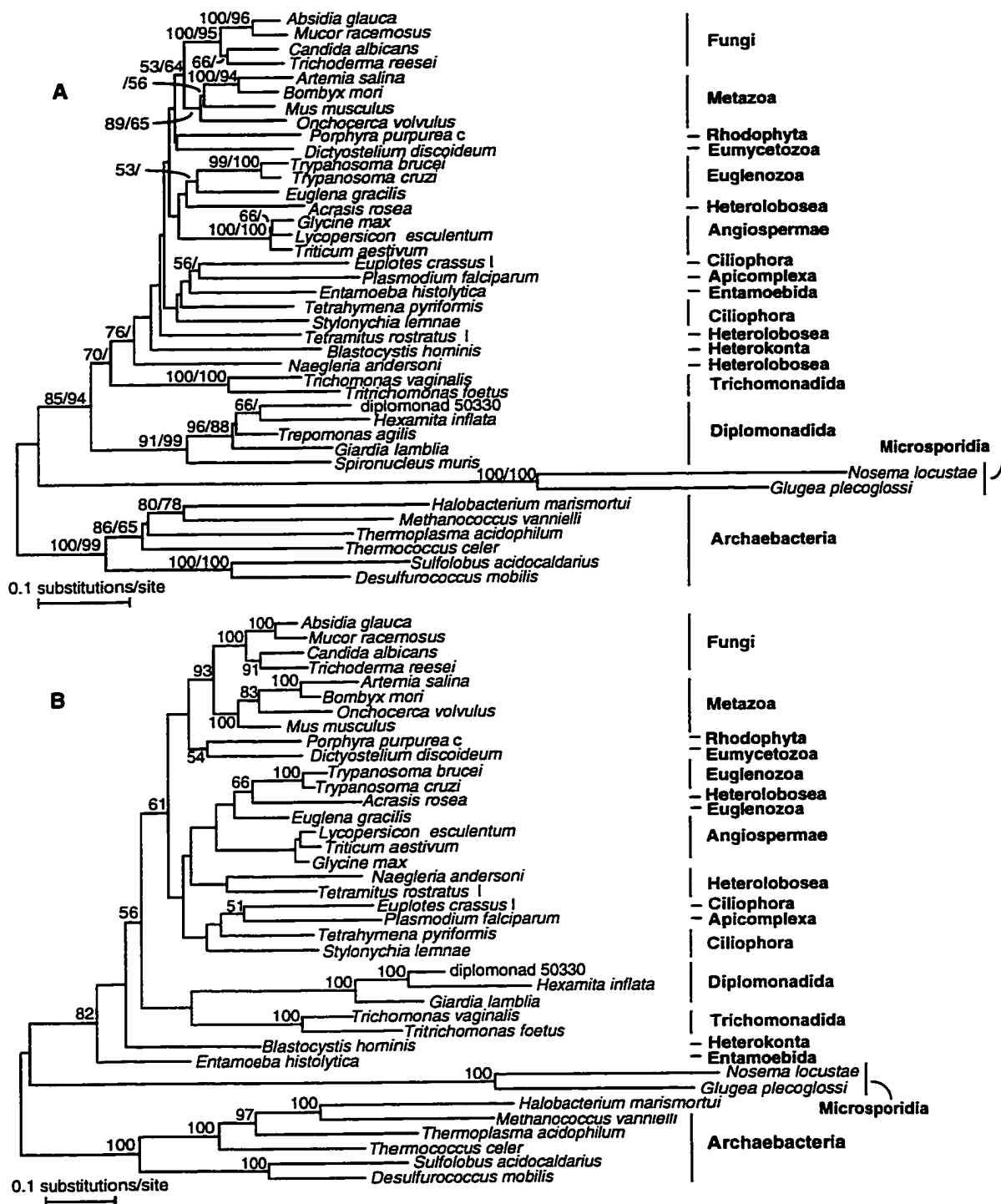


Figure 1.5 Eukaryotic phylogeny inferred from EF-1 α sequences. (A) Neighbour-joining (NJ) tree inferred from a distance matrix generated by the PROTDIST program using the Dayhoff option. Parsimony yielded 5 equally parsimonious topologies of length = 2718 steps all of which differed in minor respects to the NJ. Values at the nodes are percentage bootstrap support obtained from 300 bootstrap replicates using both NJ and parsimony methods and the values are reported in the order: NJ/parsimony. (B) Maximum likelihood tree (log-likelihood = -15217.2) inferred using the JTT-F model of amino acid substitution. Percentage bootstrap support estimated using the REL method are shown at the nodes. For all methods, bootstrap values of <50% are not shown.

All methods yielded monophyletic microsporidian, trichomonad and diplomonad groups with strong bootstrap support (Fig. 1.5A & B). Within the diplomonads, only some of the branchings were strongly supported. In distance and parsimony trees, *Spironucleus muris* appeared as the earliest offshoot followed by *Giardia lamblia* and *Trepomonas agilis* with diplomonad 50330 and *Hexamita inflata* forming a terminal clade. The branching order within the latter 4 taxa was not well supported by bootstrap analysis. This phylogeny of the diplomonads conflicts with two recently published phylogenies based on glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and SSU rRNA which show the two free-living species, *Hexamita inflata* and *Trepomonas agilis* as immediate sister groups to the exclusion of *Spironucleus muris* (Cavalier-Smith & Chao, 1996) and diplomonad 50380 (a close relative of diplomonad 50330) (Rozario *et al.*, 1996). Since both of these latter phylogenies are strongly supported and mutually congruent, it is most likely that the poorly supported distance and parsimony trees of diplomonads inferred from the EF-1 α dataset are in error. The most significant conflict with other data is the robust placement of the *Spironucleus muris* sequence as an outgroup to all other diplomonads in the EF-1 α tree. SSU rRNA phylogeny (Branke *et al.*, 1996, Cavalier-Smith & Chao, 1996) and ultrastructural considerations (Siddall *et al.*, 1992) suggest that the genus *Spironucleus* is closer to *Hexamita* than to *Giardia*. Thus, it is most likely that the EF-1 α sequence from *S. muris* is artifactually misplaced due to an acceleration of the rate of amino acid substitution in this lineage. Interestingly, the maximum likelihood tree, which lacks the *Spironucleus* and *Trepomonas* sequences, shows a robust early divergence of *Giardia lamblia* separated from a *Hexamita inflata* and diplomonad 50330 clade, in agreement with GAPDH phylogeny (Rozario *et al.*, 1996).

In contrast to the above mentioned protist groups, the monophyly of the Ciliophora (represented by *Euplotes crassus* I, *Stylonychia lemnae* and *Tetrahymena pyriformis*), the Alveolata (containing these ciliates and the apicomplexan *Plasmodium falciparum*) and the Heterolobosea (represented by *Naegleria andersoni*, *Tetramitus rostratus* I and *Acrasis rosea*) were not strongly supported by EF-1 α trees inferred with any of the phylogenetic methods (Fig. 1.5A & B) despite ultrastructural and molecular phylogenetic evidence that each of these groups is monophyletic (Page & Blanton, 1985, Hirt *et al.*, 1995, Roger *et al.*, 1996).

In order to examine whether these results were significantly preferred over the phylogeny predicted from ultrastructural and molecular phylogenetic studies, the Kishino-Hasegawa test was employed to test alternative topologies using the maximum likelihood (ML) method (Kishino & Hasegawa, 1989).

Table 1.1A shows the results of this test where the ML topology was rearranged to contain a Ciliate/Alveolate tree consistent with ciliate relationships discerned on the basis of ultrastructure and SSU rRNA phylogenetics (Hirt *et al.*, 1995). The difference in ln likelihood between this tree and the best is 1.00 SE worse, indicating that it is not significantly excluded at the 5% level (Table 1.1A) (a Z-value of 1.96 SE corresponds to the 95% confidence interval) (Kishino & Hasegawa, 1989).

Similarly, a heterolobosean topology was imposed on the ML tree (Page & Blanton, 1985, Roger, *et al.*, 1996). This test was more difficult since the maximum likelihood topology of EF-1 α has this group as polyphyletic: placing the acrasid slime mould, *Acrasis rosea* within the Euglenozoa while the vahlkampfiid amoebae, *Naegleria andersoni* and *Tetramitus rostratus* I, form a weak clade situated as sister group to a Euglenozoa/*Acrasis* / Angiospermae grouping (Fig. 1.5B). To deal with this, a Heterolobosea subtree was created by alternatively moving *Acrasis* down as a sister group to the vahlkampfiids and

moving the vahlkampfiids up to branch with *Acrasis* (Table 1.1B). Neither topology is significantly worse than the ML tree: the Heterolobosea as a sister group to the Euglenozoa/Angiospermae clade was 1.15 SE worse than the ML tree, and placing the Heterolobosea within the Euglenozoa was 1.55 SE worse. Of the two positions of the Heterolobosea, the one interrupting the Euglenozoa is least preferred. Swapping the *Euglena* and Heterolobosea branches to make the Euglenozoa monophyletic improves the likelihood of the resulting tree, although not strongly (Table 1.1B).

The branching order of major eukaryotic groups in the EF-1 α tree.

Trees from each of the phylogenetic methods placed the Microsporidia clade (containing *Nosema locustae* and *Glugea plecoglossi*) as the earliest offshoot of the eukaryotic lineage with strong bootstrap support, congruent with other analyses (Kamaishi *et al.*, 1996). However, the methods do not concur on which groups branch next. Neighbour-joining and parsimony trees have the diplomonads emerging after the Microsporidia followed by the trichomonads. Both of these branchings are moderately supported by bootstrap analysis using the neighbour-joining method (70% and 76% respectively) but poorly supported by parsimony (<50% in both cases). By contrast, the likelihood method shows *Entamoeba histolytica* and *Blastocystis hominis* as the next earliest divergences, although bootstrap support for both of these branchings is low. Interestingly, the maximum likelihood topology shows the diplomonads and trichomonads as a clade emerging just after *Blastocystis*.

Extreme differences in the rates of substitutions between different lineages are well known to cause the failure of phylogenetic methods to reconstruct the correct phylogeny (Felsenstein, 1978, Hasegawa & Fujiwara, 1993). The extremely divergent nature of the microsporidian sequences, as evidenced by their

extremely long branch (approximately 2x as long as any other branch in the trees in Fig. 1.5), may be responsible for the lack of congruence between methods in reconstructing the early branching order of eukaryotes. To investigate the possibility that the Microsporidia

Table 1.1- Imposing a Ciliate/Alveolate and a Heterolobosea tree on the EF-1 α topology

Tree*	lnL**	Δ lnL***	Z-value****
ML topology	-15217.2	0	best
<u>A. Ciliate / Alveolate tree:</u>			
(Pf, (Tp, (Sl, Ec)))	-15228.6	-11.3	1.00
<u>B. Heterolobosea trees:</u>			
((Ar, (Tr, Na), (AS, EZ)))	-15236.8	-19.6	1.15
(Eg, ((Ar, (Tr, Na)), (Tb, Tc)))	-15244.9	-27.7	1.55
((Ar, (Tr, Na)), EZ)	-15240.9	-23.7	1.44

*Species and group abbreviations are Pf: *Plasmodium falciparum*; Tp: *Tetrahymena pyriformis*; Sl: *Stylonychia lemnae*; Ec: *Euplotes crassus*; Ar: *Acrasis rosea*; Tr: *Tetramitus rostratus*; Na: *Naegleria andersoni*; Eg: *Euglena gracilis*; Tc: *Trypanosoma cruzi*; Tb: *Trypanosoma brucei*; AS: Angiospermae; and EZ: Euglenozoa. **Log-likelihood of the topology. ***difference in lnL from the ML topology. **** Δ lnL divided by its standard error.

were artifactually influencing the topology of the eukaryotic tree, analyses using each of the methods were performed with these two species removed from the dataset and are shown in Figure 1.6. All of the methods show the diplomonads as the earliest offshoot followed by trichomonads. Although the branching of the diplomonads first is not strongly supported, there is reasonable bootstrap support (80% and 72% for NJ and ML methods respectively) separating both of these groups from all other eukaryotes. In addition, the relative branching order of some crown groups, such as *Dictyostelium* and *Porphyra*, changes with the exclusion of the Microsporidia (Fig. 1.5 & 1.6). Thus it is clear that the relative

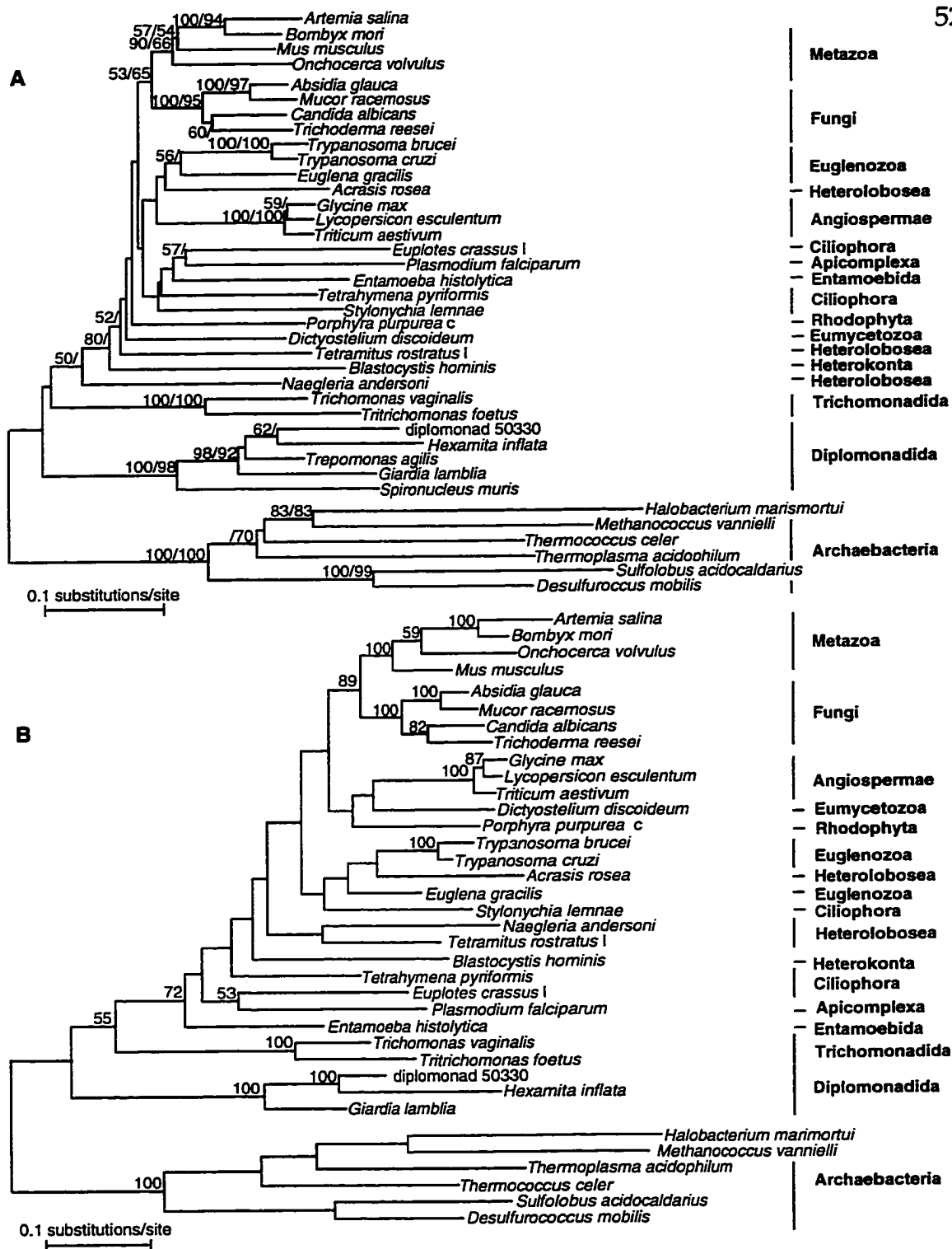


Figure 1.6 EF-1 α phylogeny with the Microsporidia removed. (A) Neighbour-joining tree inferred from pairwise distances estimated by the PROTDIST program using the Dayhoff option. Parsimony analysis yielded 4 equally parsimonious topologies of length = 2360 steps. Much of the internal structure of the trees was not conserved between equally parsimonious reconstructions. Bootstrap values on branches were determined from 300 resamplings of the dataset and are shown as NJ/parsimony. (B) Maximum likelihood tree (log-likelihood = -13580.78) inferred using the JTT-F model of amino acid substitution. Percentage bootstrap support estimated using the REL method are shown at the nodes. For all methods bootstrap values <50% are not shown.

branching order of eukaryotic groups is influenced by the presence of the Microsporidia, especially when the maximum likelihood method is employed.

After the earliest branchings, the relative branching order of groups in the EF-1 α tree appears to be very poorly resolved. As mentioned above, the monophyly of the Ciliophora, Alveolata, Heterolobosea and the Euglenozoa are not well resolved. In addition, the relative branching order of species of these taxa and the Angiospermae changes with different methods and different species samplings (Fig. 1.5A & B, Fig. 1.6A & B). One relatively consistent feature is the placement of at least one heterolobosean (most often *Acrasis*) and the Angiospermae in a clade with the Euglenozoa, although this grouping is never well-supported. In some trees (Figs. 1.5B & 1.6A), a large cluster containing the alveolates, the Euglenozoa, the Angiospermae and the Heterolobosea is also evident, but again this group is not supported by bootstrap analysis. The Metazoa and Fungi always group together in EF-1 α trees with moderate bootstrap support, congruent with ultrastructural and abundant molecular evidence that these are sister groups (Baldauf & Palmer, 1993, Cavalier-Smith, 1993a, Nikoh *et al.*, 1994). When Microsporidia are included, this clade is a sister group to a poorly supported *Porphyra/Dictyostelium* clade.

β -tubulin phylogeny.

The overall branching order of the β -tubulin trees inferred by neighbour-joining distance (Fig. 1.7A), maximum parsimony (not shown) and maximum likelihood (Fig. 1.7B) were similar to recently published phylogenies of this molecule (Edlind *et al.*, 1996). These analyses have rooted the β -tubulin tree on α - and γ -tubulin paralogs found in other eukaryotes. I chose not to follow this practice because these proteins are extremely distant to any β -tubulin sequence and subtrees of each paralog are separated from the others by long unbroken

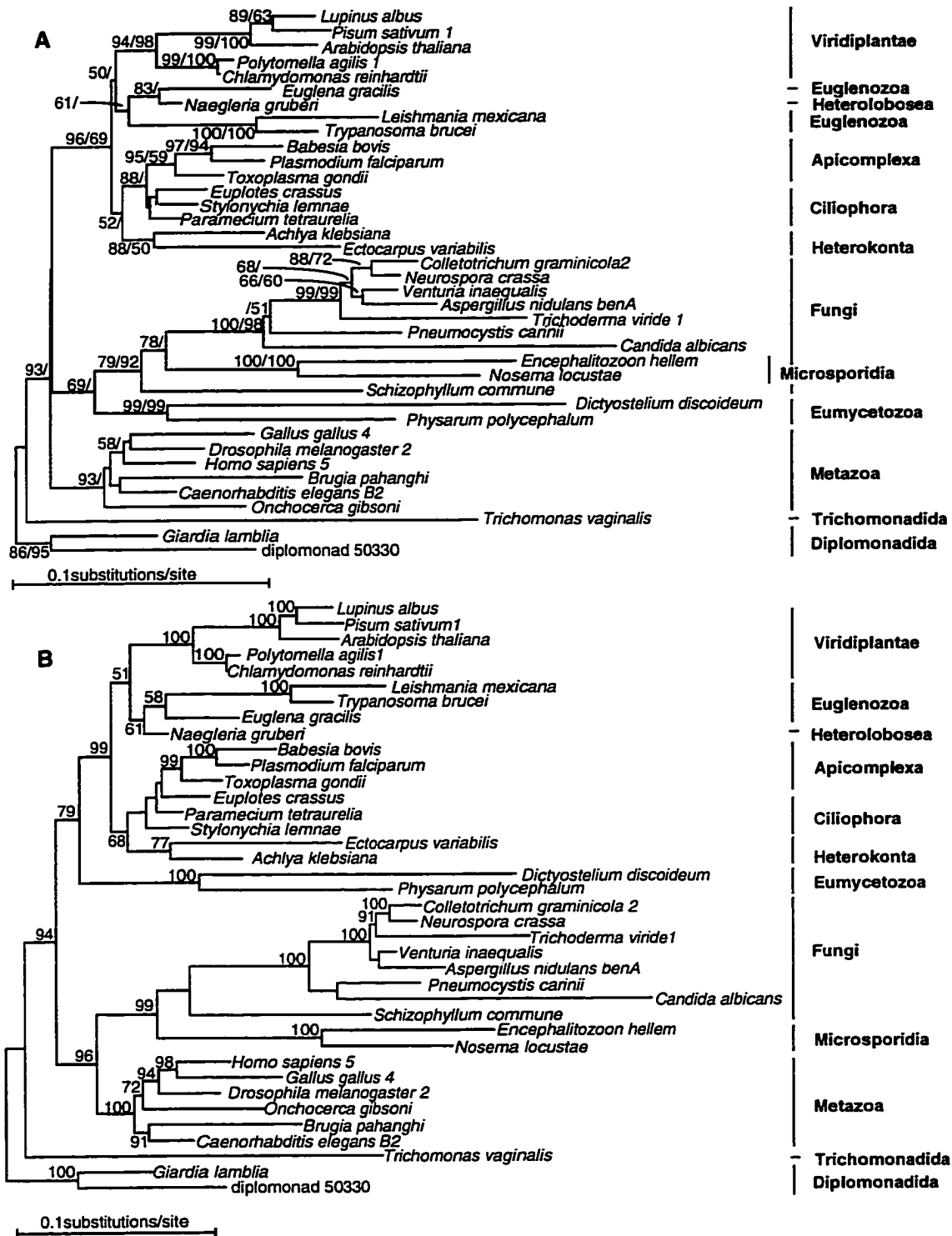


Figure 1.7 Eukaryotic phylogeny inferred from β -tubulin sequences. The trees are rooted on the diplomonad lineage. (A) Neighbour-joining topology inferred from a distance matrix obtained using the PROTDIST program on the Dayhoff setting. The 12 equally parsimonious trees of length = 1071 were similar to both of the topologies shown above and differed from each other in the placement of the Heterokonta and the branching order within the alveolates. Percent bootstrap support for nodes obtained from 300 resamplings are shown over the branches in the order NJ/parsimony. (B) Maximum likelihood topology (log-likelihood = -7573.5) inferred using the JTT-F model of amino acid substitution. Bootstrap values are estimates obtained from the RELL method. Bootstrap values of <50% are not shown.

branches in the tree (see Fig. 2B in Edlind *et al.*, 1996). This property makes them unsuitable for an outgroup, as they will likely attract any other fast rate sequences to the base of the tree artifactually (for instance the extremely divergent *E. histolytica* sequence in Edlind *et al.* (1996) likely groups with the outgroup for this reason). Instead I have chosen to root the tree on the Diplomonadida lineage since trees of V-ATPase, SSU rRNA (Cavalier-Smith & Chao, 1996), hsp70 (Gupta *et al.*, 1994), RNAPol IIA (Klenk *et al.*, 1995), EF-2 and EF-1 α (Hashimoto & Hasegawa, 1996) all concur in placing this group as one of the deepest branches in the eukaryote tree.

If this root is chosen, then the second earliest divergence after the strongly supported monophyletic diplomonad clade is *Trichomonas vaginalis*. These two groups are separated from all others by significant bootstrap values for neighbour-joining distance and likelihood methods (93% and 94% respectively in Fig. 1.7A & B). The tree then splits into 2-3 large clades that differ in exact composition depending on the phylogenetic method. Specifically, two groups, the Metazoa and the Eumycetozoa (containing the slime moulds, *Dictyostelium discoideum* and *Physarum polycephalum*), differ in their placement between neighbour-joining distance analysis on the one hand and parsimony and likelihood on the other. Neighbour-joining places the Metazoa as a poorly supported outgroup relative to the two other major clades while placing the Eumycetozoa as a sister group to the Fungi and Microsporidia with moderate bootstrap support. However, both likelihood and parsimony methods place the Metazoa as a sister group to the Fungi and Microsporidia (this grouping is supported strongly by likelihood but only moderately by parsimony), and show the Eumycetozoa as a sister group (moderately supported by both methods) to a super-clade containing the Heterokonta, Heterolobosea, Euglenozoa, Alveolata

(and therefore the Ciliophora) and the Viridiplantae (land plants and green algae).

By contrast, all methods concur on a close relationship of the Microsporidia with the Fungi. Neighbour-joining and parsimony analyses show a monophyletic microsporidian group emerging from within the Fungi, diverging after the basidiomycete *Schizophyllum commune*. Likelihood instead reconstructs a monophyletic fungal clade with the Microsporidia as a sister group. Parsimony and likelihood bootstrap analyses suggest that a monophyletic Fungi/Microsporidia clade is extremely well supported (with bootstrap values of 92% and 99% respectively) while neighbour-joining shows moderate bootstrap support (79%).

Similarly, the clade containing the Heterokonta, Heterolobosea, Euglenozoa, Alveolata and Viridiplantae is recovered by all phylogenetic methods, gaining significant bootstrap support from neighbour-joining and likelihood (96% and 99% respectively) with reduced support from parsimony (69%). The relative branching order within this grouping is also congruent among methods. Each method recovers a relationship of the heterolobosean, *Naegleria gruberi* to the Euglenozoa (the former sometimes interrupting the latter group) with this grouping forming a sister to the Viridiplantae. These three taxa form a clade to the exclusion of a heterokont/alveolate grouping. However, none of these intergroup relationships is particularly well-supported by bootstrap analyses (Fig. 1.7).

Reconciling β -tubulin and EF-1 α trees.

EF-1 α and β -tubulin trees are congruent in recovering a number of eukaryotic nodes, however, the position of the Microsporidia is strongly conflicting between trees of these datasets. As mentioned above, the conflicting

placement of the Microsporidia in either dataset could be a result of several factors. Firstly, it is possible that the conflicts are a result of lack of phylogenetic resolution. If either the EF-1 α or β -tubulin datasets are not significantly resolving the position of the microsporidia, then various placements of them should not be significantly worse than the optimal one using the Kishino-Hasegawa test. If the conflicts are significant, then either the true gene phylogenies are different (the Microsporidia have acquired one or other of these proteins via lateral transfer or one or other of the trees involves comparison between paralogous homologs) or for one or other of the genes, the phylogenetic methods are systematically recovering an artifactual topology.

