STUDIES ON THE EVOLUTION OF THE SPLICEOSOME AND ORIGIN OF THE MICROSPORIDIA

by

Naomi M. Fast

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

at

Dalhousie University
Halifax, Nova Scotia
July 1999

0-612-49256-7

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 evolution of the spliceosome and origin of the microsporidia"

by Naomi Marya Fast

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

Dated: July 8, 1999

External Examiner

Research Supervisor

Examining Committee

[redacted]

DALHOUSIE UNIVERSITY

DATE: July 8, 1999

AUTHOR:	Naomi M. Fast
TITLE:	Studies on the evolution of the spliceosome and origin of the microsporidia
DEPARTMENT:	Department of Biochemistry

DEGREE: Ph.D. CONVOCATION: October YEAR: 1999

[redacted]

Signature of Author

To

Lord and Lady Blue Moon

TABLE OF CONTENTS

Introduction

Results and Discussion

Appendices

LIST OF ILLUSTRATIONS

ABSTRACT

The spliceosome is a large ribonucleoprotein complex composed of five small nuclear RNAs and over fifty proteins and is responsible for the removal of introns from pre-messenger RNA. Spliceosomal introns are prevalent among "crown" eukaryotes, whereas none has been found in those lineages thought to diverge earliest on the eukaryotic tree. However, the number of protein-coding genes sequenced from early-diverging eukaryotes is relatively low, creating a very small sample size. Since searching for introns directly in these organisms would be a nearly impossible task, I instead searched for evidence of spliceosomal components whose presence would be necessary for intron removal. The presence and conservation of such components could indicate that they are part of functional spliceosomes mediating the removal of spliceosomal introns.

The parabasalid *Trichomonas vaginalis* was found to possess a highly conserved homolog of the spliceosomal protein PRP8. Essential in all organisms known to splice, PRP8 interacts abundantly with the pre-mRNA substrate and is likely at the catalytic centre of the spliceosome. The *T. vaginalis* homolog is highly conserved in regions of functional significance, indicating that spliceosomes could be present in trichomonads. Likewise, the microsporidian *Nosema locustae* also possesses spliceosomal components. Two highly conserved genes for U6 and U2 snRNAs were found, the genes are expressed and the two RNAs have the potential to interact with each other in the functional pairings characterized in organisms known to splice. The microsporidian snRNAs are also predicted to be part of a functional spliceosome, and introns are predicted to be present.

During the course of this work, the ancient origin of the microsporidia has been called into question. Since the origin of the microsporidia bears on the interpretation of their possession of spliceosomal machinery, the phylogenetic position of the microsporidia was addressed by sequencing genes for TATA box binding protein (TBP), triosephosphate isomerase (TPI) and the second-largest subunit of RNA polymerase II (RPB2) from *N. locustae*. Taken as a whole, phylogenetic analyses with these sequences favour a much later origin for the microsporidia, and indicate a relationship between the microsporidia and the fungi. In addition, all genes are intron-lacking. A fungal origin for the microsporidia, coupled with the presence of old intron positions in TBP and TPI genes, suggests that spliceosomal introns have been lost throughout the course of microsporidian evolution.

ABBREVIATIONS

PCR	polymerase chain reaction
kb	kilobase
bp	base pair
kDa	kilodalton
pfu	plaque forming unit
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein
PRP8	pre-mRNA processing 8
TBP	TATA box binding protein
TPI	triosephosphate isomerase
RPB1	largest subunit of RNA polymerase II
RPA2	2nd largest subunit of RNA polymerase I
RPB2	2nd largest subunit of RNA polymerase II
RPC2	2nd largest subunit of RNA polymerase III
EF-1α	elongation factor-1 alpha
EF-2	elongation factor-2
NJ	neighbor-joining
K-H	Kishino-Hasegawa
Inr	initiator element
EST	expressed sequence tag
rRNA	ribosomal RNA
SSU	small subunit
LSU	large subunit
ML	maximum likelihood

ACKNOWLEDGEMENTS

Throughout the last four years I have been lucky enough to have the assistance and support of a great number of people. In many ways, the completion of this thesis is as much a reflection of their friendship and guidance as it is a reflection of the scientific experiments it describes. I am grateful to all who have provided help along the way, and particularly to those mentioned below.

First, I would like to thank Ford Doolittle. There always seemed to be time in Ford's lab to sit down in front of the journals and discuss something (scientific or otherwise); often this time was just as productive as that spent at the bench. Ford provided me with a great deal of freedom in my research and also in combining my research with my personal life. I am extremely grateful to him for both.

I could not have spent four months working in Melbourne, Australia without the kindness of Geoff McFadden and the members of his lab. I thank all of them for their friendliness and willingness to make space for me in a relatively small lab. Geoff is an excellent role model for keeping a healthy balance among time spent at work, home, and play; I thank him for his example.

This thesis was written in Bloomington, Indiana and could not have been completed without the facilities and friendliness provided by Jeff Palmer and those in his lab. I am very grateful for their kindness.

All of the graduate students, undergrads and postdocs that I shared space and time with in the Doolittle lab provided me with assistance and friendship in one way or another - thank you! In particular, I would like to

thank Claire Richardson for being the best honours student ever, along with being an excellent friend. I am grateful to