The hypothesis that Microsporidia are a sister group to Fungi was tested using the EF-1 α dataset in two ways. Firstly, the optimal topology (Fig. 1.5A, denoted tree A in Table 1.2) was rearranged so that the Microsporidia branched with the Fungi and all other elements of this tree were held constant (tree B in Table 1.2). This method makes the implicit assumption that the presence of Microsporidia will not locally influence the topology of the tree. However, if the extremely long branch of the Microsporidia is artifactually influencing the topology of nearby nodes, then moving them to a node as far away as the Fungi may leave a highly suboptimal topology in the region from which they were moved. In order to account for this possibility, another test was performed using a tree obtained from the dataset lacking the Microsporidia (Fig. 1.6B) modified to include the Microsporidia as a sister group to the Fungi (topology C in Table 1.2). This tree should be optimized in early branches of the tree in a manner independent of the presence of the microsporidian sequences.

For the first test, the difference in likelihood of tree A and B was highly significant (3.02 SE). However, the likelihood of the second tree showing the Microsporidia/Fungi grouping, tree C, was much higher than tree B and was not

significantly worse (0.31 SE) than the optimal tree (Table 1.2). This result suggests that a placement of the Microsporidia with the Fungi is not excluded by the EF-1 α dataset. In addition, the location of the Microsporidia at the base of the eukaryotes in the optimal tree appears to strongly influence the branching order in their immediate vicinity. Clearly, this topology is not optimal when the Microsporidia are removed from the dataset (as shown by the discrepancies between the trees in Fig. 1.5B and 1.6B discussed above) or when they are located on a distant branch, as a sister group to the Fungi.

For the β -tubulin dataset, a reciprocal test was performed on several alternative topologies. Firstly, in order to make the test comparable to the EF-1 α dataset, a maximum likelihood tree of EF-1 α sequences was estimated from a dataset lacking the Archaeobacteria as outgroup sequences. The optimal topology obtained from this dataset displayed Microsporidia as an immediate sister group to *Blastocystis hominis*. It is not possible to test this topology with the β -tubulin dataset, since this gene has not yet been characterized from any *Blastocystis* species. However, recently Silberman *et al.*

Table 1.2- The position of Microsporidia in the EF-1 α tree.

Tree*	lnL**	Δ lnL***	Z-value****
Tree A-(MS, AB)	-15217.2	0	best
Tree B- (MS, FG)	-15250.1	-32.9	3.02#
Tree C- (MS, FG)	-15229.5	-12.3	0.31

*Group name abbreviations are MS: Microsporidia; AB: Archaeobacteria; and FG: Fungi.

Log-likelihood of the topology. *Difference in lnL from the ML topology. **** Δ lnL divided by its standard error. #significant at the 5% level.

have used SSU rRNA phylogenetics and ultrastructural arguments to show that the parasites of this genus are closely related to heterokont algae (Silberman *et al.*, 1996). Thus for the β -tubulin dataset, it was possible to test a comparable

topology by placing the Microsporidia as a sister group to *Ectocarpus variabilis* and *Achlya klebsiana*, the two heterokont sequences available. Since neighbour-joining trees of this EF-1 α dataset and many SSU rRNA trees show the Microsporidia to be closest to diplomonads and trichomonads, I also tested topologies consistent with these relationships.

Although, the microsporidian β -tubulin sequences are not particularly divergent (Fig. 1.3), two sets of topologies were tested against the optimal topology in a similar manner to the EF-1 α analysis. A first set (denoted set A) was derived from the optimal topology of the β -tubulin dataset, with the Microsporidia rearranged to be a sister group with the various taxa mentioned above. The second set (set B) was derived from a maximum likelihood tree estimated from a β -tubulin dataset lacking the microsporidian sequences with the Microsporidia inserted in several places to form the groups under test. The topologies of the latter set of trees should be independent of any influence exerted by the microsporidian sequences. The results of these tests on sets A and B are shown in Table 1.3.

All of the various placements of Microsporidia in the trees of set A were significantly worse than their optimal position as a sister group to the Fungi. Interestingly, the placement of Microsporidia with the Heterokonts, reflecting the maximum likelihood EF-1 α topology, has by far the lowest log-likelihood, the difference being 4.39 SE from the optimal tree. The ML tree of the β -tubulin dataset lacking the Microsporidia was very similar to the tree shown in Fig. 1.7B, indicating that the topology is not strongly influenced by the presence of the microsporidian sequences. All of the set B topologies based on this tree were also found to be significantly worse than the optimal tree, indicating that these relationships are all strongly excluded by the β -tubulin dataset.

Table 1.3- The position of Microsporidia in the β -tubulin tree

<u>Tree*</u>	<u>lnL**</u>	<u>ΔlnL***</u>	<u>Z-value****</u>
(MS, FG)	-7573.5	0	best
<u>Set A-</u>			
(MS, HK)	-7652.0	-78.5	4.39#
(MS, DM)	-7615.2	-41.6	2.72#
(MS, (DM, Tv))	-7610.5	-36.0	2.55#
<u>Set B-</u>			
(MS, HK)	-7655.5	-82.0	4.34#
(MS, DM)	-7618.8	-45.2	2.67#
(MS, (DM, Tv))	-7614.4	-40.9	2.54#

*Species and group name abbreviations are MS: Microsporidia; FG: Fungi; HK: Heterokonta; DM: Diplomonadida; Tv: *Trichomonas vaginalis*. **Log-likelihood of the topology. ***Difference in lnL from the ML topology. **** Δ lnL divided by its standard error. #significant at the 5% level

DISCUSSION

The phylogenetic position of the Microsporidia within eukaryotes.

The foregoing analysis confirms and extends a recent analysis (Kamaishi *et al.*, 1996) that showed Microsporidia as branching as the deepest eukaryotic lineage in phylogenetic trees of EF-1 α . However, using identical methodology, I have also shown that trees of another protein coding gene, β -tubulin, place this protist group as a derived eukaryotic lineage with a strong relationship to the Fungi consistent with other recent reports (Edlind *et al.*, 1996, Li *et al.*, 1996). On the surface these results seem contradictory. However, the Kishino-Hasegawa tests shown in Table 1.2 indicate that while a relationship of Microsporidia to Fungi is not optimal for the EF-1 α dataset, it is not significantly excluded. By contrast, the relationship of Microsporidia to Heterokonts (suggested by ML analysis of EF-1 α), to diplomonads or a diplomonad/trichomonad clade

(suggested by NJ analysis of EF-1 α and published SSU rRNA analyses) are all topologies significantly excluded by the β -tubulin dataset (Table 1.3). Therefore, consideration of both of these proteins suggests that a relationship of Microsporidia to Fungi is the hypothesis best supported by the data.

Two recent analyses of the α -tubulin dataset also report a strong affinity between Microsporidia and Fungi (Li *et al.*, 1996, P. Keeling & W. F. Doolittle, pers. comm.). Moreover, the presence of insertions in both microsporidian EF-1 α sequences in the same position as similar insertions unique to the Metazoa and Fungi (Fig. 1.4) are also suggestive of a relationship between these three taxa.

If this phylogenetic position for Microsporidia is correct, then several observations need to be explained. Firstly, the recent phylogenetic analysis of an EF-1 α dataset including the microsporidian *Glugea plecoglossi*, reported a robust placement at the base of the eukaryotic tree using the maximum likelihood and maximum parsimony methods (Kamaishi *et al.*, 1996). My analysis also shows similarly high bootstrap values for this placement using all three phylogenetic methods. It is difficult to see how this can be reconciled with the Kishino-Hasegawa test results. The explanation likely rests on biases introduced into the analysis by the methodology used to recover the ML tree. For instance, Kamaishi *et al.* (1996) used a semi-constrained topology to arrive at an estimate of the maximum likelihood tree. Their constraints included a Metazoa/Fungi/Plant/Euglenozoa/*Dictyostelium* constrained subtree effectively preventing the consideration of a Microsporidia/Fungal relationship. In my analysis, the bootstrap values estimated for the ML tree (Fig. 1.5B) are derived from 200 trees obtained by the heuristic "quick-add OTU" tree-searching procedure. The accuracy of these bootstrap values obtained by the RELL method is a function of how well the EF-1 α "tree-universe" is represented amongst these trees. If some nodes and combinations of nodes are not represented in this

sample of the universe, then the bootstrap values for them will be underestimated. Conversely, the values for represented nodes will likely be overestimated. Thus, it is possible that the strongly supported position of microsporidia in these ML EF-1 α trees is due to an artifact of this sort.

Secondly, both of the microsporidian EF-1 α sequences are extremely divergent. Since the branch connecting the Archaeobacterial subtree with the eukaryotes is long and unbroken it is possible that it artifactually attracts the microsporidian sequences. It has been argued that the ML method is less sensitive to the long branch attraction phenomenon than other phylogenetic methods (Felsenstein, 1978, Hasegawa & Fujiwara, 1993, Kuhner & Felsenstein, 1994). Yet, if the inequality of rates between lineages is severe enough, even this method will be inconsistent (Hasegawa & Fujiwara, 1993, Yang, 1996). In addition, Lockhart *et al.* have recently shown that the presence of invariant sites in datasets can lead both neighbour-joining and maximum likelihood methods to yield an inconsistent estimate of phylogeny when rates of evolution are different between lineages (Lockhart *et al.*, 1996). This problem is exacerbated if there are varying patterns of invariant sites between homologous genes of different functions. Examination of patterns of conservation of amino acids in the full EF-1 α alignment show that the eukaryotic subset (63 sequences with developmental isoforms excluded) displays a total of 182 positions that are invariant or vary in only one sequence. Of these positions, the Microsporidia vary at 86 while Archaeobacteria vary at 75 with the two groups sharing 47 positions variable in both. Therefore, it seems likely that the effect described by Lockhart *et al.* (1996) could be influencing the placement of the Microsporidia in the EF-1 α tree, causing them to artifactually group with the Archaeobacteria.

Another observation that requires explanation is the deep placement of Microsporidia in the SSU rRNA tree of eukaryotes. Recently, Galtier and Gouy

have analysed this dataset with a method designed to be insensitive to variation in base composition in different lineages and have argued that their results provide strong evidence that the Microsporidia are the earliest emerging eukaryotic lineage (Galtier & Gouy, 1995). However, curing an analysis from one source of error does not cure it from others. Microsporidian rRNAs also appear as quite long branches on SSU rRNA trees and are known to display many aberrant features (Cavalier-Smith, 1993). Moreover, a recent unpublished analysis (H. Philippe, pers. comm.) suggests that changes in the pattern of invariable sites in Microsporidia may also account for the deeply-branching position of these organisms in the SSU rRNA tree. Furthermore other recent analyses have suggested that many of the branching orders of eukaryotic groups are unresolved by this dataset (Kumar & Rzhetsky, 1996). Clearly, until all of these issues are resolved, one must be cautious in interpreting the placement of Microsporidia in the SSU rRNA tree.

The presence of a fused 5.8S/23S rRNA species in the Microsporidia has been suggested by some authors to be a retained prokaryotic feature in these organisms that betrays their early-branching position (Vossbrinck & Woese, 1986, Siddall *et al.*, 1992). However, as Cavalier-Smith has noted, microsporidian rRNAs tend to display a number of large unique deletions and it is possible that the processing site for splitting the large subunit rRNA into two species has been secondarily deleted in the Microsporidia (Cavalier-Smith, 1993a).

Similarly, the lack of mitochondria in Microsporidia has also been argued to be a primitive trait. However, there is a growing body of evidence that secondary loss of mitochondrial functions has occurred at least 7 times independently in eukaryotic evolution (see Chapter 3), probably as an adaptation to an anaerobic lifestyle. Thus, it is not particularly unlikely that the loss of

mitochondria has occurred in the ancestors of Microsporidia during the evolution of their intracellular parasitic habit.

While the observations discussed above do not specifically support an evolutionary link between Microsporidia and Fungi, several ultrastructural and molecular features are consistent with this relationship.

For instance, the recent discovery of U2 snRNA in Microsporidia (DiMaria *et al.*, 1996) is at odds with the lack of spliceosomal snRNAs noted in other deeply-branching groups such as the diplomonads (Niu *et al.*, 1994) and the lack of introns in both diplomonad and trichomonad genes (Roger *et al.*, 1994).

The absence of a Golgi dictyosome in Microsporidia was interpreted by Cavalier-Smith as a primitive feature shared with other archezoan groups such as diplomonads and oxymonads (discussed in Cavalier-Smith, 1993a). However, higher fungi such as ascomycetes, basidiomycetes and zygomycetes also appear to lack this organelle (Cavalier-Smith, 1987b). Placing the Microsporidia within the Fungi is therefore an equally parsimonious interpretation of the distribution of this character state.

Another possible link with the Fungi is the lack of flagella and centrioles during all stages of the life cycle of the Microsporidia, with an intranuclear spindle forming during mitosis and meiosis from nuclear plaques (Canning, 1990). Both of these features are also characteristic of the ascomycetes, basidiomycetes and some zygomycetes (Cavalier-Smith, 1987b). In addition, the meiotic cycle of Microsporidia, which was widely regarded as an aberrant, possibly primitive one-step process, has recently been argued to be fundamentally similar to some higher fungal meioses (Flegel & Pasharawipas, 1995). Moreover, the persistence of a diplokaryon state (where two nuclei are closely apposed in a single cell) in some Microsporidia is also often suggested to

resemble the dikarya observed both basidiomycete and ascomycete Fungi (Canning, 1990).

If the Microsporidia are truly related to Fungi, then it is most likely they arose from within the group sharing a common ancestor with ascomycetes, basidiomycetes and, possibly, some zygomycetes to the exclusion of chytridiomycetes. Species of the latter fungal group have centrioles, a well developed Golgi dictyosome, commonly possess flagella in some stages of their life cycles and are thought to be the most deeply branching of the fungal groups (Cavalier-Smith, 1987b, Bruns *et al.*, 1992, Berbee, 1993). Placing the Microsporidia on a common branch with the other eufungal groups would allow for a single event of loss of these features in their common ancestor after its divergence from chytridiomycetes. Characterization of tubulins and other genes from the full diversity of fungal taxa will be necessary to test this hypothesis.

The early phylogeny of eukaryotes.

If the Microsporidia are not the earliest branching eukaryotic group, then what group is? The diplomonads are consistently placed as the earliest emerging eukaryotic group in all of the EF-1 α trees obtained from each phylogenetic method when Microsporidia are removed from the dataset (Fig. 1.6). Although this placement is not strongly supported in EF-1 α analyses, a similar position for diplomonads is found in phylogenies of many other molecules including V-ATPases, RNA pol IIA, EF-2, hsp70 and SSU rRNA (see references cited in the Results). This abundant congruent evidence strongly suggests that diplomonads are truly the deepest eukaryotic lineage.

The placement of trichomonad flagellates as the next earliest branch is relatively well supported by both EF-1 α and β -tubulin trees (if one accepts the rooting on the diplomonad branch above). Once again there is congruency for

this placement with analyses of SSU rRNA (Gunderson *et al.*, 1995) and more recently the RNA pol IIA enzyme (Quon *et al.*, 1996). A link between these two groups has always been suspected because of the presence in some species of both groups of a tetrakont kinetid (found also in other groups such as oxymonads and retortamonads) and their shared lack of mitochondrial properties. However, this phylogeny does not agree with Cavalier-Smith's proposal that the Percolozoa (a large protist assemblage containing the Heterolobosea, psalteriomonads, and species of *Percolomonas* and *Stephanopogon*) are a deeper branch than the trichomonads (parabasalids) because of the absence in the former group of a Golgi dictyosome (Cavalier-Smith, 1993b). If we accept the root above, then both EF-1 α and β -tubulin show percolozoans (heteroloboseans in both cases) as a more recently emerging lineage, implying the secondary loss of a Golgi dictyosome in this group.

The deeply branching position of diplomonads and trichomonads has often been regarded as evidence that their amitochondriate nature is primitive. It is evidence in the form of a parsimony argument: since these organisms lack mitochondria and branch at the base of the eukaryotic tree, then it is most parsimonious to assume that mitochondria originated after their divergence. But, these parsimony arguments are rarely strong, and in Chapter 3, I describe strong evidence for the secondary loss of mitochondrial function in the trichomonads. In light of this and other evidence for secondary loss of mitochondrial functions, one should be cautious in interpreting the phylogenetic position of diplomonads as evidence for their archezoan nature.

Moving up the tree.

After the early branchings, the EF-1 α tree does not appear to strongly resolve either the monophyly of established protist groups nor their relative branching order. For instance the Ciliophora and the Heterolobosea are both protist groups for which there is much ultrastructural and molecular evidence (Hirt *et al.*, 1995, Page & Blanton, 1985, Roger *et al.*, 1996), yet their monophyly is not recovered by this molecule. For the Heterolobosea, there is evidence for multiple genes in each of the organisms studied and it is possible that one or several of the sequenced homologs are divergent developmentally expressed paralogs. Conclusions based on these sequences should be considered tentative until all of the homologs from each organism are characterized.

The explanation for the lack of resolution of the Ciliophora may have to do with an accelerated rate of evolution of the EF-1 α molecule in these organisms. Reduced use of actin in some ciliates (for example see Cohen, *et al.*, 1984) may have caused a corresponding reduction in constraints in the EF-1 α homologs of ciliates, since the protein is known to also function in actin-binding in some organisms (Condeelis, 1995). However, more data on the functions of EF-1 α in this group of organisms are needed before this can be considered anything other than speculation.

Despite the lack of strong resolution in the intermediate portion of the EF-1 α tree, several trends can be identified. There appears to be a relatively consistent association between the Euglenozoa, some of the Heterolobosea and the angiosperms. Occasionally these groups are united with an alveolate clade, containing the Ciliophora and the apicomplexan, *Plasmodium falciparum* (Fig 1.5B and 1.6A). Interestingly, a similar group containing the Euglenozoa, Heterolobosea, Viridiplantae, Alveolata and the Heterokonta is strongly recovered by all phylogenetic methods applied to the β -tubulin dataset. This

view of eukaryotic phylogeny is radically different than that which is obtained from SSU rRNA analyses. In SSU rRNA trees, the Euglenozoa and Heterolobosea are relatively early-branching independent groups, while alveolates, heterokonts and plants are part of a huge unresolved crown-like radiation that includes rhodophytes, metazoans, fungi and a variety of protist groups (Sogin, 1991). However, there are various lines of evidence, both from protein molecular phylogenies and ultrastructural data that favour some of the groupings found in the β -tubulin and EF-1 α trees. Below, I discuss these data and attempt to develop a picture of eukaryotic phylogeny based on a synthesis of all of the evidence. This hypothetical phylogeny is depicted in Figure 1.8.

Developing a consensus view of eukaryotic phylogeny.

As both O'Kelly and Patterson have argued, the presence of discoidal mitochondrial cristae in both the Heterolobosea (and other Percolozoa) and the Euglenozoa could be an ultrastructural synapomorphy uniting these two groups (O'Kelly, 1993, Patterson, 1994). Such a relationship is also supported by trees of α -tubulin (Li *et al.*, 1996), TPI (see Chapter 2) and iron superoxide dismutase (Fe-SOD) (see Chapter 4). The Fe-SOD and TPI datasets also show an association of this Heterolobosea/Euglenozoa clade with the apicomplexan, *Plasmodium falciparum*, providing evidence for a relationship of this group with the alveolate protists. V-ATPase trees also appear to weakly support such a relationship by grouping *Trypanosoma* and *Plasmodium* together (Gogarten *et al.*, 1996). The Euglenozoa and Heterolobosea are also shown to be related to green plants in trees of α -tubulin (Li *et al.*, 1996). Several other pieces of evidence support such a relationship. Trees of mitochondrial cpn60 show *Naegleria fowleri* grouping with plant homologs (Horner *et al.*, 1996) while the TPI gene from *Acrasis rosea*, shares a single intron position found only in plant enzymes (Chapter 2). The

euglenozoan, *Trypanosoma cruzi*, appears closely related to the green alga *Chlorella kessleri* in a recent analysis of EF-2 homologs (Nakamura *et al.*, 1996).

A connection between alveolates and green plants is suggested by the presence in *Plasmodium falciparum* of an enolase homolog bearing a highly conserved 6 amino acid insertion also found in green plant enolases but lacking in homologs from Metazoa, Fungi, the protist *Entamoeba histolytica* and bacteria (Hyde *et al.*, 1994). The common presence of cortical alveoli (membranous sacs located near the cell membrane) in alveolates and glaucophyte algae (such as *Cyanophora paradoxa*) may also support a link between alveolate protists and green plants, if Cavalier-Smith's hypothesis of a glaucophyte/plant/rhodophyte relationship is correct (Cavalier-Smith, 1987c). A green plant/glaucophyte relationship does appear to be supported by analysis of rRNA sequences (Van de Peer *et al.*, 1996) while V-ATPase trees depict a link between the green plants and rhodophytes (Gogarten *et al.*, 1996).

EF-1 α trees show a clade containing the Metazoa and the Fungi and this gene as well as β -tubulin sometimes shows a relationship between the Eumycetozoa (slime moulds) and this clade. These nested relationships are supported by two recent maximum likelihood analyses of multiple protein coding genes: Nikoh *et al.* (1994) used 23 separate protein genes to argue for a relationship between the animals and fungi while Kuma *et al.* (1995) report a relationship of this group to the cellular slime mould *Dictyostelium discoideum* by analyzing 19 protein datasets. Ribosomal RNA evidence has also been used to bolster the claim of a metazoan/fungus connection, identifying Choanoflagellates (Wainright *et al.*, 1993) and more recently Myxosporidia (Smothers *et al.*, 1994) as protist sisters of the Metazoa. If my foregoing arguments are correct regarding the placement of the Microsporidia within the

Fungi, then all of these groups together may represent a huge monophyletic assemblage.

The tree in Fig. 1.8 is an attempt to synthesize all of the evidence presented above as a tentative phylogenetic hypothesis. So far, many of the characters of "higher" eukaryotes have also been found in the diplomonad, *Giardia lamblia*. If diplomonads represent the deepest branch of eukaryotes then these features, shown in the box at the base of eukaryotes, must have evolved prior to the divergence of this group. Instead of the pattern of sequential branching of many independent protist lineages early in eukaryotic evolution as typically recovered by rRNA, this tree shows that after diplomonads, there are only a few early eukaryotic divergences, followed by a deep split into the two major superclusters of protists and multicells discussed above. Hopefully, as the taxonomic representation of rRNA and protein datasets improves, some elements of this hypothesis will come under test.

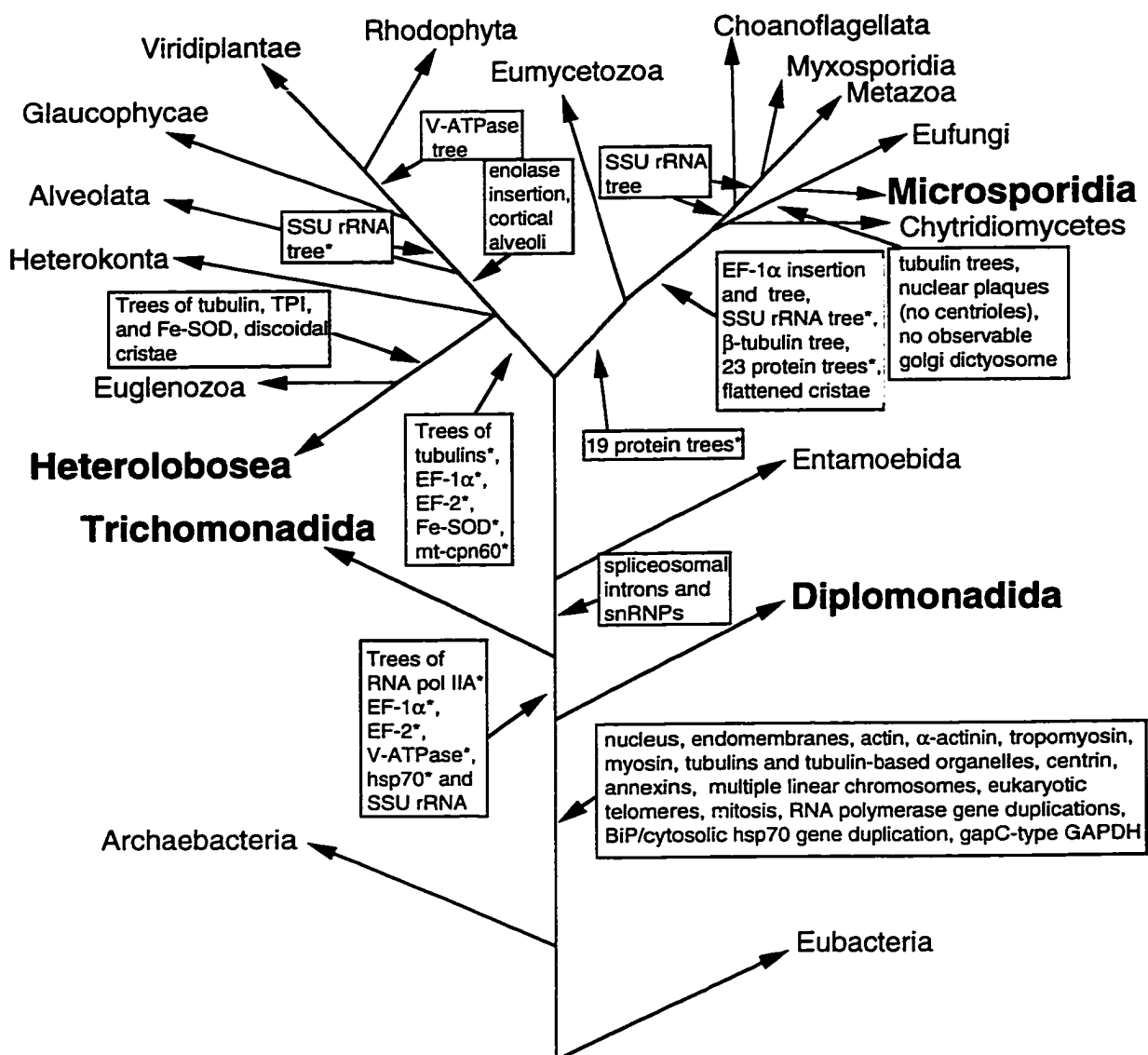


Figure 1.8. A hypothesis of relationships among eukaryotes based on selected molecular and ultrastructural data. Boxes show ultrastructural and molecular characters as well as molecular trees that support the branch indicated with an arrow. Asterisks (*) indicate that the phylogeny of the molecule supports the relationship of only some of the descendant taxa of the branch in question. Taxa written in bold text are ones for which the author has contributed molecular sequence data. Citations for molecular trees, other molecular characters and ultrastructural characters are as reported in the text except for data cited in the box at the base of the eukaryotes. These data come from Vickerman, (1989), M. Müller (pers. comm.), Klenk *et al.*, (1995), Fiedler & Simons, (1995), Gupta *et al.*, (1994), Drouin *et al.*, (1995), Henze *et al.* (1995), Rozario *et al.*, (1996), and Adam *et al.*, (1988)

Chapter 2

INTRODUCTION

The origin of spliceosomal introns is a classic molecular evolutionary puzzle that still inspires debate 19 years after their first discovery in eukaryotic genes, a debate that still centres on the initial polarization into two general theories on intron origins. The "introns-early" theory arose by combining Gilbert's ideas about how the exon/intron organization of eukaryotic genes could speed evolution by the process of "exon-shuffling" (Gilbert, 1978) with Doolittle's and Darnell's assertions that introns were *relics* of the assembly of genes in a primitive ancestor of all living cells (Darnell & Doolittle, 1986). According to this theory, modern prokaryotes lost their introns through genomic streamlining, while eukaryotes retained the primitive introneousness of their genes. By contrast, proponents of the "introns-late" models argued that most introns are likely the husks of mobile genetic elements that had the special property of splicing out of genes on the RNA level. Genes were split, the argument goes, by these mobile introns relatively recently after the origin of the eukaryotic nucleus and therefore had nothing to do with the origin of genes (Cavalier-Smith, 1991, Palmer & Logsdon, 1991).

Since the first proposals of the theories explaining intron origins, much has been learned about the diversity of intron splicing mechanisms and the genomes that introns inhabit. It is now clear that there are at least 5 different intron types, distinguishable by splicing mechanisms, their structure and phylogenetic distribution (reviewed in Lambowitz & Belfort, 1993). Of these types, only two are relevant to the original theories: group II self-splicing introns (and their degenerate group III form) commonly found in eukaryotic organellar genomes, and spliceosomal introns, which are abundant in nuclear genes and are spliced by a multimolecular RNA/protein complex called

the spliceosome. These demonstrate detailed similarity in structure and splicing mechanism of the sort that would suggest convincingly that they shared a common ancestor (Jacquier, 1990, Weiner, 1993, Lambowitz & Belfort, 1993, Lamond, 1993). Recently, the similarity noted between the catalytic core structure of group I and spliceosomal introns has been heralded as evidence that these intron types are also related (McPheeters & Abelson, 1992). However, at this level of similarity it is difficult to distinguish homology from forced moves in the evolution of an intron splicing mechanism. Until stronger evidence is found for homologous similarity between intron types, it is best to regard the origin and evolution of these elements as separate questions.

This study is specifically concerned with the issues surrounding the origin and evolution of spliceosomal introns. Over the years there have been literally hundreds of proposals of different evolutionary scenarios relating to the origin and function of spliceosomal introns and it is practically impossible to discuss the relative merits of all of them. Instead, I shall try to summarize and evaluate the central claims of the two lineages of proposals mentioned above, which have been dominant in the literature for nearly two decades.

The central claims of the theories.

The introns-early theory, often referred to as the "exon theory of genes", makes two central assertions: (1) exon-shuffling was the dominant mode of evolution in the formation of the first genes prior to the divergence of all extant life; and (2) *some* or *all* spliceosomal introns are ancestral to the genes in which they are found. According to this view, the subsequent evolution of introns has been influenced by two forces: the widespread loss of introns and their movement to neighbouring positions by "sliding" or "displacement" mechanisms (Gilbert, 1978, 1987, Doolittle, 1978, Doolittle & Darnell, 1986, Martinez *et al.*,

1989, Liaud *et al.*, 1992, Kersenach *et al.*, 1994). Although the possibility of intron insertion was discounted or ignored by advocates of introns-early for many years, some have recently conceded that insertion may occur rarely (de Souza *et al.*, 1996).

By contrast, the introns-late (insertional or recent origin) theory of spliceosomal intron evolution claims that: (1) Spliceosomal introns originated in the eukaryotic nuclear lineage; (2) Spliceosomal introns have spread within this lineage by inserting into previously unsplit genes (Cavalier-Smith, 1985, 1991, Rogers, 1989, 1990, Logsdon & Palmer, 1991, Roger *et al.*, 1994, Mattick, 1994). Intron loss is also acknowledged as a force operating in eukaryotic genome evolution (Rogers, 1990, Cavalier-Smith, 1991, Palmer & Logsdon, 1991, Roger *et al.*, 1994).

Evaluating the evidence from protein structure.

The first claim of introns-early has relied on several kinds of evidence and has fueled much of the debate on intron origins. Firstly, some authors (for instance, Gilbert, 1987) have drawn attention to examples of exon-shuffling that has occurred to produce many of the vertebrate extracellular proteins (Pathy, 1991). The origins of these genes, however, occurred relatively recently in evolution, perhaps just prior to the divergence of extant chordates 500 million years ago. Another example of exon-shuffling has recently been shown for some plant enzymes (see de Souza *et al.*, 1996), but again, the events in question have clearly occurred within the land plant lineage. These examples are actually irrelevant to the debate over intron evolution, since the central claim of introns-early is not that such shuffling has occurred sometime in evolution, but that it was the dominant mode of evolution prior to the last common ancestor of all life. To prove such a claim, one must demonstrate that some genes common to all

living organisms are chimaeric and that the recombination junctions between the protein segments correspond to present-day intron positions. Examples of ancient exon-shuffling have been suggested for proteins such as alcohol dehydrogenase (ADH), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), lactate dehydrogenase (LDH), pyruvate kinase (PK)(Duester *et al.*, 1986) and triose phosphate isomerase (TPI) (Gilbert *et al.*, 1986). However, as Patthy points out, such proposals seem dubious as these proteins do not show the hallmarks of exon-shuffling discerned from known recent vertebrate examples. For instance, one would expect that exon-shuffling would produce genes sharing homologous exons bounded by introns of the same phase (one of three possible positions in codons): a pattern not observed in these genes (Patthy, 1987). Furthermore, the paucity of this kind of evidence was highlighted by an attempt by Gilbert's group to estimate the size of an underlying "exon universe" that failed to detect any examples of exons shared among ancient genes (Dorit *et al.*, 1990).

A second, less direct, form of evidence for ancient exon-shuffling is the claim that in ancient genes, the boundaries of protein structural elements correlate with intron positions, a result expected if these elements were recombined to create the gene (Blake, 1983). For many of the suggested examples, however, the correlations have only been suggested qualitatively. A recent analysis testing several suggested correlations for TPI, PK, globins and ADH, found that, with one exception, random intron positioning with respect to protein structure could not be statistically excluded (Stoltzfus *et al.*, 1994). The exceptional case, where boundaries of compact protein modules correlate with intron positions in TPI, was recently rendered insignificant by the discovery of a wealth of new intron positions in homologs of this gene (Logsdon *et al.*, 1995).

Despite the general lack of direct evidence for the ancient exon-shuffling tenet of the introns-early theory, proponents continue to defend it, bolstering

their case with a recent global database analysis of intron phases. Patthy's demonstration that symmetrical exons (exons bounded by introns of the same phase) are the substrate and product of exon-shuffling in the evolution of vertebrate blood proteins, prompted Long and colleagues to investigate this phenomenon on a larger scale (Long *et al.*, 1995). Their study revealed that exons bounded by introns of the same phase are represented in the database in greater proportion than would be expected by random chance. Examination of the database of proteins showing ancient conserved regions (ACRs) also showed a similar tendency of exons to be symmetrical. Long *et al.* (1995) have argued that these latter observations are strong evidence for ancient exon-shuffling. However, depending on how the boundaries of ACRs were evaluated, this pattern could be the result of exon-shuffling and duplications within these regions that occurred recently in eukaryotic evolution. Even if this is not the case, the observation is not uniquely explained by exon-shuffling; it could be the result of non-random intron insertion or some weak selection pressure against consecutive introns possessing different phases. Thus, in the absence of any other evidence for ancient exon-shuffling, these observations can only be considered weak support for introns-early. Defenders of this theory have increasingly relied on evidence for its second proposition, that some or all introns in eukaryotes are ancestral to genes.

The phylogenetic distribution of intron positions in genes.

Early on, it was suggested that the presence of shared intron positions between animal, fungal and plant TPI genes indicated that introns in these positions antedated the divergences of these organismal groups and were therefore ancient (Gilbert *et al.*, 1986). More recently, several other instances of intron conservation at this level have also been elaborated (Rogers, 1990,

Stoltzfus & Doolittle, 1993, Logsdon *et al.*, 1995, de Souza *et al.*, 1996). However, introns-late proponents have disputed the relevance of this claim, since this pattern is also expected if the introns originated by events of insertion in the common ancestor of these groups. Although introns-late defenders have repeatedly and emphatically made arguments such as these in the literature (Rogers, 1990, Cavalier-Smith, 1991, Stoltzfus and Doolittle, 1993, Roger *et al.* 1994, Logsdon *et al.*, 1995), a recent review by Gilbert's group grossly distorts the introns-late position by claiming that it does not acknowledge the possibility of Precambrian intron origin (pg. 498, de Souza *et al.*, 1996). Despite this bizarre, misleading claim, introns-early workers have drawn attention to more impressive instances of intron conservation that warrant attention. These constitute examples where intron positions are shared between cytosolic and organellar homologs of genes. For instance, five coincident introns are observed in mitochondrial and cytosolic aspartate aminotransferase (AspAt) homologs (Juretic *et al.*, 1990), two are shared between mitochondrial and cytosolic malate dehydrogenase (MDH) homologs (Iwabe *et al.*, 1990) and five are shared between plastid and cytosolic glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes (Kersenach *et al.*, 1994). The case for AspAt is now discredited by the finding that cytosolic and mitochondrial homologs are immediate relatives, suggesting a recent common ancestry rather than ancient divergence at the prokaryotic/eukaryotic level (J. R. Brown, pers. comm.). However, for GAPDH it does appear that ancient divergences between the homologs have occurred. Kersenach *et al.* calculated the probability of these five GAPDH introns matching if they were chance multiple insertion events and concluded that this scenario was vanishingly unlikely (Kersenach *et al.*, 1994). However, once again, counter-arguments to these examples have been advanced. Stoltzfus and Roger *et al.* objected to the implicit assumption in the calculation by Kersenach *et al.* that all

positions in the gene are equally likely to experience intron insertion (Roger *et al.*, 1994, Stoltzfus, 1994). As with any transposable element, spliceosomal introns are likely to require a targeting sequence, and if the gene into which they are inserted is highly conserved (like the GAPDH gene), then the effective number of interruptable sites may be much smaller than the total number of nucleotide positions in the gene. On the other hand, Logsdon and Palmer have argued that once the phylogenetic distribution of these GAPDH introns was taken into account, the presence of these five introns in the common ancestor of plastid and cytosolic genes requires extensive parallel loss of introns in multiple eukaryotic and bacterial lineages, an extremely unparsimonious scenario (Logsdon & Palmer, 1994).

Arguments such as these have been more fully developed in the analysis by Logsdon *et al.* (1995) and Kwiatowski *et al.* (1995) of the phylogenetic distribution of introns in the TPI gene. Again from parsimony considerations, they argue that none of the introns observed is likely to be ancestral. As in the GAPDH example, the argument is that most intron positions are phylogenetically restricted to recently-evolved eukaryotic lineages. The suggestion that they were ancestral requires one to accept that multiple events of loss occurred in lots of independent lineages. However, if instead one postulates a loss and gain scenario where no introns were ancestral, many fewer events are required to explain the present-day intron distribution. Similar arguments were made several years ago by Dibb and Newman in evaluating the intron distribution in actin and tubulins as well as more recently by Dietmaier and Fabry who examined small G-proteins (Dibb & Newman, 1989, Dietmaier & Fabry, 1994).

Objections like these have prompted several replies from the introns-early camp. In defending their arguments made in relation to the GAPDH data, Cerff

and colleagues argue that Palmer and Logsdon failed to recognize that intron loss may be more likely than intron gain (Cerff *et al.*, 1994). In addition, de Souza *et al.* have disputed the claim that events of intron insertion require a targetting sequence, pointing to instances of intron gain in U6 snRNA genes that do not appear to require any particular flanking sequence (de Souza *et al.*, 1996).

Several additional general objections to the parsimony arguments of the introns-late camp have been advanced. Doolittle (pers. comm.) and Hurst (Hurst, 1994) have argued that if introns were ancestral to genes and enough loss occurred then it is expected that one will find introns retained only in isolated lineages: the exact observation that introns-late advocates argue is evidence for insertion. Moreover, de Souza and colleagues suggest that the debate over the phylogenetic distribution of introns is fundamentally irreconcilable since both sides can equally explain the present-day distribution of introns by invoking their own models of intron evolution (de Souza *et al.*, 1996).

The problem with parsimony.

Both of these latter objections expose a fundamental flaw in the parsimony arguments for a recent origin for introns. A general form of this parsimony argument can be stated as follows: if two hypotheses explaining the data are under test, the one that posits the fewest events to explain the data should be preferred. This argument is often justified on philosophical grounds. Hypothetico-deductivists claim that this parsimony approach is connected to Popperian falsification, arguing that the most parsimonious hypothesis (the one that invokes the least number of events) is the one least falsified by the data. Such a view is also claimed to be connected to the global scientific principle of parsimony, Ockham's razor, whereby the simplest hypothesis consistent with the data should always be preferred. Both of these justifications suggest that the use

of parsimony in hypothesis evaluation is an approach free of assumptions about how evolution proceeds. However, neither argument stands up to scrutiny. Sober (1988) has argued that extra events of character evolution are always possible explanations of the data and therefore hypotheses that postulate more than the minimum number of events cannot be falsified by them. Furthermore, he argues that neither the global principle of parsimony (Ockham's razor) nor the local one applied to systematics is assumption-free. To illustrate his point, consider an example from the intron debate. If intron loss were truly far more frequent than intron gain, then why should these events count equally in a parsimony calculation? Under this scenario, perhaps one would expect that many events of intron loss occurred for every event of intron gain and that the most parsimonious hypothesis was wrong. Yet the parsimony arguments advanced by Logsdon *et al.* (1995) and a former time-slice of this author (Roger *et al.*, 1994) rely on the assumption that both intron loss and gain events count equally. It is impossible to compare two different models of intron evolution using parsimony if they have different implications for the relative probability of the kinds of events being counted.

A likelihood approach to intron evolution.

So, are de Souza *et al.* correct in their view that it is impossible to choose between the two alternative hypotheses of intron evolution when considering the phylogenetic distribution of introns? Consideration of the principle of likelihood shows that they are not. This principle states that the likelihood of a hypothesis is the probability of observing the data if the hypothesis were true (Edwards, 1972, Sober, 1988). One can use this principle to choose between different models of intron evolution by considering which one confers a higher probability on the observed phylogenetic distribution of introns. Likelihood has been used

extensively in molecular systematics to infer what phylogenies confer the greatest probability on the distribution of sequences (Felsenstein, 1981, Goldman, 1990). In this Chapter, I develop an approach to the question of intron evolution by employing stochastic models of intron evolution coupled with a phylogeny of the genes containing them to calculate the probability of observing a particular intron distribution.

Developing explicit models of intron evolution in GAPDH and TPI genes.

In this study, I have chosen to focus on GAPDH and TPI, two glycolytic genes that have figured prominently in the debate over intron evolution. In both cases, introns-early advocates have suggested that there are ancestral introns retained in the modern homologs of these genes.

An introns-early model that applies to these genes can be broken into two different models. As mentioned above, for many years introns-early advocates denied the possibility of intron insertion. Instead the forces of intron loss and sliding (or displacement) were postulated to explain the present-day distribution of introns. This view, which holds that *all* introns are ancestral to the gene, I will refer to as the hard introns-early model. More recently, Gilbert's group has accepted that intron gain does occur, and that the forces determining the intron distribution are a combination of intron loss, sliding and gain (de Souza *et al.*, 1996). This soft introns-early model differs from an introns-late model in claiming that at least *some* introns are ancestral to genes. By contrast, an introns-late model is based on the view that introns are gained and lost, but does not see a role for intron-sliding and argues that no introns are ancestral to the gene.

Of the various forces governing intron evolution, the processes of intron sliding and displacement are currently the most controversial. The former mechanism involves movement of splice boundaries on the DNA level while the

latter involves the reintroduction (by reverse splicing) of recently spliced introns at neighbouring sites on the RNA-level. These processes were originally invoked because introns in homologs were sometimes observed to occupy similar but not identical positions in the gene. If both of these introns were ancestral, then the exons they bounded could only have been a few bases in length, too small to encode shufflable units of protein structure. So instead, these neighbouring introns are viewed by introns-early proponents as homologous, having moved to slightly different positions during evolution. However, direct evidence for intron sliding, in the form of clearly homologous introns in non-identical positions, has never been observed. Moreover, in an attempt to evaluate the evidence for sliding in TPI, Logsdon *et al.* tested whether introns were significantly clustered as one would expect if intron sliding were a *bona fide* phenomenon. They found that the size distribution of exons in this gene does not deviate from the exponential distribution expected from random intron placement (Logsdon *et al.*, 1995). Comparable tests carried out on the GAPDH dataset had similar results (Roger, Yang & Doolittle, unpublished data). This suggests that for at least TPI and GAPDH datasets, intron sliding is not a detectable phenomenon. For this reason, I have chosen not to explicitly incorporate this process into the intron models. It should be noted, however, that sliding is actually a special case of intron gain and loss, whereby neighbouring positions are correlated. Thus, the gain and loss models do account for this process in an indirect way, although the assumption that all sites are independent is violated in this case.

To model intron evolution, I have used the program DNAML developed for evaluating the likelihood of phylogenetic trees of DNA sequences (see the Materials and Methods for a more complete description of this method). The model in this program allows for transitions between all four bases in DNA, with special provision for differing rates for transitions and transversions as well as

different base frequencies. To evaluate intron data, the model was reduced to considering two character states: intron presence and absence coded as purines and pyrimidines respectively (R and Y). In this model, each of these states can change to the other with a particular ratio of rates (which is defined by the ratio of the frequencies of the two character states). In addition, the branch lengths of the tree are individually optimized to maximize the probability of the data. The hard introns early model was simulated by setting the ratio of the rate of gain to loss very close to zero. Soft introns-early and introns-late models were evaluated by increasing this ratio over a range of values. At the optimal gain:loss ratio, these models can be discriminated between by maximum likelihood reconstruction of ancestral sequences.

In order to improve the taxonomic representation of the GAPDH and TPI datasets, I sought to obtain homologs of these genes from several independent, putatively early-branching eukaryotic lineages. For GAPDH, homologs were obtained from two heteroloboseans, *Naegleria andersoni* and *Acrasis rosea* in addition to one microsporidan, *Nosema locustae*. Homologs of TPI were cloned from one heterolobosean, *Acrasis rosea*, the diplomonad 50380, *Entamoeba histolytica* and the trichomonad, *Trichomonas vaginalis*.

RESULTS

New TPI and GAPDH sequences.

TPI products were obtained using the TF-1/TR-1 primer pair from *Acrasis rosea*, *Trichomonas vaginalis* and *Entamoeba histolytica*. This pair along with TF-2/TR-1 variant designed against the *Giardia lamblia* homologs were used on the diplomonad 50380 DNA. A product corresponding to TPI was obtained only with the latter pair. Products from *T. vaginalis*, *E. histolytica* and diplomonad 50380 were translated and entered into an alignment with 39 other homologs (a

portion of this alignment is shown in Fig. 2.1). Each encoded a protein colinear with other TPI genes in the alignment, indicating their lack of introns. The sequence of the TPI product from *Acrasis rosea*, by contrast, appeared to be interrupted by a single intron 38 bases in length possessing typical GT-AG spliceosomal intron boundaries. This phase 0 intron occurs between a cysteine and an asparagine codon in alignment positions 14 and 15, exactly matching the position of the first intron found in all land-plant genes (Fig. 2.1). Although there is no direct evidence that this is an intron (i.e. there is no cDNA sequence), its removal restores the reading frame and allows precise alignment of the amino acid sequence.

The GAPN and GAPC primers generated PCR products from *A. rosea*, *N. andersoni* and *N. locustae* that were clearly homologous to GAPDH genes from other organisms and covered more than 90% of the coding region. The inferred amino acid sequences of these products were free of stop codons and aligned with a minimum of gaps to GAPDH sequences from 87 other organisms suggesting that none of them was interrupted by introns. A selection of sequences from this GAPDH alignment is shown in Fig. 2.2.

Compilation of intron datasets.

The alignments of TPI and GAPDH were made up of sequences of genomic DNA clones where the presence or absence of introns can be determined. The intron positions for TPI have recently been compiled by Logsdon *et al.* (1995) and at this time, these 21 positions remain the only ones identified in homologs of this enzyme. For each sequence in my alignment the presence or absence of a particular intron position was determined by inspection of the GenBank entry and by reference to Logsdon *et al.* (1995). These 21 positions were converted into intron sequences for parsimony and likelihood

```

                                                    75
d80 MN-GTIK---FINDHAEVLKSIKNN--VEVVMAPTALHASLLQHLK----DSHVCVAAQN
T.va ANPKTVE---EAEKLIEMLNKAKVEGN-VEVVVAAPIFLPTLQOKLR-----KDWKVSAEN
E.hi CN-GTLASITETLTKGVAASVDAELPKK-VEVTVGVVFFIYIPKVQQLAGEANGANILVSAEN
A.ro CN-GTQE---SVDKLVKILNDAKDVDGKIDVVVAPTLIHLAKVHESLRK-----DFHVSAQN
T.br MSKQPPIAAANWKN--GSQQ---SLSELIDLFNSTSINHD-VQCVVASTFVHLAMTKERLS----HPKFVIAAQN
E.co MRHPLVMGNWKN--GSRH---MVHELVSNLKRELKAGVA-CAVAIAPPEMYIDMAKREAEGS----HIMLGAQN
B.bu MRKTFLAGNWKMH-YTSA---EASIVAKKIATEVKTLLKDVVIMITPPFTALSKVSECIKGS----NILLGAQN
                                                    150
d80 VYDQKPGAFTGELAVEMLVDAGIKYAIIGHSERRRIMGESNEQSAKKTLRALE-ANITVLFICIGETLEERNANKV
T.va VFTKPDGAFTGEVTVPMIKSFGIEWTILGHSERRDILKEDDEFLAAKAKFALE-NGMKIICYCGEHLSEREAGKA
E.hi AWTKS-GAYTGEVHVGMVLDVCQVPVYVILGHSERRQIFHESNEQVAEKVKVAID-AGLKVIAICIGETEQAQRIANQT
A.ro -FVAESGAYTGEVSVSMLKIDIGLHYAIVGHSERRSLYHETDEVAHKKVAVD-AGLTAIACIGETLQERENKT
T.br AIAKS-GAFTGEVSLPILKDFGVNWIIVLGHSERRAYYGETNEIVADKVAAVA-SGFMVIAICIGETLQERESGRT
E.co VNLNLSGAFGTGETSAAMLKDIGAQYIIGHSERRTYHKEDELIAKKFAVLKE-QGLTPVLCIGETEAEAEAGKT
B.bu MSYMESGARTSEISPSMLLEFGVEYVILGHSECRLYLAETDEIINKILAGLKHPPFKYLILCVGETLDERDSGKT
                                                    225
d80 DEVNFAQLAALKAVITPQQWVDVVIAYEPVWSIGTGCVVASPEQAQEVHASIRNWLKKEISTEVAEMTRIQQGGSV
T.va SEFVSAQIEKMIPAIPAGRWDDVVIAYEPIWAIIGTGKVVASTQDAQEMCKVTRDILAAKVGADIANKVRILYGGSV
E.hi EEVVAQKKAINNAI SKEAWKNI ILAYEPVWAIIGTGKTATPDQAQEVHQYIRKWMTENISKEVAEATRIQQGGSV
A.ro NEVTTRQLQAYANVIKD--WDKVVIAYEPVWAIIGTGKVVATPDQAQVHADLREWLKRNVEEVADKVRILYGGSV
T.br AVVVLTIQAATAAKLKKADWAKVVIAYEPVWAIIGTGKVVATPDQAQEAHALIRSWVSSKIGADVAGELRILYGGSV
E.co EEVCARQIDAVLKTQGAFAFEGAVIAYEPVWAIIGTGKVSATPAQAQAVHKFIRDHIAKV-DANIAEQVLIQQGGSV
B.bu LEVVLNQVKKGLNVCSESDIQRIILAYEPVWAIIGTGKTATKKEAEVHKAIRLEITKLYTKSASDNIIIQYGGSV
                                                    268
d80 NGKNCAELSKCADIDGFL
T.va KPNNCNELAACPDVDGFL
E.hi NPANCNELAKKADIDGFL
A.ro KGDNAEVLIKEKDIDGFL
T.br NGKNARTLYQQRDVNGFLVGGASLKPEFVDIIKATQ*
E.co NASNAAELEFAQPDIDGALVGGASLKADAFIVKAAEAQA*
B.bu NSNNVKELMNEPNIDGALIGGASLKAESFLSIINNVL*

```

Figure 2.1 An alignment of TPI sequences obtained in this study with homologs from other organisms. The species name abbreviations are d80: diplomonad 50380; *T.va*: *Trichomonas vaginalis*; *E.hi*: *Entamoeba histolytica*; *A.ro*: *Acrasis rosea*; *T.br*: *Trypanosoma brucei*; *E.co*: *Escherichia coli*; *B.bu*: *Borrelia burgdorferi*. The bolded taxon labels indicate the sequences obtained in this study. Asterisks (*) indicate stop codons. The sequences in the alignment are selected from a full alignment of 43 TPI homologs.

75

N.an GRLVFRASLERTDVEIVAINDIMMTPPEYMIYMIKYDVTVHGKFGHGK-LEYT-DKSIIVNG
A.ro GRLVMRASLERDDVEIVGVNDIMLDPKYMAYLFKYDSVHGTFKGT-VDFK-EGALIVNG
N.lo GKIVYQVFKRNIRVSV-INDPFAKPEDIEYALKYDITTFGRSGAK-VHRS-GNRVTVGD
U.ma MSQVNIGINGFGRIGRIVFRNSVVENTANVVAINDPFDLEVMVYMLKYDSTHGVFNGD-ISTK-DGKLIIVNG
G.la MPIRLGINGFGRIGRMALRASLNIDGVQVVAINDPFTDCEYMEYMLKYDITVHGRFDGT-IAHS-EDSITVNG
Z.ma4 MAKIKIGINGFGRIGRIVARVALQSDDELVAVNDPFISTDYMTYMFKYDITVHGQWKHHEVKV-DSKTLLFG
H.va MMSEFVRVGLNGFGRIGRNVFRASLHSDDVEIVGIND-VMDDSEIDYFAQYDSVMGELEGA-SVDDGVLTVDGTD
150

N.an RE-VHVLCDERDPEQLPWGHHGVEYVVESTGIFTKLDTASKHLKGGAKRVVISAPA----DTPTFMVGNHHEYKP
A.ro LE-TKVFAEKEPSKLPWGLKVDYVVESTGIFLDDKKSCEEHLKGGAKRVVISAPAK---DDTPMYVGVNEDTYS
N.lo IE-TKILSERSPANINWE--NADVVEASGVFLTLDCEGHLNT-ARRVITAPSP---NASMYVGVNHCEYKG
U.ma KS-IAVFAEKDPSNIPWQAGAHYVVESTGVFTTIDKASAHIKGGAKKVVISAPSA---DAPMYVCGVNLDAYDP
G.la NK-ISVFKSMKPEEIPWGTQVDIVLECTGRFTTKDAELHITGGCKRVIISAPSA---DAPMFVCGCNLETYDP
Z.ma4 EKEVAVFGCRNPEEIPWGSVGAEYVVESTGVFTDQEKAAHLKGGAKKVVISAPSK---DAPMFVGVNEKEYKS
H.va FE-AGIFHETDPTQLPWDDLVDVAFEATGIFRTKEDASQHLDAGADKVLISAPPKGDEFPVKQLVYGVNHDEYDG
225

N.an E-MTVINNASCPTNCLAPIASVLEHENFGLLEGLMTTVHAVTATQPTVDAPSKKDWRRGGRAGYNIIPSSSTGAAGA
A.ro G-QTVISNASCPTNCLAPLASHDKYTIIEGLMTTVHATTATQKTVDGPGQRGDWRFRGGAAFNIIPASTGAARA
N.lo E--RIISNASCPTNCLAAIAKVVHESFGLIEGLMTTVHATTNSQRAVDTCIAK--RTRKRC-FNIIPASTGAAGA
U.ma K-AQVVSNASCPTNCLAPLAKVIHDKFGIVEGLMTTVHATTATQKTVDGPPSAKDWRRGGRAAAANIIPSSSTGAAKR
G.la STMKVVISNASCPTNCLAPLAMVVKFKGIEGLMTTVHAVTATQLPVDGPPSKKDWRRGGRSCGANVIPSSTGAAGA
Z.ma4 D-INIVSNASCPTNCLAPLAKVINDKFGIVEGLMTTVHATTATQKTVDGPPSKKDWRRGGRASFNIPSSSTGAAGA
H.va ED--VVSNASCPTNSITPVAKVLDEEFGINAGQLITVHAYTGSQNLMDGP?NGKPRRRRAAAENIIPSTGAAQA
300

N.an VGLVIPSLNGKLTGMAFRVPTADVSVVDLTCRLEKPATKKQIDEAMKKAESERFKGILKYTDDEEVVSSDFVHDS
A.ro VGSVIPSLKGLTGMSFRVPTSDVSVVDLTVRIEKGANKQEIDKTLKEAANSERWKNIFAYTDDDDVSTDFIHDH
N.lo LSKVIPTLEGKMTGMAFRVFPVNVSVVDLTVRLEKASLEATLEKVKNAK-GEMKGVLCYTEDEVVSGDYNGCS
U.ma VGKVIPSLNGKLTGMAFRVPTTNSVVDLTARLEKASYSDEIKAEVKRASE-NELKGIILGYTEDAVVSQDFIGNS
G.la VGKVLPALNGKLTGMAFRVFPVVDVSVVDLTCLEKDATYDEICAEIKRGE-NELKGIIMTYTNEDEVVSSDFLSTT
Z.ma4 VGKVLVPLNGKLTGMSFRVPTVDVSVVDLTVRLEKSATYDEIKAAVKEEAE-GSLKGIILGYVEEDLVSTDFQGD
H.va ATEVLPELEGLDGMARVFPVNGSITFEVVDLDDDVTESDVNAAFEDAAA-GELEGLVGVTSDDVVSSDILGDP
348

N.an ASSTYDSKASISLNDNFVKVVA
A.ro HTSTYDSNASIFLNDNFIKLIA
N.lo LSCIFDYKASIALNDKFFKLV
U.ma HSSIFDAAAGISLNNNFVKLVSWYDNEWGYSN-RCLDLLVFMAQKDSA*
G.la STCNFDSKAGIMLSRFVKLVAWYDNEFGYAN-KLVELAKYVGSKGC*
Z.ma4 RSSIFDAKAGIALNGNFVKLVSWYDNEWGYSSTRVVDLIRHMNSTN*
H.va YSTQVDLQSTNVVSG?MTKILLTWYDNEYGFSN-RMLDVAEYITE*

Figure 2.2 An alignment of GAPDH sequences obtained in this study with homologs from other organisms. The species name abbreviations are *N.an*: *Naegleria andersoni*; *A.ro*: *Acrasis rosea*; *N.lo*: *Nosema locustae*; *U.ma*; *Ustilago maydis*; *G.la*: *Giardia lamblia*; *Z.ma4*: *Zea mays* gapC4; *H.va*; *Haloarcula vallismortis*. The bolded taxon labels indicate the sequences obtained in this study. Asterisks (*) indicate stop codons. The sequences in the alignment are selected from a full alignment of 90 GAPDH homologs.

analyses. In this study, all positions in the gene were considered interruptable by introns. To develop the intron position sequences, 248 codons of the alignment were used extending from alignment position 3-260, yielding a total of 744 positions in the dataset (several alignment gaps were not included in this dataset). For each TPI homolog, interrupted and uninterrupted positions were enumerated to produce the intron position sequence. Positions lacking some of the sequences were coded as missing information in the datafile. All positions possessing an intron in at least one taxon are shown in Fig. 2.3.

Intron positions for GAPDH were compiled in a similar manner by referring to Kersenach *et al.* (1994), by inspection of GenBank entries and consultation with J. Logsdon (pers. comm.). Since the publication of this paper, two new intron positions have been published: one identified in the slime mould, *Dictyostelium discoideum* (Roger *et al.*, 1996) and another in the rhodophyte alga, *Gracilaria verrucosa* (Zhou & Ragan, 1995). Upon inspection, several of the distinct positions identified by Kersenach *et al.*, appeared to be located at the same position in the alignment used in this study. Table 2.1 shows the comparison of intron positions from Kersenach *et al.* (1994) with the positions used in this study. In total, 46 distinct positions occupied by introns were identified and are shown in Fig. 2.4. Intron position sequences for GAPDH homologs were created as described above. For this dataset a total of 340 codons corresponding to 1020 interruptable positions were used to develop intron position sequences. These positions extend from position 3 to 344 in the GAPDH alignment (one alignment gap position was not included in this dataset). Missing positions in some sequences were coded as missing information as above.

Species	Intron positions																					
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2
<i>H. sapiens</i>	-	-	+	-	-	+	+	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-
<i>P. troglodytes</i>	-	-	+	-	-	+	+	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-
<i>M. mulatta</i>	-	-	+	-	-	+	+	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-
<i>G. gallus</i>	-	-	+	-	-	+	+	-	-	+	-	-	+	+	-	-	-	-	-	-	-	-
<i>C. tarsalis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>D. melanogaster</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
<i>H. virescens</i>	-	-	+	+	-	-	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	??
<i>C. elegans</i>	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-
<i>S. mansoni</i>	-	-	+	-	-	+	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+
<i>A. nidulans</i>	+	-	-	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	-	+
<i>S. cerevisiae</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. pombe</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. cinereus</i>	-	-	+	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+
<i>Z. mays</i>	-	+	+	-	-	+	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-	+
<i>O. sativa</i>	-	+	+	-	-	+	+	-	-	+	-	-	+	-	-	-	-	-	-	-	-	+
<i>G. verrucosa</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>P. falciparum</i>	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>E. histolytica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	??
<i>A. rosea</i>	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	??
<i>T. brucei</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. cruzi</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>L. mexicana</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. vaginalis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	??
<i>G. lamblia</i> GS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>G. lamblia</i> WB	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dip. 50380	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	??
<i>B. megaterium</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>B. stearothermophilus</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>B. subtilis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>L. lactis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. glutamicum</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>M. leprae</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. maritima</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>E. coli</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>H. influenzae</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>V. marinus</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>V. sp.</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. sp.</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>B. burgdorferi</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>M. flocculare</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>M. hyorhinis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>M. sp.</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>P. woesii</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 2.3 Positions occupied by introns in TPI homologs. Intron position numbers correspond to the positions identified by Logsdon *et al.* (1995). Pluses (+) indicate presence of an intron at that position in the sequence, hyphens (-) indicate the lack of an intron at that position and question marks (?) denote missing sequence data at the position.

Species	Intron positions																																							
	1111111111							2222222222							3333333333							44444444																		
	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6				
<i>H. vallismortis</i>																																								
<i>P. sativum gapA</i>												+																												
<i>A. thaliana gapA</i>																																								
<i>Z. mays gapA1</i>																																								
<i>P. sativum gapB</i>																																								
<i>A. thaliana gapB</i>																																								
<i>C. reinhardtii gapA</i>																																								
<i>G. verrucosa gapA</i>																																								
<i>C. crispus gapA</i>																																								
Syn. sp.gap1																																								
Syn. sp.gap2																																								
<i>A. variabilis gap1</i>																																								
<i>A. variabilis gap2</i>																																								
<i>A. variabilis gap3</i>																																								
<i>B. megaterium</i>																																								
<i>B. stearothersophilus</i>																																								
<i>B. subtilis</i>																																								
<i>C. glutamicum</i>																																								
<i>C. pasteurianum</i>																																								
<i>S. aureofaciens</i>																																								
<i>C. freundii</i>																																							??	
<i>E. coli gapA</i>																																								
<i>E. coli gapB</i>																																								
<i>E. coli gapC</i>																																								
<i>S. typhimurium</i>																																							??	
<i>Sal. sp.</i>																																								
<i>S. marcescens</i>																																							??	
<i>Ser. sp.</i>																																							??	
<i>E. aerogenes</i>																																							??	
<i>K. pneumoniae</i>																																							??	
<i>H. influenzae</i>																																								
<i>Z. mobilis</i>																																								
<i>R. sphaeroides</i>																																								
<i>T. thermophilus</i>																																								
<i>T. maritima</i>																																								
<i>T. vaginalis</i>																																								
<i>N. locustae</i>																																							??	
<i>G. lamblia</i>																																								
dip. 50380																																							??	
<i>T. agilis</i>																																							??	
<i>H. inflata</i>																																							??	
<i>E. histolytica</i>																																							??	
<i>N. andersoni</i>																																							??	
<i>A. rosea</i>																																							??	

Figure 2.4 (continued on the next page)

Species	Intron positions																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
<i>T. borreli</i> gap1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. borreli</i> gap2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. cruzi</i> gapG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>T. brucei</i> gapG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>L. mexicana</i> gapG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>E. gracilis</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*
<i>T. brucei</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>L. mexicana</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>D. discoideum</i>	?	?	?	?	?	?	?	?	?	+	-	-	-	-	-	-	-	-	-	-	-	-	-	??
<i>P. infestans</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. cerevisiae</i> gap1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. cerevisiae</i> gap2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. cerevisiae</i> gap3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Z. rouxii</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>K. lactis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>E. nidulans</i>	-	-	-	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+
<i>P. anserina</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. purpurea</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. gleosporioides</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>U. maydis</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. heterostrophus</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. parasitica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. lunata</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. commune</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>P. chrysosporium</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>A. bisporus</i> I	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>A. bisporus</i> II	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>G. verrucosa</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. crispus</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Z. mays</i> gapC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Z. mays</i> gapC4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>P. sativum</i> gapC1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>A. thaliana</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. reinhardtii</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>S. mansoni</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. elegans</i> gap1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. elegans</i> gap2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. elegans</i> gap3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. elegans</i> gap4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. briggsae</i> gap2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>C. briggsae</i> gap3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>D. hydei</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>D. melanogaster</i> gap1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>D. melanogaster</i> gap2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>H. sapiens</i> gapC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>G. gallus</i>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 2.4 (continued from previous page) Positions occupied by introns in GAPDH homologs. Intron position numbers correspond to the positions identified in Table 2.1. Pluses (+) indicate presence of an intron at that position in the sequence, hyphens (-) indicate the lack of an intron at that position and question marks (?) denote missing sequence data at the position. The asterisk (*) indicates that the introns found in the *Euglena gracilis* gapC have not been included in the analysis. These introns probably represent a novel intron type and their relationship to spliceosomal introns is currently unknown (Henze *et al.*, 1995).

Table 2.1 A comparison of GAPDH intron positions used in this study with those reported in Kersenach *et al.* (1994).

Intron#	Kersenach#	Codon#	Kersenach codon#
(2) ‡	1 †	(2-1) §	-3-1 ¶
1	2	1-2	-2-2
2	4	2-1	-1-1
3	3	4-0	-1-0
(3)	5	(4-0)	1-0
4	6	6-0	3-0
5	7	10-1	7-1
6	8	10-2	7-2
7	9	12-1	9-1
8	*	14-2	*
9	*	15-1	*
10	10	18-0	15-0
11	11	22-2	19-2
12	12	33-0	30-0
13	13	35-2	32-2
14	14	43-0	40-0
15	15	44-0	41-0
16	16	78-0	73-0
17	17	81-2	76-2
18	18	82-2	77-2
19	19	90-2	85-2
20	20	100-0	95-0
21	21	102-0	97-0
22	22	109-2	104-2
23	23	110-0	105-0
24	24	112-0	107-0
25	25	116-0	111-0
26	26	121-0	116-0
27	27	150-0	144-0
28	28	151-2	145-2
29	29	166-0	160-0
30	30	173-1	166-1
31	31	180-0	173-0
32	32	187-1	180-1
33	33	190-0	183-0
34	34	220-0	213-0
35	35	224-0	217-0
36	36	230-1	223-1
37	37	233-2	226-2
(37)	38	(233-2)	227-0
38	39	253-0	246-0
39	40	257-0	250-0
40	41	267-2	260-2
41	42	276-1	269-1
42	43	285-0	278-0
43	44	295-2	288-2
44	45	317-2	310-2
45	46	325-2	318-2
46	47	334-1	326-1

‡-Introns are numbered in order of occurrence from N- to the C-terminus of the GAPDH alignment. Parentheses indicate positions not considered distinct in this study.

†-numbering system used by Kersenach *et al.* (1994). §-Codon number and phase of intron respectively relative to the *Zea gapC4* enzyme. ¶-Codon and phase numbers in Kersenach *et al.* relative to the *Bacillus stearothermophilus* enzyme. Asterisks indicate positions unknown at the time of publication of Kersenach *et al.* (1994).

Phylogenies of the TPI and GAPDH genes.

Two kinds of trees of TPI and GAPDH homologs were constructed. For each dataset, the PROTDIST program was used to estimate a distance matrix from pairwise comparisons of the aligned amino acid sequences. For TPI and GAPDH, 238 and 350 alignment positions were considered respectively. Neighbour-joining (NJ) trees (inferred using the NEIGHBOR program) were then constructed based on the distance matrix from each dataset and are shown in Figs 2.5 and 2.7.

A second set of trees of each of the genes was developed based on the assumption that the NJ trees could be in error. For the TPI dataset this assumption is probably valid since bootstrap analysis (not shown) showed that the relative branching order of major groups in the NJ tree was not significantly resolved and hence the topology may not precisely reflect the true phylogeny. This assumption is also likely true of the GAPDH topology since previous analyses have suggested that many of the branches in trees of this dataset are also not well-supported and may be artifacts as a result of systematic biases such as unequal rates of evolution (Roger *et al.*, 1996). The second set of trees consisted of corrected, "hypothetical", gene trees. These were created by modifying the neighbour-joining topologies to reflect the organismal relationships proposed in Chapter 1 (Fig. 1.8). For the eubacterial portion of the hypothetical TPI tree, the relationships between eubacterial groups described in Olsen *et al.* (1994) were followed. However, several cases were identified where the GAPDH gene tree appeared to deviate strongly from an organismal tree. Previous analyses of GAPDH have shown that several ancient events of gene duplication likely occurred prior to the diversification of the eubacteria, producing several paralogous gene trees (Martin *et al.*, 1993, Henze *et al.*, 1995). Thus, the branching order within the eubacterial portion of the neighbour-joining

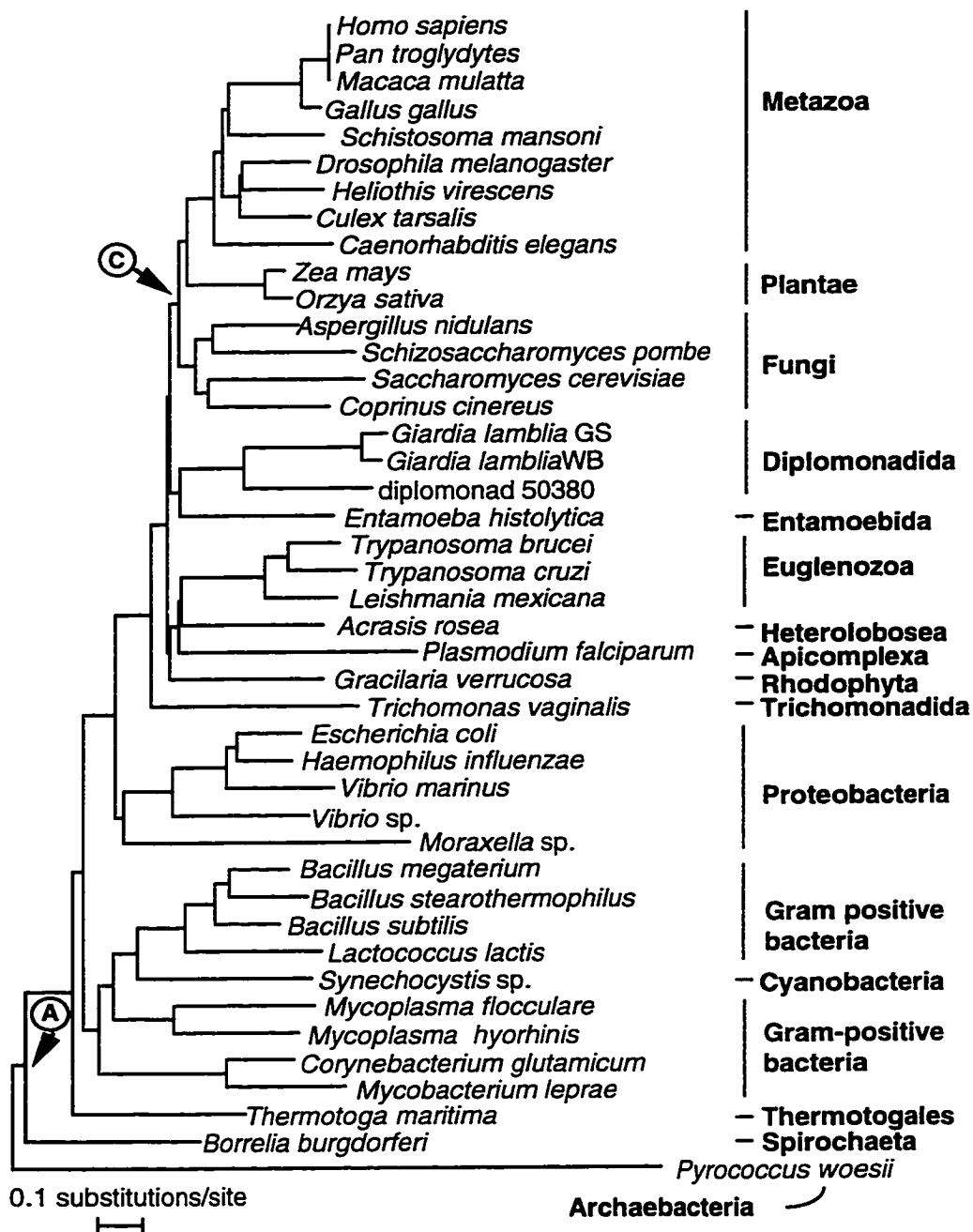


Figure 2.5 A phylogeny of TPI sequences using the neighbour-joining method. The distance matrix was inferred using the PROTDIST program with the Dayhoff setting. Bootstrap analysis (not shown) indicates that most of the internal structure of the eukaryotic portion of the trees is not significantly supported. The intron position sequences of labelled nodes were evaluated (Table 2.3). Node A is an immediate descendant of the root node and node C is the latest common ancestor of animals, plants and fungi. This latter node is also the latest common ancestor of animals and fungi (D).

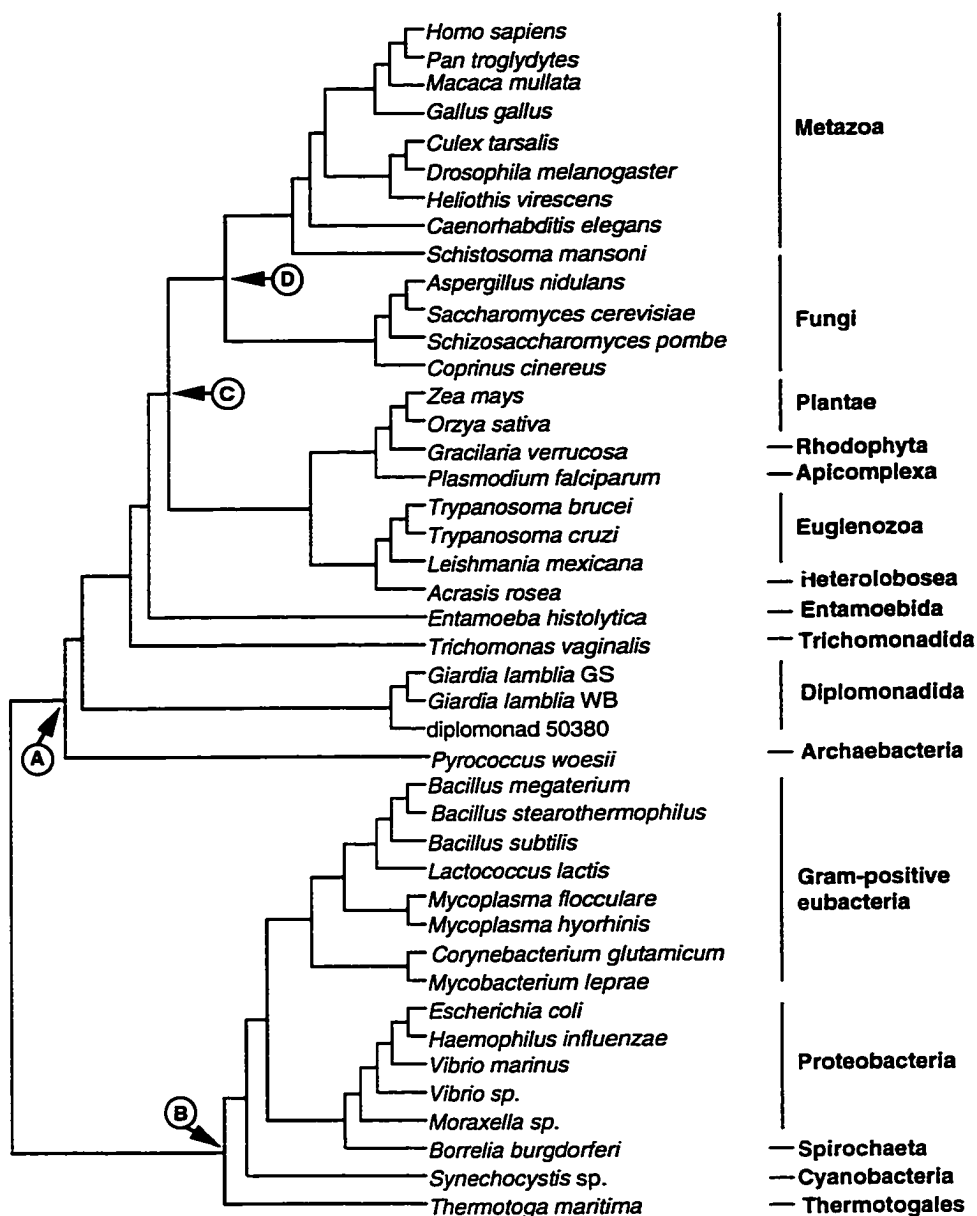


Figure 2.6 A hypothetical phylogeny of the TPI gene. The tree is based on the eukaryotic relationships proposed in Fig. 1.8 and prokaryotic relationships are taken from Olsen *et al.* (1994). The intron position sequences of labelled nodes were evaluated (Table 2.3). Node A is the latest common ancestor of eukaryotic and archaeobacterial sequences, node B is the latest common ancestor of the eubacterial sequences, node C is the latest common ancestor of animals, plants and fungi and node D is the latest common ancestor of animals and fungi.

tree was not modified to reflect organismal relationships in the hypothetical tree. In addition, the observation that several eubacterial paralogs (such as cyanobacterial *gap1* homologs and the γ -Proteobacterial *gapA* sequences (Fig. 2.6) branch near and within the eukaryotic portion of the GAPDH tree has led Martin *et al.* to propose that eukaryotes acquired their present day GAPDH enzyme via gene transfer from eubacteria (Martin *et al.*, 1993, Henze *et al.*, 1995). Although this scenario is quite complex, there are several lines of supporting evidence (Henze *et al.*, 1995, Roger *et al.*, 1996). Thus, the hypothetical GAPDH tree was constructed to reflect this scenario. The final hypothetical trees are shown in Figs. 2.6 and 2.8.

Parsimony calculations.

To rigorously evaluate the parsimony arguments made by introns-late advocates, for both GAPDH and TPI datasets and accompanying trees, the minimum numbers of events required by each introns theory to explain the intron distribution were evaluated. To do these calculations, it was necessary to fix the ancestral intron states implied by the theories. This was accomplished by constructing a hypothetical ancestral sequence to reflect the desired ancestral state for each intron position implied by each theory. This ancestral sequence was then joined by a branch to the root node of the tree. The roots of the GAPDH trees were placed on the archeobacterial branch consistent with an endosymbiotic origin for the eukaryotic homologs. The hypothetical TPI tree was based on the gene phylogeny tracking the universal phylogeny and was rooted according to gene duplication rootings of the universal tree (Iwabe *et al.*, 1989, Gogarten *et al.*, 1989). However, it was not possible to root the TPI neighbour-joining tree in this way since the *Pyrococcus* sequence was located on a branch far away from eukaryotic enzymes (Fig. 2.5). Recent work by Keeling and Doolittle (pers.

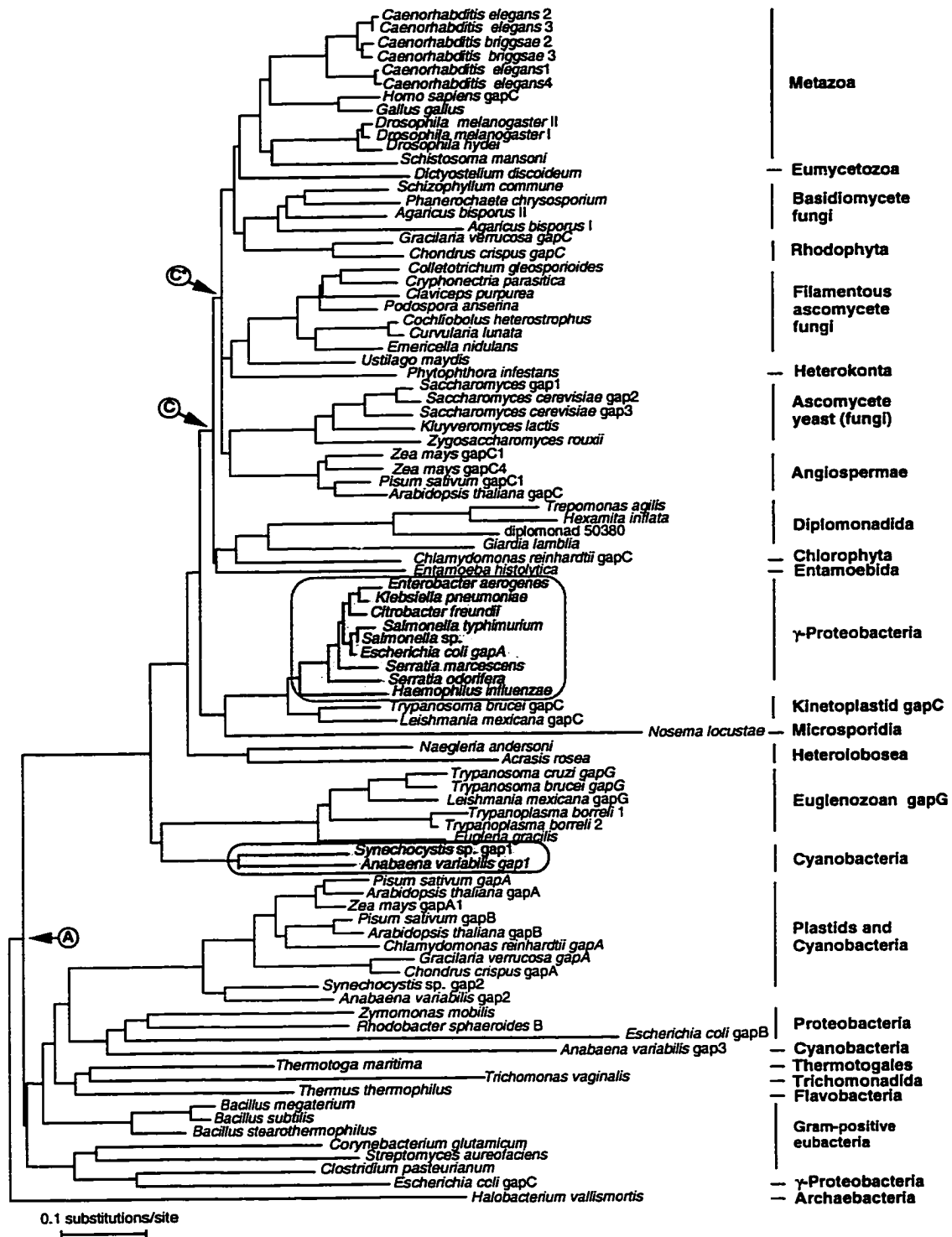


Figure 2.7 Phylogeny of GAPDH using the neighbour-joining method. The distance matrix was estimated using the PROTDIST program with the Dayhoff setting. Shaded boxes depict the eubacterial sequences placed near or within the eukaryotic subtree that suggest an endosymbiotic origin for the eukaryotic cytosolic GAPDH enzyme. The intron position sequences of labelled nodes were evaluated (Table 2.3). Node A is the common ancestor of eubacterial and eukaryotic sequences, node C is the latest common ancestor of animals, plants and fungi and node C* is the common ancestor of these three groups excluding *Chlamydomonas reinhardtii*. Node C* is also the latest common ancestor of animals and fungi (node D).

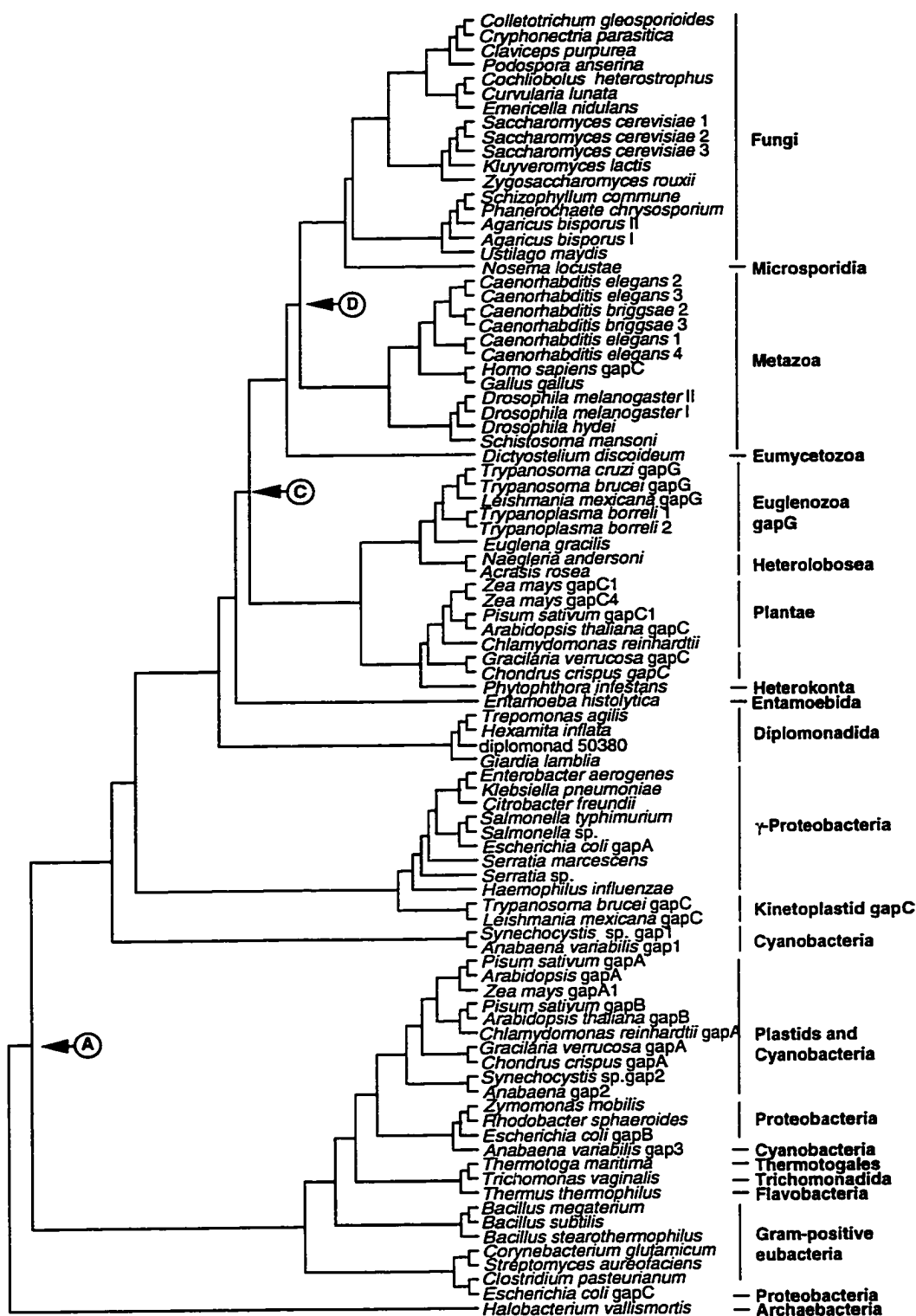


Figure 2.8 A hypothetical phylogeny of GAPDH. The eubacterial portion of this tree is derived from Fig. 2.7. The eukaryotic portion of the tree was corrected to match the hypothetical organismal phylogeny in Fig. 1.8 as well as reflect the endosymbiotic origin of the eukaryotic enzymes (Roger *et al.*, 1996). The intron position sequences of labelled nodes were evaluated (Table 2.3). Node A is the common ancestor of eubacterial and eukaryotic sequences, node C is the latest common ancestor of animals, plants and fungi and node D is the latest common ancestor of animals and fungi.

comm.) suggests that this anomalous topology, like the GAPDH tree, may be explained by an endosymbiotic origin for the eukaryotic TPI enzyme . To reflect this possibility, the root of the TPI/NJ topology was placed on the archaeobacterial branch .

The results of these tests are shown in Table 2.2. In accordance with introns-late arguments, for each dataset and an accompanying tree, the introns-late model required the fewest numbers of events to explain the intron data. Analysis of these datasets without the ancestral states fixed and allowing for events of gain and loss, yielded exactly the same reconstructions (and numbers of events) for each position as the introns-late test (not shown). The next most parsimonious hypothesis was the soft introns-early interpretation, followed by the hard introns-early interpretation. Hard introns-early required more than double the number of events compared to the other two views. Curiously, the number of events required by soft-introns early exceeds introns late by precisely the number of intron positions known (Table 2.2). Closer examination of the maximum parsimony reconstructions (MPRs) of the character changes over the tree revealed that this pattern was due to loss of each ancestral intron on the branch leading from the hypothetical ancestor to the root node. Thus, although each known intron position was forced to be ancestral in this test, none of the present day introns were reconstructed to be homologous to these . Instead, each intron appeared to have originated a second time in a recent lineage. In fact, except for the events of loss on the ancestral branch, the MPRs of each position for soft introns-early were identical to the corresponding introns-late reconstructions.

Table 2.2- A comparison of the minimum number of events required for each introns theory to explain the TPI and GAPDH intron data.

Dataset/Tree*	IL**	Hard IE***	Soft IE****
TPI/NJ‡	32	244	53
TPI/hypothetical†	37	206	58
GAPDH/NJ§	73	502	119
GAPDH/hypothetical§	69	551	115

*Abbreviations are TPI: triosephosphate isomerase; GAPDH; glyceraldehyde-3-phosphate dehydrogenase; NJ: neighbour-joining topology (shown Figs. 2.5 & 2.6); and hypothetical: hypothetical gene trees (shown in Figs. 2.7 & 2.8). **IL: the introns-late theory where loss and gain are allowed but with no introns held as ancestral. ***Hard IE: the hard introns-early theory where all introns are ancestral and can only be lost. ****Soft IE: the soft introns-early theory where all introns are ancestral but both loss and gain are allowed. † This tree is rooted on the archaeobacterial branch in accordance with a recent theory of the endosymbiotic origin for this gene. ‡ This tree is rooted in accordance with the recent rootings of the universal tree where the root falls between a eukaryote/archaeobacterial and a eubacterial clade. § These trees are rooted on the archaeobacterial branch in accordance with the proposed endosymbiotic origin of the eukaryotic enzymes.

In the analysis of Logsdon *et al.* (1995), of the TPI introns 5 positions, introns #3, #7, #9, #13 and #18 (Fig. 2.3), were suggested to be relatively "old" introns, having inserted into the common ancestor of animals and plants. In my study, the MPRs on the hypothetical and neighbour-joining TPI topologies show all of these introns as having inserted multiple times in the separate lineages that contain them. Rearrangement of the TPI topology to match that reported by Logsdon *et al.* (1995) (where taxa overlap), yielded MPRs that similarly showed multiple events of gain and loss at all of these positions except #3, which showed a single intron gain in a common ancestor of plants, *Plasmodium*, animals and fungi. Thus, with the dataset considered in this study, the maximum parsimony method in most cases does not reconstruct these putatively "old" introns as the result of single insertion events in the Precambrian era, as suggested by Logsdon *et al.* (1995).

Maximum likelihood analysis.

In order to test the various intron theories using a likelihood approach, the maximum likelihood ratio of rates of intron gain to loss was estimated for both the TPI and GAPDH datasets and their respective trees. As a starting estimate, the ratio of the overall frequencies of intron-interrupted to intron-uninterrupted sites were chosen to reflect the ratio of gain rate to loss rate. By varying these values over a range of 3 orders of magnitude, likelihood curves were developed for both of the two trees for each intron dataset (shown in Fig. 2.9). The maximum value on this curve was taken as the maximum likelihood estimate of this gain:loss ratio.

For both datasets, the hypothetical trees had overall greater likelihoods than the trees obtained by the neighbour-joining method, indicating that the topological corrections were somewhat successful. For TPI, the ML estimate of the gain:loss rate ratio was 0.0033 for the neighbour-joining topology and 0.0025 for the hypothetical topology (Fig. 2.9A). Both neighbour-joining and hypothetical trees had a gain:loss rate ratio of 0.0033 for the GAPDH intron dataset.

The hard introns-early theory predicts that the rate of intron gain is negligible and, therefore, the ML ratio of gain to loss rates will not significantly differ from zero. To test this, the 95% confidence intervals for these ML estimates were obtained by using the likelihood ratio test. These confidence intervals extend to ratios having likelihoods within $0.5\chi^2$ (with 1 degree of freedom) of the maximum likelihood value (Edwards, 1972). All of the ML estimates of the gain:loss rate ratio are significantly different from zero by this criterion. For TPI, these values extend from 0.0020-0.0048 for the neighbour-joining tree and 0.0013-0.0036 for the hypothetical topology. Confidence intervals of 0.0017-0.0054 and 0.0024-0.0067 were obtained with the GAPDH dataset for the neighbour-joining

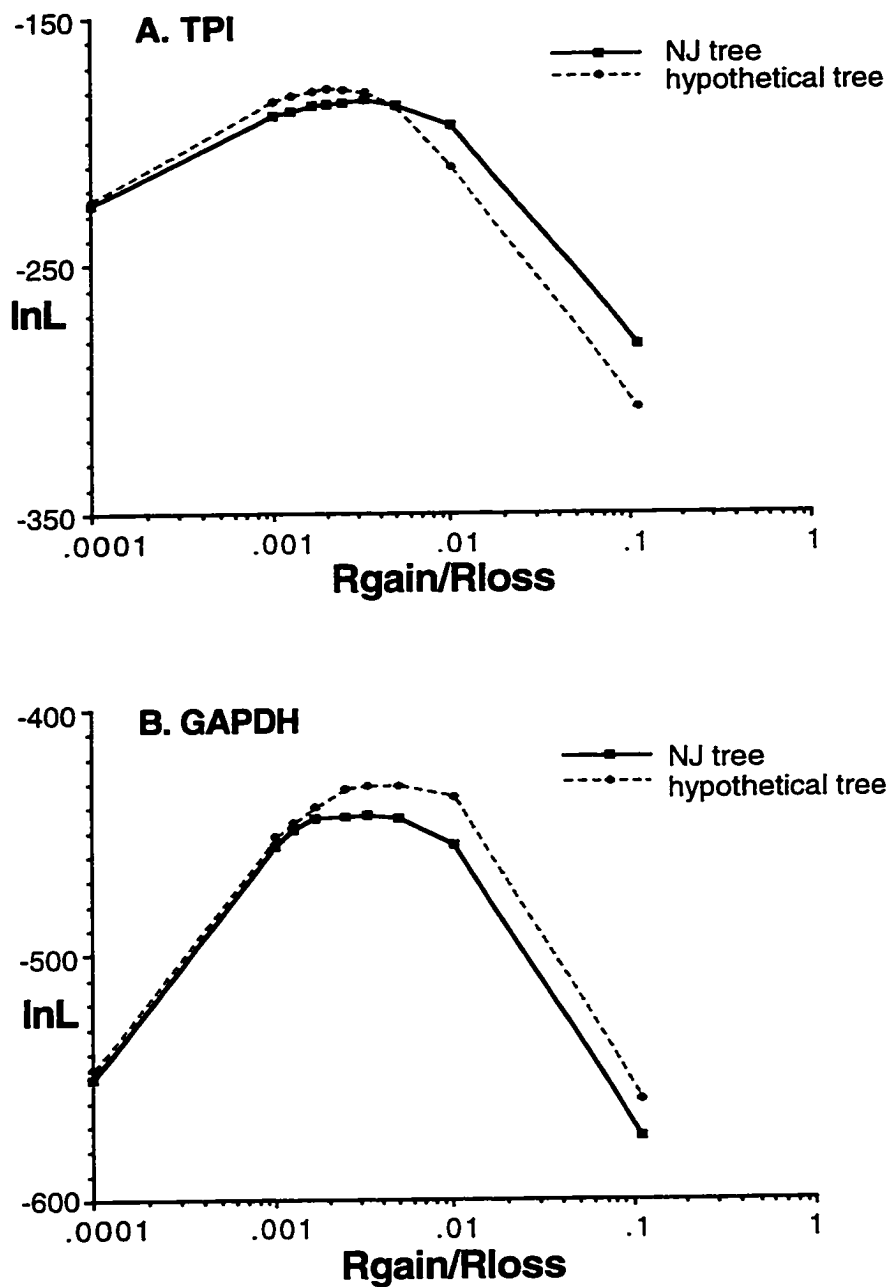


Figure 2.9 Estimation of the maximum likelihood ratios of the rate of intron gain to loss. The plot is log-likelihood ($\ln L$) versus the ratio of the instantaneous rate of intron gain to the instantaneous rate of loss ($R_{\text{gain}}/R_{\text{loss}}$). (A) Calculations using the TPI intron dataset and, (B), the GAPDH intron dataset. For both datasets, hypothetical and neighbour-joining trees were used in the maximum likelihood estimation procedure.

and hypothetical trees respectively. These data indicate that intron gain has probably occurred in the evolutionary history of these genes and that the hard introns-early view, in the absence of intron-sliding, is untenable.

Reconstruction of ancestral sequences.

In order to differentiate between a soft introns-early and introns-late views, it is necessary to determine whether the ancestral GAPDH or TPI genes had introns. This may seem like a futile task; we cannot directly observe the gene structure of the common ancestor of all cells. However, like the parsimony method, the likelihood method can be used to obtain estimates of the ancestral sequences of genes, or in this case, indicate whether they contained introns. To do this, one calculates the posterior probability of each state (either intron-interrupted or uninterrupted) for a particular position at internal nodes on the tree. This is accomplished by using an unpublished method implemented in the program DNAML 4.0 (J. Felsenstein, pers. comm.). Very similar methods for likelihood estimation of ancestral sequences have recently been published by Yang *et al.* (1995).

Since the DNAML program does not allow definition of the root node of a tree, it was not possible to directly estimate the intron position sequences of the ancestral TPI and GAPDH genes. However, the proposed roots of these trees all fall between two nodes that are defined. These are the nodes labelled A and B on each of the trees shown in Figs. 2.5-2.8. If a soft introns-early position were true, then for some modern introns to have been present in the ancestral genes, either of these descendant nodes must also have possessed them. For the two GAPDH trees and the neighbour-joining TPI topology, node B corresponds to the extant intron-lacking archaeobacterial intron sequences of these genes. Thus, in these

cases it is only necessary to reconstruct the ancestral sequence at node A. The reconstructions of the ancestral nodes were estimated and are shown in Table 2.3.

For each of the two trees for each dataset the ancestral sequences in node A and B (where it was necessary to test it) were reconstructed as completely lacking introns and, for each case, the posterior probability of the reconstruction at every position was greater than 95%. (Table 2.3). This method clearly favours the introns-late view that no introns were present in the ancestral GAPDH and TPI genes, even provided that the rate of intron loss vastly exceeds the rate of gain. The suggestions of Kersenach *et al.*, that the five introns conserved across the eukaryotic cytosol/plastid GAPDH boundary are ancestral (Kersenach *et al.*, 1994), are thus not supported by this analysis.

In order to evaluate the performance of the method, several other reconstructions were evaluated. The reconstructed intron position sequences of the latest common ancestor of animals, plants and fungi (node C) and the latest common ancestor of animals and fungi (node D) are shown in Table 2.3 for each dataset. Node C for the GAPDH/NJ dataset was reconstructed as completely lacking introns. However, this node is actually quite deep within the GAPDH tree since, in this analysis, *Chlamydomonas* (considered a plant) does not branch with higher plants (Fig. 2.7). If one excludes this sequence from consideration, node C* becomes the common ancestor of animals, plants and fungi in the neighbour-joining tree and this sequence appears to possess two introns, although the probabilities are <0.95 for each position. These positions (#6 and #15) correspond to positions #8 and #15 of Kersenach *et al.* that were suggested to be homologous introns shared by animals and fungi and animals, plants and fungi respectively. Clearly, the ML reconstructions are reflecting the intuitive assignments of homology for intron positions in this case. However, the hypothetical GAPDH topology shows no common ancestral introns for animals,

A strict consensus view of the ancestral sequences obtained from both trees of GAPDH, suggests that the common ancestor of animals, plants and fungi did not contain any introns while the common ancestor of animals and fungi contained two (position #6 and #15). In the case of GAPDH, the reconstructions are very sensitive to the topological differences between the neighbour-joining and the hypothetical topology.

By contrast, quite similar estimates of the sequences of ancestors were obtained from the analysis of neighbour-joining and hypothetical trees of TPI. Node C was reconstructed to have 6 introns and 8 introns in the neighbour-joining and hypothetical trees respectively. Both methods concur on five of these positions. These five positions, #3, #7, #9, #13 and #18, correspond exactly with the positions suggested to be ancestral to TPI genes of animals, plants and fungi by Logsdon *et al.* (1995). Again, in this case it appears that the ML reconstructions are reflecting assignments of homology made on the basis of human intuition. However, for each tree in this dataset there are several intron positions that are reconstructed to be ancestral and that are only found in two or even one of the descendant groups and were not suggested to be ancestral. The reconstructions at these positions can be considered predictions of introns that should be found in these descendant groups.

For both datasets, nodes C and D are the deepest nodes in the eukaryotic tree that are reconstructed to have any introns. The putatively deeply-branching groups in these trees not only lack introns but apparently the common ancestor they share with the "higher" eukaryotic groups also lacked them. It should be noted however, that the neighbour-joining trees of both datasets do not concur with each other nor with the consensus view developed in Chapter 1 of what eukaryotic groups are putatively deeply-branching. Until the true phylogeny of

these genes is better understood, the present results should be treated with a measure of caution.

DISCUSSION

This study represents the first attempt to apply objective methods to evaluate the implications of the introns-early and introns-late theories with respect to the phylogenetic distribution of intron positions in genes. Several years ago, Nyberg and Cronhjort also attempted to use a likelihood approach to this problem. Unfortunately, their results were irrelevant to the debate, since they compared a model where introns are only lost to a model where they are only gained (Nyberg & Cronhjort, 1992). This latter model does not adequately capture the introns-late perspective, since advocates of this view have accepted the validity of a process of intron loss for many years now (Rogers, 1989, Cavalier-Smith, 1991, Roger *et al.* 1994,).

The parsimony analyses indicate that the introns-late theory does indeed yield the most parsimonious interpretation of the intron distribution in genes such as GAPDH and TPI. Both introns-early views require extra events to have occurred in the evolution of these genes. However, examination of the most parsimonious reconstructions of intron evolution for TPI reveals that, for the most part, intron positions shared between animals, plants and fungi are not reconstructed to be ancestral to these groups. Yet, ironically, supporters of both introns-early and late camps appear to agree that these introns were ancestral to these groups (Logsdon *et al.*, 1995, de Souza *et al.*, 1996). This nicely illustrates the serious pitfalls of using pure parsimony arguments in the debate over alternative models of intron evolution. Clearly, if the cost of intron gain was allowed to increase relative to loss in a weighted parsimony calculation, then at some point, these introns *would* be reconstructed as ancestral. However, there is

no *a priori* way of assigning these costs and hence, advocates of each theory could settle on costs that caused their particular theory to become the most parsimonious, returning the argument to stalemate.

In sharp contrast, likelihood methods can objectively address the debate since there *is* a maximum likelihood ratio of the rates of intron loss to gain. Using a likelihood framework, I have shown that for the GAPDH and TPI datasets this loss:gain rate ratio appears to be in the range of 200-500:1. Clearly this estimate suggests that intron loss is vastly more frequent than intron gain for this dataset. Yet, the inverse of this ratio is significantly different from zero, suggesting that a model where introns are only lost (hard introns-early) is excluded. In addition, maximum likelihood reconstructions of intron position sequences suggest, with >95% probability, that no introns were ancestral to either the GAPDH or the TPI gene, favouring an introns-late view for both. It should be noted that in these analyses, I have assumed that all positions in these genes are equally likely to suffer intron insertion or loss. This same assumption was used by Kersenach *et.al.* to show that the chance matching of the five GAPDH introns shared by plastid and cytosolic enzymes by events of parallel insertion was extremely unlikely (Kersenach, *et al.*, 1994). The maximum likelihood results indicate that even with this assumption, the extreme improbability of chance matching is outweighed by the vanishing improbability of these introns being retained from the common ancestor of these sequences and lost scores of times in many independent lineages.

The ML reconstructions also suggest that no intron insertion appears to have occurred in these genes prior to the common ancestor of animals, plants and fungi. This result is consistent with the view that early-branching eukaryotic groups such as diplomonads and trichomonads may primitively lack spliceosomal introns (Roger & Doolittle, 1992; Cavalier-Smith, 1993). Moreover,

the reconstructions imply that widespread intron acquisition and loss in these genes appears to have started in the common ancestor of animals, plants and fungi and has continued to operate since this divergence. Although there was quite a disparity between reconstructions using different tree topologies for relatively recent nodes, the results concur to some degree with published assignments as to which intron positions of these genes are ancestral to the animal, plant and fungal groups, apparently made on the basis of intuition (Logsdon *et al.*, 1995, Kersenach *et. al.*, 1994). Therefore, intuitive reasoning like this may rely on principles more akin to likelihood than to parsimony.

Possible problems with the method.

In the absence of simulation studies, it is difficult to evaluate the performance of this likelihood method for dealing with the problem of intron evolution. However, it is clear that problems with the method may come from several sources.

The maximum likelihood ratio of the rate of intron loss to gain in all cases was in the range of 200-500:1. This ratio seems inordinately large. An explanation for this comes from the fact that only 3-5% of positions in these genes are interrupted by introns. Yet some of these positions appear to have suffered several independent gains of introns in the trees while some fraction of the other 95-97% of positions have not. On the face of it, this seems like evidence that the rate at which all sites gain introns may not be equal, perhaps because of the propensity for introns to insert at particular "target" sites. If this is true, then the implicit assumption of this method, that all sites evolve at a constant rate, is violated. This violation may introduce a systematic bias into the method; if a large proportion of uninterrupted sites are immune to intron insertion, then the method will severely underestimate the probabilities of intron loss and gain (the

branchlengths) in the tree. Moreover, the fact that they are invariable and intron-lacking will bias the maximum likelihood estimate of the intron loss rate to be far greater than it is in actuality, explaining the large ratio observed. This may, in turn, bias the method into reconstructing ancestral sequences with too many introns. Clearly, such a bias would favour the introns-early theory. However, since the method found no support for either version of this theory, this bias may not have adversely affected the tests of the alternative theories. By contrast, the reconstructions of intron position sequences at more recent nodes could be affected. A similar violation of models of DNA and protein evolution has been discussed in the literature and several alternative models have been put forward that allow for rate heterogeneity between sites (Yang, 1993, Felsenstein & Churchill, 1995). Application of these kinds of models to the intron data may allow for improved reconstructions of ancestral intron position sequences and better estimates for the ratio of the rate of intron loss to gain.

In addition to the assumption of rate homogeneity, this method assumes that the evolutionary process producing the intron distribution is at equilibrium; simply put, the model of evolution should not be changing over the tree. The violation of this assumption by the intron datasets may also be quite serious and the validity of the results of this study depends on how robust the method is to this kind of violation. The violation of this assumption is most serious when evaluating the hard introns-early theory. An adequate representation of this theory requires that the evolutionary process of intron evolution *never* comes to equilibrium; if introns can only be lost, then their frequency will be ever decreasing moving from the root of the tree to the tips. In addition, the introns-late theory explicitly suggests that introns did not evolve until sometime in eukaryote evolution. The application of a model that assumes constant loss and gain over both eukaryotic and prokaryotic groups will also therefore be

inappropriate, because, according to this theory, spliceosomal introns *could not* be inserted or lost in prokaryotic genes. Notwithstanding these problems, the density of introns appears to vary vastly across taxa, indicating that the relative rate of loss and gain has probably changed over the tree, once again leading to non-equilibrium conditions. The analogous problem in developing models of DNA evolution is the problem of varying base composition in different species. However, for DNA this variation in base composition never extends to an order of magnitude difference, yet intron densities can vary by at least 100 fold (J. Logsdon, pers. comm.). Clearly, an equilibrium model of intron evolution is not a very good approximation to reality when dealing with the huge evolutionary divergences of this study. However, since the violations of the model come from both introns-early and late theories, it is not clear whether the comparisons of these theories in this study have been biased in favour of either view. In order to deal with the problem of non-stationarity, intron models that do not have this constraint are being developed by this author in a collaboration with Z. Yang. Preliminary results obtained by this method concur with this study in suggesting that an introns-late explanation is the most likely for TPI and GAPDH intron datasets (Roger & Yang, unpublished data).

A final problem with this analysis may come from the lack of information in the intron sequences. In this study there were only 47 and 22 site patterns for the GAPDH and TPI datasets respectively. Yet for these datasets there were 177 and 83 parameters respectively in the form of branchlengths that were estimated and optimized from these data. This paucity of data will cause the estimates of these parameters to have large errors associated with them. This error will be reflected in the ancestral sequence reconstructions that depend on the estimated branchlengths for the calculation of posterior probabilities. Clearly, to solve this problem more genomic sequence data from a wide variety of organisms is

needed. If enough intron datasets are developed for a reasonable cross-section of taxa, then the data can be concatenated and more sensitive likelihood analyses like these can be performed.

Conclusions and perspectives.

If these potential problems with the intron model described in this chapter can be corrected or shown to be relatively unimportant by simulation studies, then with the wealth of new genomic sequence data, we may soon be able to apply them to large datasets. Analyses of these datasets may yield more concrete information about what periods in eukaryotic genome evolution were characterized by episodes of intron gain, loss or a combination of the two. Such information may also help in identifying taxa in which there is a lot of intron flux. The identification and study of such organisms is essential to improve our understanding of the mechanisms of intron gain and loss and will, in turn, allow us to make improvements in the intron models.

The results reported in this chapter represent a first attempt towards developing and testing several alternative intron models. This analysis clearly found the introns-late view was favoured for the GAPDH and TPI enzymes. If further analyses of other datasets have similar results, then the debate over introns-early versus intron-late theories may soon come to a close. Hopefully, a period of fruitful study of the dynamics of intron evolution will ensue.

Chapter 3

INTRODUCTION

The Archezoa hypothesis, as it was first proposed by Cavalier-Smith in 1983, suggested that several protist groups may have split from the main eukaryotic lineage prior to the endosymbiotic origin of mitochondria (Cavalier-Smith, 1983a, 1983b). Over the years Cavalier-Smith and others have suggested a number of amitochondrial protist groups that may belong in the Archezoa, including: diplomonads, retortamonads, oxymonads, Microsporidia, archamoebae and trichomonads. Testing whether an organism is truly archezoan relies on evidence of two sorts. Firstly, if an organism is primitively amitochondrial, then one expects that on molecular phylogenies it will diverge prior to all mitochondrion-containing eukaryotes. Secondly, an archezoan should not display any features that were acquired by eukaryotes as a result of mitochondrial endosymbiosis.

While the archezoan nature of retortamonads and oxymonads cannot yet be evaluated since no molecular data exist for representatives of either group, relevant information has been accumulating for the other groups. It is becoming increasingly clear that proposed archamoebae such as *Entamoeba histolytica*, *Phreatamoeba balamuthi* and *Pelomyxa* sp. are probably not Archezoa. Each of these organisms has been shown to branch from within mitochondrial eukaryotes in rRNA trees, implying that they have secondarily lost mitochondria (Sogin, 1991, Hinkle *et al.*, 1994, Morin & Mignot, 1995). The presence of genes of mitochondrial origin in *Entamoeba histolytica* has confirmed that mitochondrial functions have been lost in this lineage (Clark & Roger, 1995). Moreover, the placement of Microsporidia within the Fungi (see Chapter 1, Edlind *et al.*, 1996, Li *et al.*, 1996) in tubulin trees suggests that this protist phylum may also be derived from mitochondrion-bearing ancestors. By contrast, phylogenetic evidence

supports the view that diplomonads and trichomonads are Archezoa; they have been shown to branch prior to mitochondriate groups in several molecular phylogenies (Sogin, 1991, Klenk *et al.*, 1995, Gunderson *et al.*, 1995, Cavalier-Smith & Chao, 1996, Quon *et al.*, 1996, see also Chapter 1). Thus, the archezoan nature of trichomonads and diplomonads, until recently, was widely accepted (Patterson & Sogin, 1992, Müller, 1993, Margulis, 1996).

Nevertheless, there are dissenters from this view. Ironically, since 1987 Cavalier-Smith has argued vociferously against the primitively amitochondrial nature of trichomonads (Cavalier-Smith, 1987a). The disagreement hinges partly on differing interpretations of the origin of hydrogenosomes, unusual energy-generating organelles found in trichomonad cells. Hydrogenosomes function in the metabolism of pyruvate produced by glycolysis, generating ATP by substrate-level phosphorylation and evolving molecular hydrogen (Steinbuchel & Müller, 1986). Like mitochondria, they possess a double-membrane envelope and divide autonomously by fission (Müller, 1993). However, it is unclear whether trichomonad hydrogenosomes share a common ancestor with mitochondria or instead descend from a distinct endosymbiotic event: phylogenetic analyses of genes encoding hydrogenosomal proteins have so far failed to yield a strong link with the mitochondrial or any other specific eubacterial lineage (Müller, 1993, Johnson *et al.*, 1993, Hrdy & Müller, 1995, Müller, 1996).

Hydrogenosomes are also found in several other protist groups, and their origins are equally obscure. In the Ciliophora, their presence in several phylogenetically isolated anaerobic groups suggests that they may have evolved three to four times independently (Embley *et al.*, 1995). The presence of membranous invaginations in some of these ciliate hydrogenosomes suggests a morphological connection with mitochondria, organelles that are lacking in these

cells (Finlay & Fenchel, 1989, Embley & Finlay, 1994). Hydrogenosomes also appear to have arisen separately in chytridiomycete rumen fungi such as *Neocallimastix* (Yarlett *et al.*, 1986). In this case, the presence of a single bounding organellar membrane has been suggested by Cavalier-Smith to be indicative of a peroxisomal origin (Cavalier-Smith, 1987a). Most recently, hydrogenosome-like organelles have been reported in the amoeboflagellate genera *Psalteriomonas* and *Lyromonas* (Broers *et al.*, 1990, Brul *et al.*, 1994, Brul & Stumm, 1994). Once again, it appears that these double membraned organelles could be derived from mitochondria, although there is no direct evidence to support this claim (Brul & Stumm, 1994).

In this study, we sought to clarify the origin of hydrogenosomes in the trichomonad lineage. In so doing, we hoped to settle the controversy over whether the amitochondrial nature of these protists was a primitive or derived state. Previously, we reported the existence of two genes of mitochondrial origin that are retained in the amitochondriate amoeba *Entamoeba histolytica* (Clark & Roger, 1995). One of these genes, chaperonin 60 (cpn60), seemed an obvious gene to search for in trichomonads. In most eukaryotes, homologs of chaperonin 60 aid in the refolding of proteins after their import into mitochondria and plastids (Stuart *et al.*, 1994). We reasoned that if trichomonad hydrogenosomes were of endosymbiotic origin, then a cpn60 protein may perform a similar function in trichomonad hydrogenosomes. Phylogenetic analysis of the sequence of this gene could identify the closest eubacterial or organellar relatives of the hydrogenosome, thereby settling the question of its origin.

RESULTS

Isolation of a cDNA clone of the *T. vaginalis* cpn60 gene.

Our approach to finding a cpn60 gene in *T. vaginalis* was essentially the same as that was used in previous collaborative work on *E. histolytica* (Clark & Roger, 1995). Degenerate oligonucleotide primers were employed to amplify a 1.0 kb fragment of the gene with the polymerase chain reaction, and this DNA fragment was used to screen a cDNA library.

The complete sequence of the largest cDNA clone was obtained. The restriction site used in the cDNA cloning was fused directly to the 5' end of the coding region and no start codon was present in the sequence, indicating that the cDNA was truncated. The presence of a polyA tract 35 bp downstream of a UAA stop codon and the distinctive pattern of codon usage suggests that this cDNA was derived from *T. vaginalis* and not a bacterial contaminant. A Southern blot of *T. vaginalis* DNA digested with various restriction endonucleases, with this cDNA as a probe, revealed a single hybridizing band, implying that the protein is encoded by a single copy gene (data not shown) and confirming that *T. vaginalis* was its source.

Alignment and phylogenetic analysis.

The predicted *T. vaginalis* cpn60 protein, 544 amino acids in length, was entered into a previously reported alignment (Clark & Roger, 1995). To improve the taxonomic representation of the dataset near the node of interest, sequences from the α -Proteobacteria *Cowdria ruminantium*, *Bradyrhizobium japonicum* (groEL3) and *Brucella abortus* were added to the dataset. The final alignment contained the partial *E. histolytica* sequence, 4 eukaryotic mitochondrial homologues, 13 sequences from proteobacteria, 2 spirochetes and a chlamydia. Distance and parsimony analyses were based on 519 positions of this alignment.

For the maximum likelihood analyses, shared missing data were removed from the dataset yielding 501 alignment positions for analysis. However, for the analysis where the partial *Entamoeba histolytica* sequence was included, all positions missing in this sequence were eliminated from the alignment, leaving a final dataset of 362 positions. The maximum likelihood analysis required the use of a semi-constrained tree. Using the distance tree as a guide, several groups were constrained to be monophyletic: the spirochetes, the γ - and β -Proteobacteria, the non-Rickettsiales α -Proteobacteria and a mitochondrial cpn60 subtree was constrained consisting of animals, fungi and plants. All possible topologies containing these sub-trees were then evaluated and the trees of highest log-likelihood were determined for all datasets examined.

Preliminary phylogenetic analyses based on the full species dataset using neighbour-joining distance, maximum parsimony and maximum likelihood methods were performed (Fig. 3.1A).

The three methods generated similar trees. Mitochondrial sequences were specifically related to the Rickettsiales group (comprised of *Ehrlichia chaffeensis*, *Cowdria ruminantium* and *Rickettsia tsutsugamushi*) of the α -Proteobacteria, similar to previously published phylogenies (Viale *et al.*, 1994, Clark & Roger, 1995). In both neighbour-joining distance and maximum likelihood analysis, the *T. vaginalis* sequence formed a clade with the mitochondrial and *E. histolytica* cpn60 homologs. By contrast, maximum parsimony yielded five trees of equal length, all of which placed the *T. vaginalis* sequences as a specific sister group to the Rickettsiales sequences.

The bootstrap majority rule consensus trees from all three methods indicated that the *T. vaginalis* / *E. histolytica* / mitochondrial clade was the preferred topology in every case, including parsimony (Fig. 3.1A). For maximum likelihood, the support for this grouping was strong (90%) while distance and

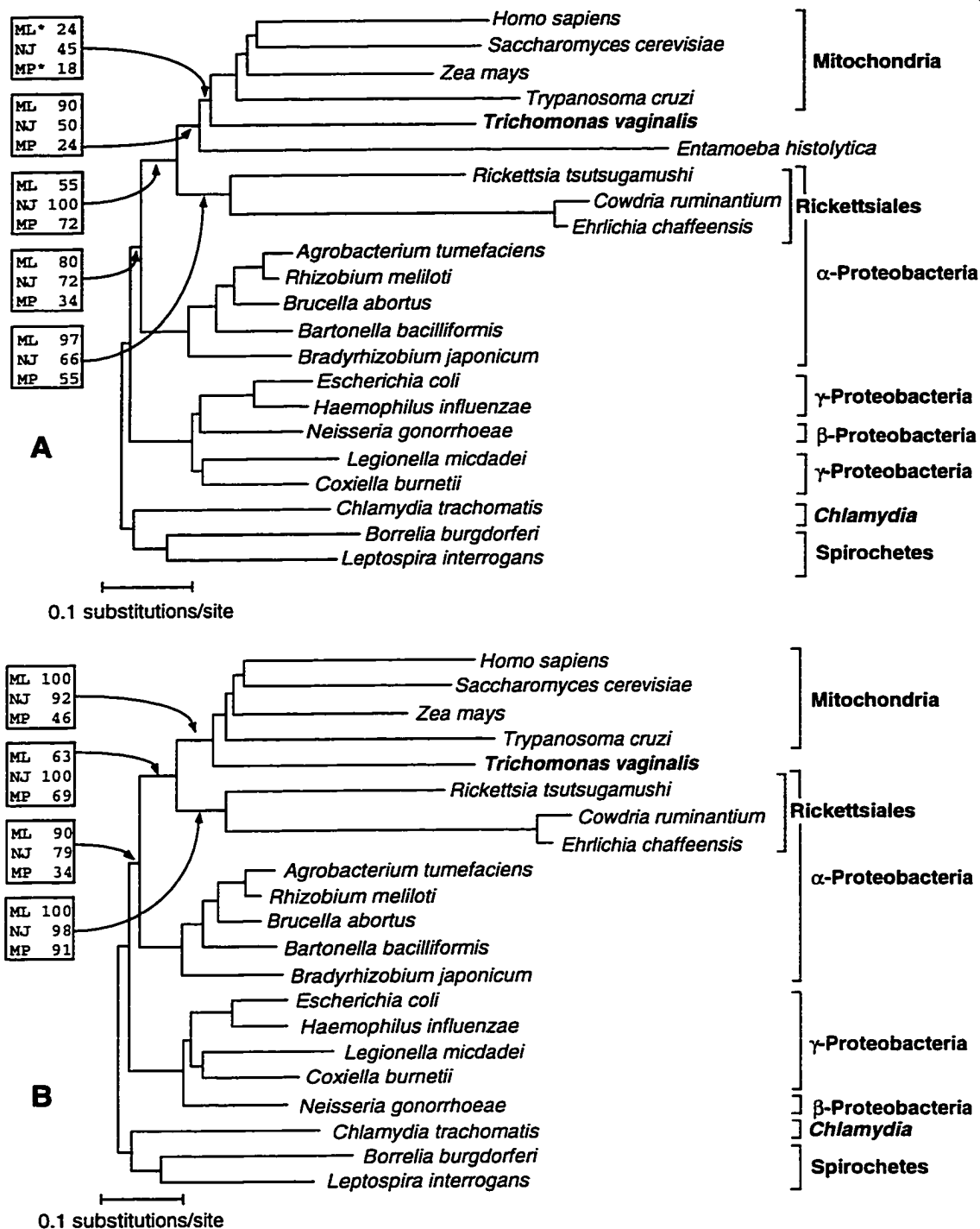


Figure 3.1 Phylogenies of cpn60 homologs. The trees shown are derived from neighbour-joining analysis of a PAM-corrected distance matrix. Percentage bootstrap support is shown above selected branches in boxes, from bootstrap analyses employing the ML (protein maximum likelihood), NJ (neighbour-joining distance), and MP (maximum parsimony) methods. For NJ and MP, 500 bootstrap replicates were performed. The ML bootstrap values were obtained using the REL method with 10,000 iterations. (A) Cpn60 tree derived from the full alignment. The maximum likelihood tree (\ln likelihood = -8933.6) differed from the neighbour-joining tree by the placement of the *T. vaginalis* and *E. histolytica* sequences as sister groups. Parsimony generated five trees of length = 2369 all of which differed principally from the tree shown by the placement of *T. vaginalis* as a sister group of the Rickettsiales species. Asterisks (*) indicate that the method used did not recover this node in the majority of bootstrap replicates. (B) Cpn60 tree with the *E. histolytica* sequence excluded. Maximum likelihood yielded a tree of identical topology (\ln likelihood = -12181.7), while parsimony generated three trees of length = 2209. Two of these differed from the neighbour-joining tree by the placement of the α -Proteobacteria (excluding the Rickettsiales) as a sister group to the γ - and β -Proteobacteria. The third differed by placing *T. vaginalis* as an immediate relative to the Rickettsiales (see text).

parsimony methods yielded significantly weaker support (50% and 24% respectively).

Previous analysis of the *cpn60* gene from *E. histolytica* showed that the extremely divergent nature of this sequence sometimes resulted in an affinity for the rickettsia, *Ehrlichia chaffeensis*, an artifactual result likely due to the long branch attraction phenomenon (Felsenstein, 1978, Hasegawa & Fujiwara, 1993, Clark & Roger, 1995,). We suspected, therefore, that the presence of the divergent *E. histolytica* sequence in the dataset may have been responsible for the poorly supported *T. vaginalis* / *E. histolytica* / mitochondria node in distance and parsimony analysis. In order to study the placement of the *T. vaginalis* sequence in the *cpn60* tree without the confounding influence of the *E. histolytica* sequence, we chose to exclude the latter from the subsequent analysis.

Analysis of the dataset without the *E. histolytica* sequence using neighbour-joining distance and maximum likelihood analyses indicated that the *T. vaginalis* *cpn60* sequence clustered with those of mitochondrial origin to the exclusion of all other sequences. As expected, the exclusion of *E. histolytica* caused bootstrap values for this relationship to increase for all three methods (Fig. 3.1B), with highly significant bootstrap values (>90%) for neighbour-joining and maximum likelihood analyses. However, maximum parsimony analysis still yielded relatively poor bootstrap support for this relationship. Moreover, three equally parsimonious trees were found, two of which displayed the *T. vaginalis* / mitochondria grouping whereas a third placed *T. vaginalis* as a specific sister group to the Rickettsiales (not shown).

To understand this result, we examined the impact of the inclusion and exclusion of Rickettsiales species on the bootstrap support for the *T. vaginalis* / mitochondria node and the alternative *T. vaginalis* / Rickettsiales node (Fig. 3.2). Deletion of the *Rickettsia* sequence causes bootstrap support for the

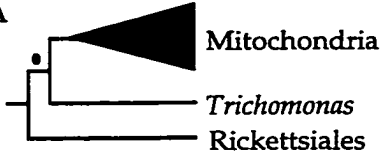
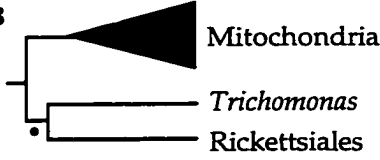
Method	Species of Rickettsiales included			Topology
	Ec, Cr, Rt	Ec, Cr	Rt	
NJ	92	99	68	A 
MP	46	67	41	
ML	100	99	89	
NJ	7	0	32	B 
MP	30	6	46	
ML	0	1	11	

Figure 3.2 The impact of the sampling of Rickettsiales species on the bootstrap support for two alternative topologies of the *cpn60* tree. The dataset excluding the *E. histolytica* sequence, was used to examine the bootstrap support for two alternative clades each indicated by a black dot (●) on the two trees. (A) The *T. vaginalis* / mitochondria clade found by neighbour-joining, maximum likelihood and two of the three maximum parsimony trees. (B) The *T. vaginalis* / Rickettsiales clade displayed by one of the three maximum parsimony trees (see text). Percentage bootstrap support for each clade is indicated to the left of the trees. Three different combinations of Rickettsiales species were used in the dataset. Species abbreviations are: Ec, *Ehrlichia chaffeensis*; Cr, *Cowdria ruminantium* and Rt; *Rickettsia tsutsugamushi*. For each combination of species, bootstrap support for the clade was evaluated using NJ (neighbour-joining distance), MP (maximum parsimony) and ML (protein maximum likelihood) methods.

T. vaginalis / mitochondria node (Fig. 3.2A) to increase in both neighbour-joining and parsimony analysis and support for the alternative *T. vaginalis* / Rickettsiales node (Fig. 3.2B) to decrease (the maximum likelihood bootstrap value was not strongly affected). Conversely, deletion of the highly similar *Ehrlichia* and *Cowdria* sequences causes bootstrap support for *T. vaginalis* / mitochondria to decrease for all methods with the alternative node receiving the majority of the remaining bootstrap support. It is clear from this that the affinity of the *T. vaginalis* sequence for the Rickettsiales is largely due to the presence of the *Rickettsia tsutsugamushi* sequence in the dataset. However, the effect is most apparent when maximum parsimony and neighbour-joining methods are used.

Maximum likelihood, by contrast, appears to be far less sensitive to this species sampling effect, in each case providing strong support for the *T. vaginalis* / mitochondria node. Since the maximum likelihood method has been shown to be more robust under conditions of substitution rate inequality between lineages (Hasegawa & Fujiwara, 1993, Kuhner & Felsenstein, 1994), we suggest that these conditions are the likely source of the *T. vaginalis* / Rickettsiales affinity observed in some of the parsimony and distance analyses. In any case, the *T. vaginalis* / mitochondria relationship is clearly preferred in 8 out of 9 of the phylogenetic analyses shown in Fig. 3.2, suggesting that this is likely the true gene phylogeny and that the alternative topology is artifactual.

DISCUSSION

Bozner recently used heterologous antibodies to immunolocalize a cpn60 homolog in trichomonads of the genus *Tritrichomonas*, showing that the cellular distribution of the cross-reacting protein is most consistent with a hydrogenosomal location (Bozner, 1996). Since we detected no other homologs of cpn60 in *T. vaginalis*, the gene we report probably encodes a hydrogenosomal

protein. After this study was completed, Horner *et al.* (1996) and Bui *et al.* (1996) also reported the sequence of this gene. The latter group showed that the protein contains a N-terminal leader sequence that is processed after import into hydrogenosomes, confirming our inferences made on the basis of the immunological data.

In other eukaryotes, cpn60 is known to function in the refolding of proteins following their transit across organellar membranes (Stuart *et al.*, 1994), suggesting that the *T. vaginalis* homolog may perform a similar function in the hydrogenosome. One other protein involved in protein refolding after organellar import is a specific isoform of the molecular chaperone hsp70. In two recently published papers, Germot *et al.* (1996) and Bui *et al.* (1996) report the existence of a gene encoding a mitochondrial isoform of hsp70 in *T. vaginalis* and also conclude that it likely has a hydrogenosomal location.

There are three possible origins, not mutually exclusive, for the *T. vaginalis* chaperonins. They could be derived from either the mitochondrial symbiont genome, the genome of the symbiont that gave rise to the hydrogenosome, or they could have been acquired by lateral transfer from another organism with which the ancestral trichomonad formed a transient symbiosis that did not result in the formation of an endosymbiotic organelle. Whichever of these possibilities is correct, the organism of origin for the chaperonin genes must have been very closely related, if not identical, to the mitochondrial endosymbiont.

Several distinct scenarios for the origin of the hydrogenosome are possible:

1. Trichomonad hydrogenosomes might have evolved directly from mitochondria by the loss of mitochondrial DNA and the electron transport chain (Cavalier-Smith, 1987a). If this is true, then proteins found in these hydrogenosomes but lacking in mitochondria must have been secondarily

acquired to complete the conversion. For hydrogenosomal enzymes such as pyruvate:ferredoxin oxidoreductase, found in the cytosol of amitochondrial eukaryotes such as *Giardia lamblia* and *E. histolytica* (Müller, 1996), this may have only required the acquisition of a targeting peptide onto the N-terminus of the protein. However, it is unclear how enzymes such as hydrogenase, unique to hydrogenosomes but lacking in mitochondria and the cytosol of other eukaryotes (Müller, 1996), were acquired by the ancestral trichomonad. The hydrogenosomal chaperonins in this case are derived from those of the mitochondrion. This scenario is supported by circumstantial evidence that hydrogenosomes in other eukaryotes appear to have arisen by conversion of mitochondria. For instance, hydrogenosomes of some ciliates bear mitochondrial cristae-like structures (Fenchel & Finlay, 1989, Embley & Finlay, 1994) while those of *Psalteriomonas lanterna* are enveloped by a layer of endoplasmic reticulum (ER) (Broers *et al.*, 1990, Brul *et al.*, 1994) in exactly the same arrangement that mitochondria are found to be associated with the ER in related heterolobosean amoeboflagellates (Page & Blanton, 1985).

2. A second view holds that hydrogenosomes and mitochondria are derived from a single endosymbiotic ancestor, which had all of the characteristics of both descendants (Johnson *et al.*, 1993). The lineage leading to trichomonads may have diverged from that leading to mitochondriate eukaryotes before the constituents of the present-day mitochondrion became fixed, with the two lineages retaining different functions of their shared ancestral symbiont. This view is supported by the finding that most hydrogenosomal enzymes, where comparative data exist, tend to be more similar to their eubacterial than their archaeobacterial homologs (Müller, 1996), consistent with an endosymbiotic origin. This

scenario also implies that selection for aerobic metabolism need not have been the sole force driving the initial integration of the symbiont, as is often suggested for mitochondria (Sagan, 1967, Margulis, 1970, Cavalier-Smith, 1987a).

3. A third possibility is that two independent endosymbioses involving closely related α -Proteobacteria occurred early after the divergence of trichomonads from the rest of the eukaryotes, giving rise (perhaps because of different selection pressures) to the hydrogenosome in the former case and mitochondria in the latter. In this scenario, the two organelles, and their chaperonins, share a pre-endosymbiosis common ancestry. However, the conversion from an endosymbiotic bacterium to an organelle likely requires many rare mutations to occur in succession (Cavalier-Smith, 1987a). Since this scenario requires that two such conversions occurred independently from the same bacterial lineage, it seems less probable.
4. It is also possible that an ancestral trichomonad possessed both the hydrogenosome and the mitochondrion but the two organelles had quite distinct endosymbiotic origins (Müller, 1993). Mitochondria were subsequently lost and certain proteins, including cpn60 and hsp70, were co-opted for use in the hydrogenosome. A distinct endosymbiotic origin for the hydrogenosome may explain the biochemical similarity noted between hydrogenosomes and some anaerobic bacteria (Müller, 1980).
5. Finally, it is possible that the hydrogenosome is not of endosymbiotic origin and the chaperonin genes were derived from a lateral transfer event from a mitochondrion-containing eukaryote or an unknown proteobacterial endosymbiont. In this case, the chaperonin genes are not indicative of the origin of the hydrogenosome as a whole.

In our opinion, scenarios 1 and 2 are the most likely and we believe that trichomonad hydrogenosomes and mitochondria share a common endosymbiotic origin. Regardless of which scenario is true, however, the phylogenetic affinities of the *cpn60* sequence in particular make it clear that an ancestor of trichomonads had an intimate relationship with an organism closely related to the mitochondrial symbiont that persisted long enough for gene transfer from its genome to the cell nucleus to take place.

From this example it is clear that the lack of mitochondrial functions coupled with a deeply-branching position for an organism in phylogenetic trees are not sufficient evidence that the organism evolved prior to the endosymbiotic origin of mitochondria. Two other amitochondrial protist groups occupy the deepest branches in SSU rRNA trees: diplomonads and microsporidia (Sogin, 1991, Cavalier-Smith, 1993). However, the case for the archezoan nature for these two groups is also weakening. Recent evidence from phylogenetic analyses of tubulins suggest that the rRNA trees may be in error and the Microsporidia may actually be related to fungi (see Chapter 1, Edlind *et al.*, 1996, Li *et al.*, 1996). If this phylogenetic placement is correct, then this group must also have suffered secondary loss of mitochondria. In the diplomonad *G. lamblia*, a 60 kDa protein which cross-reacts with anti-mitochondrial *cpn60* antibodies has been described (Soltys & Gupta, 1994), although sequence data are not currently available. Moreover, it has been suggested that eukaryotic glyceraldehyde-3-phosphate dehydrogenase (GAPDH) genes derive from mitochondrial endosymbiosis (Martin *et al.*, 1993, Roger *et al.*, 1996). Their presence in *G. lamblia* (Markós *et al.*, 1993, Henze *et al.*, 1995), other diplomonads (Rozario *et al.*, 1996) and Microsporidia (Chapter 2), could therefore also imply the loss of mitochondria from these organisms. However, better evidence for mitochondrial loss is needed before firm conclusions can be made. A concentrated search for

endosymbiotically-derived genes in these amitochondrial groups may help to decide whether the Archezoa are extinct.

Chapter 4

INTRODUCTION

Superoxide dismutases fall into two evolutionarily distinct families of enzymes: Copper, Zinc-SODs and Iron-/Manganese-SODs. These enzymes are important for the detoxification of superoxide (O_2^-) radicals produced as a by-product of enzymatic oxidation reactions (Beyer *et al.*, 1991). Smith and Doolittle have shown that the Fe- and Mn-SOD enzymes constitute an ancient superfamily of proteins with homologs found in both eukaryotic and prokaryotic cells (Smith & Doolittle, 1992). Specifically, Mn-SOD appears to be a ubiquitous enzyme that is found in many eukaryotes (where it is targeted to mitochondria), archaeobacteria and a wide variety of eubacterial cells. By contrast Fe-SOD appears to have a much narrower phylogenetic distribution, found typically in eubacterial lineages such as the Cyanobacteria and their endosymbiotic descendants, chloroplasts, the Bacteroides and the Proteobacteria. Smith and Doolittle's analysis showed that the Mn-SOD enzyme was likely present in the common ancestor of all cells and was retained in eukaryotes, Archaeobacteria and Eubacteria. However, within the eubacterial lineage, a duplication appears to have occurred to produce the Fe-SOD enzyme family. As expected, some organisms that diverged after the duplication event, such as *E. coli*, still possess both Mn- and Fe- SODs.

Because Fe-SOD appeared to be an exclusively eubacterial (and plastid) enzyme, its recent discovery in the amitochondriate non-photosynthetic protist *Entamoeba histolytica* came as a surprise (Tannich *et al.*, 1991). Smith *et al.* (1992) rationalized this finding by suggested that the gene was acquired by a rare event of lateral transfer from prokaryote to eukaryote. Several years later, Clark and this author showed that *Entamoeba*, although lacking recognizable mitochondria, appears to have retained two genes of mitochondrial origin: mitochondrial

chaperonin 60 and pyridine nicotinamide transhydrogenase (Clark & Roger, 1995). In this report, we suggested that the Fe-SOD in this organism could represent a third gene acquired via mitochondrial endosymbiosis, since within the SOD tree it clusters with the Proteobacteria, the immediate relatives of the mitochondrial endosymbiont.

This study is an attempt to address this hypothesis. I reasoned that if Fe-SOD was transferred to the nucleus after the endosymbiotic origin of mitochondria, then it must have been ancestrally present in mitochondriate eukaryotes. If this was so, then other eukaryotic groups should possess the enzyme. Secondly, an endosymbiotic origin for this enzyme would be most clearly demonstrated if the eukaryotic homologs clustered specifically with α -Proteobacteria, the closest relatives of mitochondria. To test these postulates, a PCR approach was used to amplify a portion of the SOD gene from several protists: the diplomonad, *Giardia lamblia*; the trichomonad, *Trichomonas vaginalis*; the heterolobosean, *Naegleria andersoni*; and the microsporidian *Encephalitozoon hellem*. SODs were also amplified from two α -Proteobacteria: *Rhizobium etli* and *Agrobacterium tumefaciens*.

RESULTS

The SODF1 and SODR1 primers produced products of approximately 450 bp from each organism tested. While this study was in progress, Fe-SODs from three other protists appeared in the database: *Plasmodium falciparum*, *Trypanosoma cruzi* and *Leishmania donovani*. These sequences, the sequences obtained in this study and six other α -proteobacterial SOD sequences obtained by P. Keeling, were entered into a full alignment of both Mn- and Fe- SODs. Preliminary phylogenetic analysis using this dataset showed that the products from *G. lamblia*, *T. vaginalis*, *N. andersoni* and *A. tumefaciens* fell within the Fe-SOD family

of proteins. By contrast, the *R. etli* and *E. hellem* products showed clear Mn-SOD affinities, even though the latter sequence was extremely divergent. Our analysis confirmed Smith and Doolittle's contention that the Fe-SOD family forms a coherent group falling within the eubacterial Mn-SOD subtree (Smith & Doolittle, 1992). Since this study focuses on the Fe-SOD family, further analysis of the Mn- enzymes was not undertaken.

A portion of the Fe-SOD alignment is shown in Fig. 4.1. Clearly, the enzyme is not particularly well suited to phylogenetic analysis because of its small length. Nevertheless, parsimony and distance analyses were carried out on a dataset of 180 positions from 32 aligned Fe-SOD protein sequences. A maximum parsimony tree is shown in Fig. 4.2. The relative branching order of the major groups was extremely unstable, and differed substantially between methods. Furthermore, bootstrap analysis using both methods showed that most of the internal structure of the Fe-SOD tree was not well supported (Fig. 4.2). The α - and γ -Proteobacteria both form polyphyletic groups in this tree, despite good evidence for their monophyly from SSU rRNA analyses (Olsen *et al.*, 1994). Eukaryotic sequences branch in various places in the tree. The *Giardia* sequence does not group with any other eukaryotes, but instead forms a moderately supported clade with the γ -proteobacterium *Bordatella pertussis*. *Entamoeba histolytica* and *Trichomonas vaginalis* sequences are also intermingled with some γ -proteobacterial sequences, but these groupings receive only weak bootstrap support. In sharp contrast, the SODs from *Plasmodium falciparum*, the kinetoplastids and *Naegleria andersoni* do form a monophyletic group in both distance and parsimony trees. Interestingly, *N. andersoni* and the kinetoplastids group together with extremely high bootstrap support in both analyses. The α -proteobacterium, *Ehrlichia chaffeensis*, is an outgroup to this eukaryotic clade, but once again, the grouping is not well supported. Plastids and Cyanobacteria

75

T.va QAYIDTANKLIVGS-GLEGKSIEEVIQKA-----QGPLFNNV
N.an KGYAVKLNELAQETETALAGKTIEEILLNF-----KGKAFNLS
G.la QTYVTNLNLIKGT-EFENLSLEEIVKKA-----SGGIFNNA
L.do MVFSIPPLPWGYDGLAAKGLSKQQVTLHYDKHHQGYVTKLNAAAQTNALATKSIEEIRTE-----KGPIFNLA
T.cr MVFSIPPLPWGYDGLAAKGLSKQQVTLHYDKHHQGYVTKLNAAAQTNALATKSIEEIRTE-----KGPIFNLA
P.fl MVITLPLKLYALNALSPH-ISEETLNFHYNKHHAGYVKNLNTLIKDT-PFAEKSLLDIVKES-----SGAIFNNA
E.hi MSFQLPQLPYAYNALEPH-ISKETLEFHDKHHATYVKNLNGLVKGT-EQEHKTLEELIKQKP----TQATYNNA
A.tu KAYVDNGNKLAAEA-GLSDLSLEEIVKKSFG--TNAGLFNNA
E.co MSFELPALPYAKDALAPH-ISAETIEYHYGKHHQTYVTNLNLIKGT-AFEGKSLEEIRSS-----EGGVFNNA
A.ni MSYELPALPFDYTALAPY-ITKETKEFHDKHHAAYVNNYNNAVKDT-DLDGQPIEAVTKAIAGDASKAGLFNNA
B.gi MTHELISLPHYAVDALAPV-ISKETVEFHHEKHLKTYVDNLNKLIIGT-EFENADLNTIVQKS-----EGGIFNNA

150

T.va AQHFNHSFFWKCLTPEK--VDVPSKVVDVLAASFESVEKFKETFTAKASTVFGSGWCYLYKNKDG--KCEIGQYSN
N.an AQVFNHNFYQSIGPNAG-GVPTGKIAAEIEKSFSGFDKFKAEFNTQAVNHFSGG*VWLVQRKDNNNLSVSTHD
G.la AQVWNHTFFWNLSLSPQGG-GAPTALADAINAKWGSFDFKKEEFKTAAGTAVGTFGSGGAWLVKKADG--SLDLVSTSN
L.do AQIFNHTFYWESMCPNGG-GEPTGKLADEINASFSGFAKFKEEFTNVAVGHFGSGLAWLVKDTNSGKLVYDTHD
T.cr AQIFNHTFYWESMCPNGG-GEPTGKVADEINASFSGFAKFKEEFTNVAVGHFGSGLAWLVKDTNSGKLVYQTHD
P.fl AQIWNHTFYWDSMGPDCC-GEPHGEIKEKIQEDFGSFNNFKQFQSNILCGHFGSGGWLALNNNN--KLVLQTHD
E.hi AQAWNHAFFYWKCMCGCG--VKPSEQLIAKLTAAFGGLEEFKKFKTEKAVGHFGSGGWCWLVEHDGK--LEIIDTHD
A.tu AQHYNHVHFVKWLVKKGDDGGNKLPGKLEQAFASDLGGYDKFKADFIAAGTTQFGSGGAWVSVKNGK--LEISKTPN
E.co AQVWNHTFYWNCLAPNAG-GEPTGKVAEALAAASFSGFADFKAQFTDAAIKNFGSGGWIWLVKNSDG--KLAIVSTSN
A.ni AQAWNHSFYWNSIKPNGG-GAPTALADKLAADFSGSFENFVTEFKQAAATQFGSGGAWLVLIDNG--TLKIKTTGN
B.gi GQTLNHNLYFTQFRPGK-GAPKGLGEAIDKQFGSFEKFKEEFNTAGTTLFGSGGWWLASDANG--KLSIEKEPN

208

T.va AANPVKDG-HKPLLTIV
N.an GGCPLTDG-LVPVLTIC
G.la ARTPLTTD-AKALLTI
L.do AGCPLTEPNLKPLLTCDVWEHAYYVDYKNDLAGYVQAFLN-VVNWKNVERQL*
T.cr AGCPLTEPNLKPLLTCDVWEHAYYVDYKNDRAAYVQTFWN-VVNWKNVERQL*
P.fl AGNPIKDNITGIPILTCDIWEHAYYIDYRNDRAAYVKAWWN-LVNWNFANENLKKAMKK*
E.hi AVNPMING-MKPLLTCDVWEHAYYIDTRNNRAAYLEHWWN-VVNWKFVEEQL*
A.tu GENPLVHG-AEPIILGV
E.co AGTPLTTD-ATPLLTVDVWEHAYYIDYRNARPGYLEHFWA-LVNWEFVAKNLAA*
A.ni ADTPIAHG-QTPLLTVDVWEHAYYLDYQNRDPDYISTFVEKLANWDFASANYAAAIA*
B.gi AGNPVRKG-LNPLLGFDVWEHAYYLYQNRADHLKDLWS-IVDWDIVESRY*

Figure 4.1 An alignment of Fe-SOD sequences obtained in this study with homologs from other organisms. The species name abbreviations are *T.va*: *Trichomonas vaginalis*; *N.an*: *Naegleria andersoni*; *G.la*: *Giardia lamblia*; *L.do*: *Leishmania donovani*; *T.cr*: *Trypanosoma cruzi*; *P.fl*: *Plasmodium falciparum*; *E.hi*: *Entamoeba histolytica*; *A.tu*: *Agrobacterium tumefaciens*; *E.co*: *Escherichia coli*; *A.ni*: *Anacystis nidulans*; and *B.gi*: *Bacteroides gingivalis*. The bolded taxon labels indicate the sequences obtained in this study. Asterisks (*) indicate stop codons. The sequences in the alignment are selected from a full alignment of 32 Fe-SOD homologs.

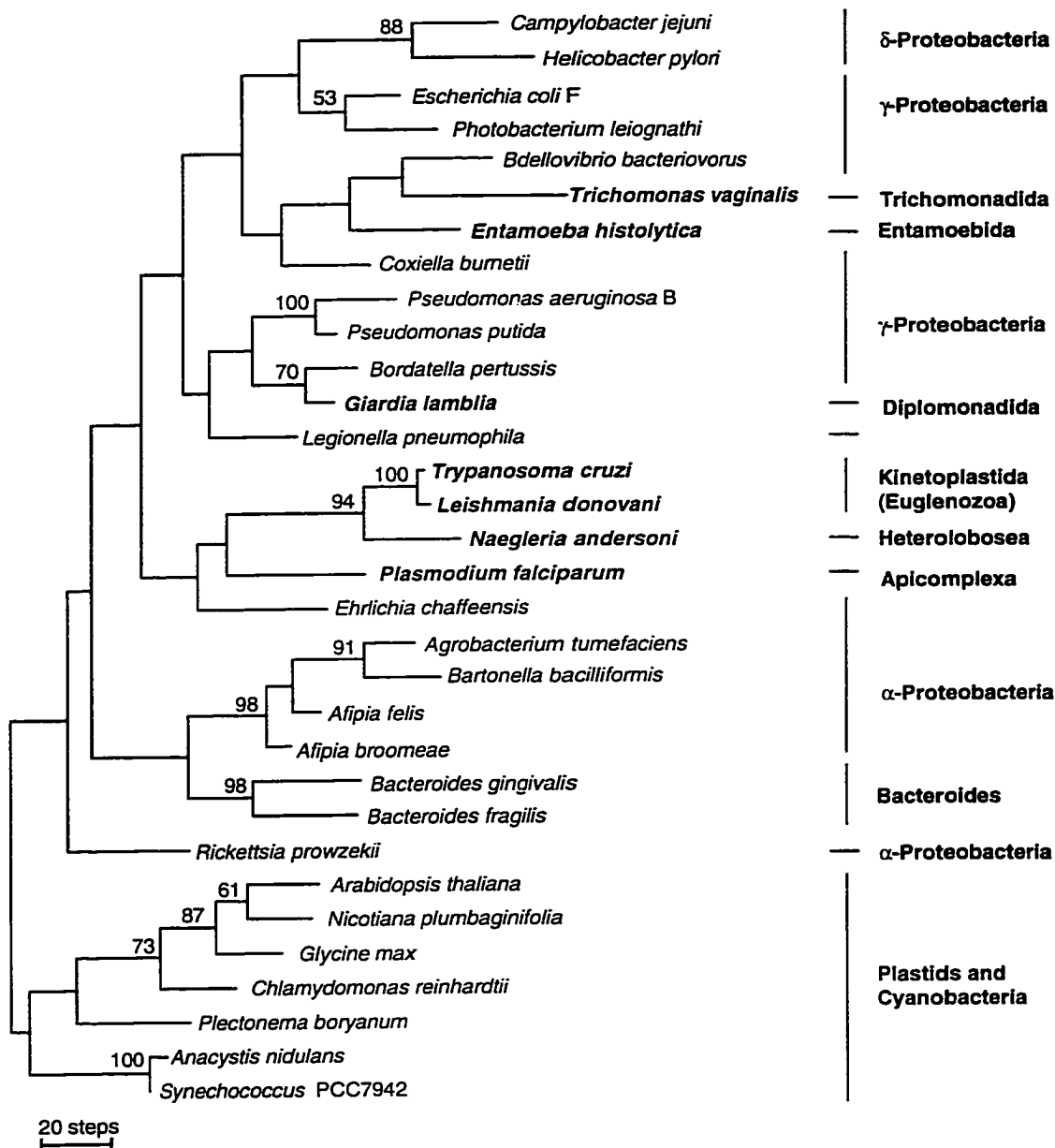


Figure 4.2 A tree of Fe-SOD inferred using the parsimony method. One of two equally parsimonious trees of steps is shown (length = 1239 steps). Bolded taxon labels denote eukaryotic sequences. The percentage bootstrap support from 300 bootstrap replicates are shown above the branches. Values of <50% are not shown. Neighbour-joining distance analysis yielded a tree which differed markedly from the tree above. However, the differences occurred principally at nodes poorly supported by bootstrap analysis using both distance and parsimony methods.

group together consistent with an endosymbiotic origin for the former nuclear-encoded enzymes. The separation of this group from the Proteobacteria and the Bacteroides was consistent between methods, but poorly supported. The Bacteroides sequences, however, fell within the Proteobacteria with their placement differing between distance and parsimony trees. Their branching position was not well supported in either case.

DISCUSSION

The finding of Fe-SOD genes in representatives of five distinct protist groups contradicts earlier suggestions that this enzyme is present only in eubacteria and plastids (Beyer *et al.*, 1991). Generalizations like this are often made based on data from animals, plants and fungi. It should be remembered that the protists comprise most of the phylogenetic diversity of eukaryotes and generalizations about eukaryotes should only be made when enough of this diversity is sampled (Cavalier-Smith, 1993a).

The phylogenetic analysis of the Fe-SOD family shows principally that this molecule retains little phylogenetic information. The causes of this are likely twofold. Firstly, the small size of the protein severely limits the total amount of information it can contain. Secondly, there appear to be extremely conserved invariant sites as well as highly variable sites in this molecule (Fig 4.1). Both of these kinds of sites are less informative than ones that evolve at a more moderate pace; invariant positions by definition do not contain phylogenetic information, while highly variable sites are often saturated by multiple overlapping substitutions. In addition, the presence of rate heterogeneity between sites is a violation of both the distance model employed and the assumptions of parsimony. Clearly, the odds are stacked against the recovery of an accurate phylogeny under these conditions.

Since the SOD tree is so poorly resolved the conclusions about the origin or origins of the eukaryotic enzymes can only be tentative. The hypothesis under test was that the *Entamoeba histolytica* enzyme was derived from mitochondrial endosymbiosis. I have confirmed that Fe-SOD enzymes are present in other eukaryotic groups as expected if this hypothesis were true. However, there is no clear evidence that all of these eukaryotic sequences are derived from the same source. The phylogenetic placement of the eukaryotic Fe-SOD enzymes within the proteobacterial/Bacteroides grouping is at odds with an origin for this protein by shared common ancestry with the eubacteria. It is more likely that they were derived from one or several events of eubacterial to eukaryotic lateral transfer. The number of these events cannot be determined precisely, but the tree only implies a maximum of four. Moreover, since the α -Proteobacteria do not appear to cluster strongly with any of the eukaryotic enzymes, a mitochondrial origin of even some of these enzymes cannot be confirmed. The most one can say is that a monophyletic, mitochondrial origin for these eukaryotic Fe-SOD enzymes is consistent with the data but not strongly endorsed by them. Perhaps once the full sequences of the eukaryotic and α -proteobacterial SODs described in this study are obtained, a clearer picture of the phylogeny will emerge.

A mitochondrial origin for these enzymes is favoured on *a priori* grounds. It is well known that many genes were transferred to the eukaryotic nucleus from the α -proteobacterial endosymbiont that gave rise to mitochondria (reviewed in Gray, 1992). Thus, the presence of proteobacterial-like enzymes in eukaryotes is best explained by appeal to this known source of such genes. Curiously, three of the eukaryotic Fe-SOD genes known are found in amitochondrial protists. Since there is now strong evidence for the secondary loss of mitochondrial functions in *E. histolytica* (Clark and Roger, 1995) and *T. vaginalis* (Chapter 3, Bui *et al.*, 1996, Horner *et al.*, 1996, Germot *et al.*, 1996), the endosymbiotic scenario for the origin

of the Fe-SODs of these organisms is not excluded by their amitochondrial nature. However, the explanation for the amitochondrial nature of diplomonads such as *G. lamblia* is still controversial. Immunological data suggest that this organism may contain a chaperonin 60 gene (Soltys & Gupta, 1994), but no sequence is yet available. The presence of a typical eukaryotic GAPDH gene in diplomonads has also been interpreted as evidence that these organisms are secondarily amitochondrial since it was suggested that this enzyme had a mitochondrial origin (Martin *et al.*, 1993, Henze *et al.*, 1995, Rozario *et al.*, 1996, Roger *et al.*, 1996)). Thus, although the evidence is weak, a mitochondrial origin for the *Giardia* Fe-SOD is possible.

In addition to the mitochondrial hypothesis, other scenarios should be considered. If the eukaryotic Fe-SODs originated polyphyletically as the phylogeny weakly suggests, then it is worthwhile to speculate how and why this may have occurred. The fact that the protein is small, soluble and does not interact with other cellular factors makes the likelihood of multiple independent lateral transfers relatively high. All that may be necessary is the insertion of the gene into the genome of the recipient near appropriate promoter sequences to allow transcription to take place.

It is unclear what the selective pressure would be on the recipient organism to acquire and maintain this protein. However, in eubacteria there appears to be a correlation between the possession of an Fe-SOD and an anaerobic lifestyle. This correlation plays out for the eukaryotes: all of the protists so far found to have these genes also tend to inhabit anaerobic climes. This correlation in the protists may either be due to selective forces acting directly on their anaerobic nature or due to the fact that the bacteria in their environment that donate the genes tend to be anaerobic and thus tend to possess Fe-SOD genes. Why exactly a correlation between anaerobiosis and the presence

of Fe-SOD should exist in the first place is unclear. Several studies of the enzymes in *Bacteroides*, *Streptococcus* and *Propionibacterium* indicate that either Fe- or Mn- can be used as the cofactor ion (summarized in Beyer *et al.*, 1991). The use of the particular ion, in these cases, appears to depend on its availability in the culture medium. These observations suggest that perhaps the selective forces responsible for the origin and maintenance of the various SODs may have to do with the availability of their cofactor ions in their environment. If a protist inhabits an anaerobic environment, then there may not be any available manganese ions, rendering its Mn-SOD inactive (unless it shares with the few organisms described above a SOD that can accommodate more than one kind of metal ion). The damaging presence of superoxide radicals in these organisms would thus create a substantial selective force for the acquisition of a SOD protein that was active in this environment. Clearly, iron ions are available in anaerobic environments and the acquisition of an Fe-SOD gene from a eubacterium would solve the problem. For these reasons, the possession of an Fe-SOD, in addition to a Mn-SOD, would also be advantageous for organisms that alternate between aerobic and anaerobic habitats.

Further studies of the Fe-SOD gene in other protists may help clarify its phylogenetic distribution amongst eukaryotic groups. Additional sequences may also improve the phylogenetic estimates obtained from this dataset since improved species sampling often helps to identify homoplasies in the dataset (Lecointre *et al.*, 1993). Once our understanding of the function and phylogeny of Fe-SODs improves, we may be able to speculate more intelligently about the selective forces that drove its acquisition and maintenance in the protists.

References

- Adachi, J. & Hasegawa, M. (1992) *Comput. Sci. Monographs*, (Inst. Statist. Math., Tokyo), No. 27.
- Adam, R. D., Nash, T. E. & Wellems, T. E. (1988) *Nucleic Acids Res.* **16**, 4555-4567.
- Allsopp, A. (1969) *New Phycologist* **68**, 591-612.
- Baldauf, S. L. & Palmer, J. D. (1993) *Proc.. Natl. Acad. Sci. U.S.A* **90**, 11558-62.
- Berbee, M. L. (1993) *Can. J. Bot.* **71**, 1114-1127.
- Bhattacharya, D., Stickel, S. K. & Sogin, M. L. (1991) *J. Mol. Evol.* **33**, 525-36.
- Blake, C. C. F. (1983) *Nature* **306**, 535-537.
- Bonen, L., Cunningham, R. S., Gray, M. W. & Doolittle, W. F. (1977) *Nucleic Acids Res.* **4**, 663-671.
- Bonen, L. & Doolittle, W. F. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 2310-2314.
- Bonen, L. & Doolittle, W. F. (1976) *Nature* **261**, 669-673.
- Bovee, E. C. & Jahn, T. L. (1973). in *The Biology of Amoeba*. ed. Jeon, K. W. (Academic Press Inc., New York), pp. 38-76.
- Bozner, P. (1996) *J. Parasitol.* **82**, 103-111.
- Branke, J., Berchtold, M., Breunig, A. & König, H. (1996) *Europ. J. Protistol.* **32**, 227-233.
- Breyer, W., Imlay, J. & Fridovitch, I. (1991) *Prog. Nucleic Acid Res.* **40**, 221-253.
- Broers, C. A. M., Stumm, C. K., Vogels, G. D. & Brugerolle, G. (1990) *Europ. J. Protistol.* **25**, 369-380.
- Brul, S. & Stumm, C. K. (1994) *Trends Ecol. Evol.* **9**, 319-324.
- Brul, S., Veltman, R. H., Lombardo, M. C. P. & Vogels, G. D. (1994) *Biochim. Biophys. Acta* **1183**, 544-546.
- Bruns, T. D., Vilgalys, R., Barns, S. M., Gonzalez, D., Hibbett, D. S., Lane, D. J., Simon, L., Stickel, S., Szaro, T. M., Weisburg, W. G. & et, a. (1992) *Mol. Phylogenet. Evol.* **1**, 231-41.

- Bui, L., Bradley, P. J. & Johnson, P. J. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 9651-9656.
- Canning, E. U. (1990). in *Handbook of Protozoists*, eds. Margulis, L., Corliss, J. O., Melkonian, M. & Chapman, D. J. (Jones & Bartlett, Boston), pp. 53-72.
- Cavalier-Smith, T. (1975) *Nature* **256**, 463-469.
- Cavalier-Smith, T. (1983a). in *Endocytobiology II*, eds. Schwemmler, W. & Schenk, H. E. A. (De Gruyter, Berlin), pp. 265-279.
- Cavalier-Smith, T. (1983b). in *Endocytobiology II*, eds. Schwemmler, W. & Schenk, H. E. A. (De Gruyter, Berlin), pp. 1027-1034.
- Cavalier-Smith, T. (1985) *Nature* **315**, 283.
- Cavalier-Smith, T. (1987a) *Ann N Y Acad Sci* **503**, 55-71.
- Cavalier-Smith, T. (1987b) *Symp. Br. Mycol. Soc.* **13**, 339-353.
- Cavalier-Smith, T. (1987c) *Evol. Trends Plants* **2**, 75-78.
- Cavalier-Smith, T. (1991) *Trends Genet.* **7**, 145-148.
- Cavalier-Smith, T. (1993a) *Microbiol. Rev.* **57**, 953-994.
- Cavalier-Smith, T. (1993b). in *Entocytobiology V*, eds. Ishikawa, H., Ishida, M. & Sato, S. (Tübingen University Press, Tübingen), pp. 399-406.
- Cavalier-Smith, T. & Chao, E. (1996) *J. Mol. Evol.* (in press).
- Cerff, R., Martin, W. & Brinkmann, H. (1994) *Nature* **369**, 527-528.
- Clark, C. G. & Roger, A. J. (1995) *Proc. Nat'l Acad. Sci. U. S. A.* **92**, 6518-6521.
- Cohen, J., Garreau de Loubresse, N. & Beisson, J. (1984) *Cell Motil.* **4**, 443-468.
- Condeelis, J. (1995) *Trends Biochem. Sci.* **20**, 169-170.
- Cunningham, R. S., Gray, M. W., Doolittle, W. F. & Bonen, L. (1977) in *Acides Nucléiques et Synthèse des Proteines Chez Les Végétaux.*, eds. Bogorad, L. & Weil, J.-H. (Centre National de la Recherche Scientifique, Paris), pp. 243-248.
- Daniels, E. W. & Breyer, E. P. (1967) *J. Protozool.* **14**, 167-179.
- Darnell, J. E. & Doolittle, W. F. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1271-1275.

- de Duve, C. (1973) *Science* **182**, 85.
- de Souza, S. J., Long, M. & Gilbert, W. (1996) *Genes to Cells* **1**, 493-505.
- Dibb, N. J. & Newman, A. J. (1989) *EMBO J.* **8**, 2015-2021.
- Dietmaier, W. & Fabry, S. (1994) *Curr. Genet.* **26**, 497-505.
- DiMaria, P., Palic, B., Debrunner-Vossbrinck, B. A., Lapp, J. & Vossbrinck, C. R. (1996) *Nucleic Acids Res.* **24**, 515-522.
- Dorit, R. L., Schoenback, L. & Gilbert, W. (1990) *Science* **250**, 1377-1382.
- Drouin, G., Moniz de Sá, M. & Zuker, M. (1995) *J. Mol. Evol.* **41**, 841-849.
- Duester, G., Jornvall, H. & Hatfield, G. W. (1986) *Nucleic Acids Res.* **14**, 1931-1941.
- Edlind, T. D., Li, J., Visvesvara, G. S., Vodkin, M. H., McLaughlin, G. L. & Katiyar, S. K. (1996) *Mol. Phylogenet. Evol.* **5**, 359-367.
- Edwards, A. W. F. (1972). *Likelihood*. (Cambridge University Press, Cambridge).
- Embley, T. M. & Finlay, B. J. (1994) *Microbiol.* **140**, 22-235.
- Embley, T. M., Finlay, B. J., Dyal, P. L., Hirt, R. P., Wilkinson, M. & Williams, A. G. (1995) *Proc R Soc Lond B Biol Sci* **262**, 87-93.
- Embley, T. M., Hirt, R. P. & Williams, D. M. (1994) *Phil. Trans. R. Soc. Lond. B* **345**, 21-33.
- Felsenstein, J. (1978) *Syst. Zool.* **27**, 401-410.
- Felsenstein, J. (1981) *J. Mol. Evol.* **17**, 368-376.
- Felsenstein, J. (1985) *Evolution* **39**, 783-791.
- Felsenstein, J. (1993). PHYLIP, Phylogeny Inference Package (University of Washington, Seattle) Version 3.57c.
- Felsenstein, J. & Churchill, G. A. (1996) *Mol. Biol. Evol.* **13**, 93-104.
- Fiedler, K. & Simons, K. (1995) *Trends Biochem. Sci.* **20**, 177-178.
- Finlay, B. J. & Fenchel, T. (1989) *FEMS Microbiol. Lett.* **65**, 311-314.

- Flegel, T. W. & Pasharawipas, T. (1995) *Can. J. Microbiol.* **41**, 1-11.
- Galtier, N. & Gouy, M. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92**, 11317-11321.
- Germot, A., Philippe, H. & Le Guyader, H. (1996) *Proc. Natl. Acad. Sci. USA* (in press).
- Gilbert, W. (1978) *Nature* **271**, 501.
- Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 901-905.
- Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46**, 151-154.
- Gogarten, J. P., Hilario, E. & Olendzenski, L. (1996) *Symp. Soc. Gen. Microbiol.* **54**, 1-26.
- Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, N. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6661-6665.
- Goksøyr, J. (1967) *Nature* **214**, 1161.
- Goldman, N. (1990) *Syst. Zool.* **39**, 345-361.
- Gray, M. (1983) *Bioscience* **33**, 693-699
- Gray, M. (1992) *Int. Rev. Cytol.* **141**, 233-356.
- Gray, M. W., Sankoff, D. & Cedergren, R. J. (1984) *Nucleic Acids Res.* **12**, 5837-5852.
- Griffin, J. L. (1988) *J. Protozool.* **35**, 300-315.
- Gunderson, J., Hinkle, G., Leipe, D., Morrison, H. G., Stickel, S. K., Odelson, D. A., Breznak, J. A., Nerad, T. A., Müller, M. & Sogin, M. L. (1995) *J. Euk. Microbiol.* **42**, 411-415.
- Gupta, R. S., Aitken, K., Falah, M. & Singh, B. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 2895-2899.
- Hasegawa, M. & Fujiwara, M. (1993) *Mol. Phylogenet. Evol.* **2**, 1-5.
- Hasegawa, M. & Hashimoto, T. (1993) *Nature* **361**, 23.
- Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N. & Miyata, T. (1993) *J. Mol. Evol.* **36**, 380-8.

- Hasegawa, M. & Kishino, H. (1994) *Mol. Biol. Evol.* **11**, 142-145.
- Hashimoto, T. & Hasegawa, M. (1996) *Adv. Biophys.* **32**, 73-120.
- Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K. & Hasegawa, M. (1994) *Mol. Biol. Evol.* **11**, 65-71.
- Henze, K., Badr, A., Wettern, M., Cerff, R. & Martin, W. (1995) *Proc. Nat'l. Acad. Sci. U. S. A.* **92**, 9122-9126.
- Hickey, D. A., Benkel, B. F. & Abukashawa, S. M. (1989) *J. Theor. Biol.* **137**, 41-53.
- Hinkle, G., Leipe, D., Nerad, T. A. & Sogin, M. L. (1994) *Nucleic Acids Res.* **22**, 465-469.
- Hinkle, G. & Sogin, M. L. (1993) *J. Euk. Microbiol.* **40**, 599-603.
- Hirt, R. P., Dyal, P. L., Wilkinson, M., Finlay, B. J., Roberts, D. & Embley, T. M. (1995) *Mol. Phylogenet. Evol.* **4**, 77-87.
- Horner, D. S., Hirt, R. P., Kilvington, S., Lloyd, D. & Embley, T. M. (1996) *Proc. R. Soc. Lond. B* **263**, 1053-1059.
- Hrdy, I. & Müller, M. (1995) *J Mol Evol* **41**, 388-396.
- Hurst, L. D. (1994) *Nature* **371**, 381-382.
- Hyde, J. E., Sims, P. F. G. & Read, M. (1994) *Parasitol. Today* **10**, 25.
- Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9355-9359.
- Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. & Miyata, T. (1990) *J. Mol. Evol.* **31**, 205-210.
- Jacquier, A. (1990) *Trends Biochem. Sci.* **15**, 351-354.
- John, P. & Whatley, F. R. (1975) *Nature* **254**, 495-498.
- Johnson, P. J., Lahti, C. J. & Bradley, P. J. (1993) *J Parasitol* **79**, 664-70.
- Juretic, N., Mattes, U., Ziak, M., Christen, P. & Jaussi, R. (1990) *Eur. J. Biochem.* **192**, 119-126.

- Kamaishi, T., Hashimoto, T., Nakamura, Y., Nakamura, F., Murata, S., Okada, N., Okamoto, K., Shimizu, M. & Hasegawa, M. (1996) *J. Mol. Evol.* **42**, 257-263.
- Keeling, P. J. & Doolittle, W. F. (1996) *EMBO J.* **15**, 2285-2290.
- Keister, D. (1983) *Trans. Roy. Soc. Trop. Med. Hyg.* **77**, 487-488.
- Kersanach, R., Brinkmann, H., Liaud, M. F., Zhang, D. X., Martin, W. & Cerff, R. (1994) *Nature* **367**, 387-9.
- Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29**, 170-179.
- Klenk, H.-P., Zillig, W., Lanzendorfer, M., Grampp, M. & Palm, P. (1995) *Arch. Protistenkd.* **145**, 221-230.
- Kuhner, M. K. & Felsenstein, J. (1994) *Mol. Biol. Evol.* **11**, 459-468.
- Kuma, K., Nikoh, N., Iwabe, N. & Miyata, T. (1995) *J. Mol. Evol.* **41**, 238-246.
- Kumar, S. & Rzhetsky, A. (1996) *J. Mol. Evol.* **42**, 183-193.
- Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K. & Ayala, F. J. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 8503-8506.
- Lambowitz, A. M. & Belfort, M. (1993) *Ann. Rev. Biochem.* **62**, 587-622.
- Lamond, A. I. (1993) *Curr. Biol.* **3**, 62-64.
- Lecointre, G., Philippe, H., Lê, H. L. V. & Le Guyader, H. (1993) *Mol. Phylogenet. Evol.* **2**, 205-224.
- Leipe, D., Gunderson, J. H., Nerad, T. A. & Sogin, M. L. (1993) *Mol. Biochem. Parasitol.* **59**, 41-48.
- Li, J., Katiyar, S. K., Hamelin, A., Visvesvara, G. S. & Edlind, T. D. (1996) *Mol. Biochem. Parasitol.* **78**, 289-295.
- Liaud, M.-F., Brinkmann, H. & Cerff, R. (1992) *Plant Mol. Biol.* **18**, 639-651.
- Liu, Q. Y., Baldauf, S. L. & Reith, M. E. (1996) *Plant. Mol. Biol.* **31**, 77-85.
- Lockhart, P. J., Larkum, A. W. D., Steel, M. A., Waddell, P. J. & Penny, D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 1930-1934.
- Logsdon, J., Jr., Tyshenko, M. G., Dixon, C., D-Jafari, J., Walker, V. K. & Palmer, J. D. (1995) *Proc. Natl. Acad. Sci. U. S. A.* **92**, 8507-8511.

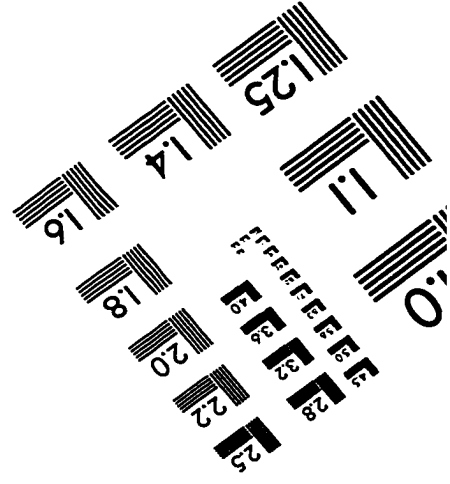
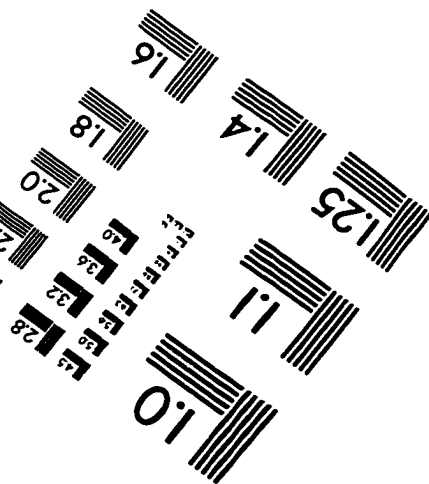
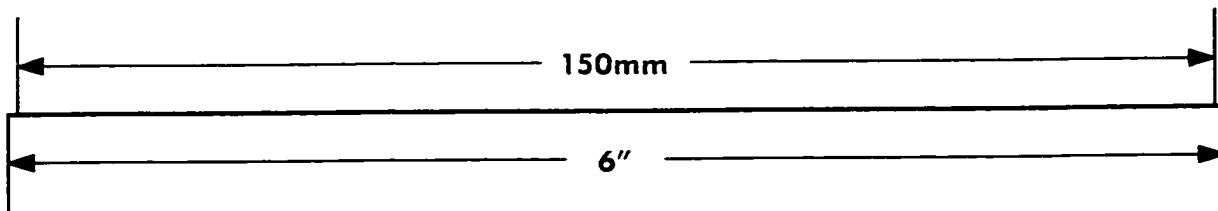
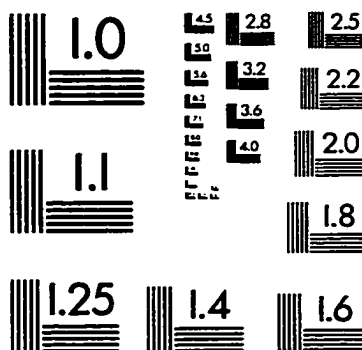
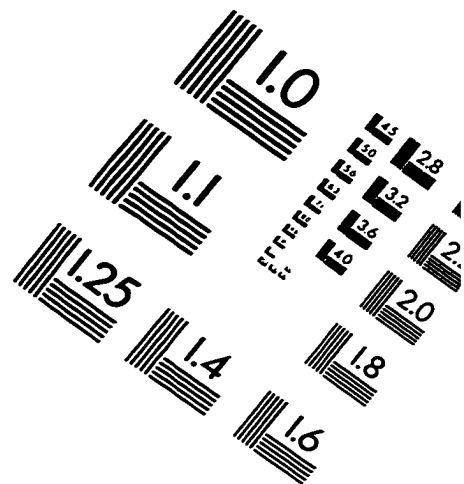
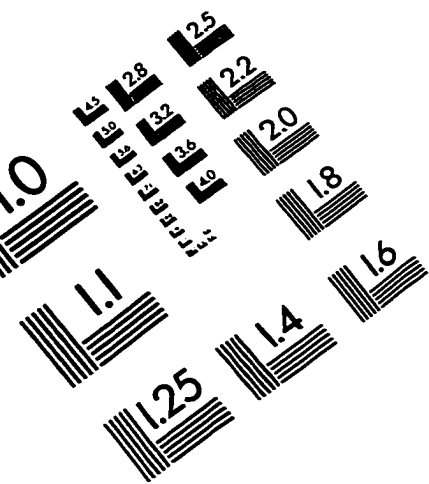
- Logsdon, J. M. & Palmer, J. D. (1994) *Nature* **369**, 526-528.
- Logsdon, J. M., Jr. & Palmer, J. D. (1994) *Nature* **369**, 526; discussion 527-8.
- Long, M., Rosenberg, C. & Gilbert, W. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 12495-12499.
- Margulis, J. (1970). *Origin of Eukaryotic Cells*. (Yale University Press, New Haven).
- Margulis, L. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 1071-1076.
- Markós, A., Miretsky, A. & Müller, M. (1993) *J. Mol. Evol.* **37**, 631-43.
- Martin, W., Brinkmann, H., Savonna, C. & Cerff, R. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90**, 8692-6.
- Martinez, P., Martin, W. & Cerff, R. (1989) *J. Mol. Biol.* **208**, 551-565.
- Mattick, J. S. (1994) *Curr. Biol.* **4**, 823-831.
- McPheeters, D. S. & Abelson, J. (1992) *Cell* **71**, 819-831.
- Morin, L. & Mignot, J.-P. (1995) *Eur. J. Protistol.* **31**, 402.
- Müller, M. (1980) *Symp. Soc. Gen. Microbiol.* **31**, 127-142.
- Müller, M. (1993) *J Gen Microbiol* **139**, 2879-89.
- Müller, M. (1996). in *Anlässlich der 14. Wissenschaftlichen Jahrestagung der Deutschen Gesellschaft für Protozoologie*, eds. Schlegel, M. & Hausmann, K. (Christian Gottfried Ehrenberg-Festschrift, Delitzsch). pp. 63-76.
- Nakamura, Y., Hashimoto, T., Kamaishi, T., Adachi, J., Nakamura, F., Okamoto, K. & Hasegawa, M. (1996) *J. Biochem.* **119**, 70-79.
- Nass, M. M. K. & Nass, S. (1963) *J. Cell. Biol.* **19**, 593-691.
- Nikoh, N., Hayase, N., Iwabe, N., Kuma, K. & Miyata, T. (1994) *Mol. Biol. Evol.* **11**,
- Niu, X. H., Hartshorne, T., He, X. Y. & Agabian, N. (1994) *Mol. Biochem. Parasitol.* **66**, 49-57.
762-768.
- Nyberg, A. M. & Cronhjort, M. B. (1992) *J. Theor. Biol.* **157**, 175-190.

- O'Kelly, C. J. (1993) *J. Euk. Microbiol.* **40**, 627-636.
- Olsen, G. J., Woese, C. R. & Overbeek, R. (1994) *J. Bacteriol.* **176**, 1-6.
- Page, F. C. & Blanton, R. L. (1985) *Protistologica* **21**, 121-132.
- Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opin. Genet. Dev.* **1**, 470-7.
- Patterson, D. J. (1994). in *Progress in Protozoology*, eds. Hausmann, K. & Hulsman, N. (Gustav Fischer Verlag, Stuttgart). pp. 1-14.
- Patterson, D. J. & Sogin, M. L. (1992). in *The Origin and Evolution of Prokaryotic and Eukaryotic Cells*, eds Hartman, H. & Matsuno, K. (World Scientific Publishers, Singapore). pp. 13-46.
- Patthy, L. (1987) *FEBS Lett.* **214**, 1-7.
- Patthy, L. (1991) *Curr. Opin. Struct. Biol.* **1**, 351-361.
- Quon, D. V. K., Delgadillo, M. G. & Johnson, P. J. (1996) *J. Mol. Evol.* **43**, 253-262.
- Raff, R. A. & Mahler, H. R. (1972) *Science* **177**, 575-582.
- Raff, R. A. & Mahler, H. R. (1973) *Science* **180**, 516-517.
- Ris, H. & Plaut, W. (1962) *J. Cell. Biol.* **12**, 383-391.
- Roger, A. J. & Doolittle, W. F. (1993) *Nature* **364**, 289-290.
- Roger, A. J., Keeling, P. J. & Doolittle, W. F. (1994) *Soc. Gen. Physiol. Ser.* **49**, 27-37.
- Roger, A. J., Smith, M. W., Doolittle, R. F. & Doolittle, W. F. (1996) *J. Euk. Microbiol.* **43**, 475-485.
- Rogers, J. H. (1989) *Trends Genet.* **5**, 213-216.
- Rogers, J. H. (1990) *FEBS Lett.* **268**, 339-343.
- Rozario, C., Morin, L., Roger, A. J., Smith, M. W. & Müller, M. (1996) *J. Euk. Microbiol.* **43**, 330-340.
- Sagan, L. (1967) *J. Theor. Biol.* **14**, 225-274.
- Sapp, J. (1994). *Evolution by association*. (Oxford University Press, New York).
- Schlegel (1994) *Trends Ecol. Evol.* **9**, 330-335.

- Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199**, 395-403.
- Siddall, M. E., Hong, H. & Desser, S. S. (1992) *J. Protozool.* **39**, 361-367.
- Silberman, J. D., Sogin, M. L., Leipe, D. & Clark, C. G. (1996) *Nature* **380**, 6573.
- Smith, M. W. & Doolittle, R. F. (1992) *J. Mol. Evol.* **34**, 175-184.
- Smith, M. W., Feng, D. F. & Doolittle, R. F. (1992) *Trends Biochem. Sci.* **17**, 489-93.
- Smothers, J. F., von Dohlen, C. D., Smith, L. H. & Spall, R. D. (1994) *Science* **265**, 1719-1721.
- Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution and Inference*. (MIT Press, Boston).
- Sogin, M. L. (1989) *Amer. Zool.* **29**, 487-499.
- Sogin, M. L. (1991) *Curr. Opin. Gen. Develop.* **1**, 457-463.
- Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & A., P. D. (1989) *Science* **243**, 75-77.
- Soltys, B. J. & Gupta, R. S. (1994) *J. Parasitol.* **80**, 580-590.
- Stanier, R. Y. (1970). *Symp. Soc. Gen. Microbiol.* **20**, 1-38
- Stanier, R. Y. & van Niel, C. B. (1962) *Arch. Mikrobiol.* **42**, 17-35.
- Steinbuchel, A. & Müller, M. (1986) *Mol. Biochem. Parasitol.* **20**, 57-65.
- Stoltzfus, A. (1994) *Nature* **369**, 526-527.
- Stoltzfus, A. & Doolittle, W. F. (1993) *Curr. Biol.* **3**, 215-217.
- Stoltzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. & Doolittle, W. F. (1994) *Science* **265**, 202-207.
- Stuart, R. A., Cyr, D. M., Craig, E. A. & Neupert, W. (1994) *Trends Biochem. Sci.* **19**, 87-92.
- Swofford, D. L. (1993). PAUP; Phylogenetic Analysis Using Parsimony. (Illinois Natural History Survey, Champaign). Version 3.1.1.

- Tannich, E., Bruchhaus, I., Walter, R. D. & Horstmann, R. D. (1991) *Mol. Biochem. Parasitol.* **49**, 61-72.
- Taylor, F. J. R. (1974) *Taxon* **23**, 229-258.
- Taylor, F. J. R. (1976) *Taxon* **25**, 377-390.
- Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids. Res.* **22**, 4673-4680.
- Van de Peer, Y., Rensing, S. A., Maier, U.-G. & De Wachter, R. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 7732-7736.
- Viale, A. M. & Arakaki, A. K. (1994) *FEBS Lett.* **341**, 146-151.
- Vickerman, K. (1990). in *Handbook of Protoctista*, eds. Margulis, L., Corliss, J. O., Melkonian, M. & Chapman, D. J. (Jones & Bartlett, Boston), pp. 200-210.
- Vossbrinck, C. R., Maddox, J. V., Friedman, S., Debrunner-Vossbrinck, B. A. & Woese, C. R. (1987) *Nature* **326**, 411-414.
- Vossbrinck, C. R. & Woese, C. R. (1986) *Nature* **320**, 287-288.
- Wainright, P. O., Hinkle, G., Sogin, M. L. & Stickel, S. K. (1993) *Science* **260**, 340-342.
- Weiner, A. M. (1993) *Cell* **72**, 161-164.
- Whatley, J. M. (1976) *New Phytol.* **76**, 111-120.
- Woese, C. R. (1977) *J. Mol. Evol.* **10**, 93-96.
- Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J. & Woese, C. R. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4443-4447.
- Yang, Z. (1993) *Mol. Biol. Evol.* **10**, 1396-1401.
- Yang, Z. (1996) *J. Mol. Evol.* **42**, 294-307.
- Yang, Z., Kumar, S. & Nei, M. (1995) *Genetics* **141**, 1641-1650.
- Yarlett, N., Orpin, C. G., Munn, E. A., Yarlett, N. C. & Greenwood, C. A. (1986) *Biochem J* **236**, 729-39.
- Zhou, Y. & Ragan, M. A. (1995) *Curr. Genet.* **28**, 324-332.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved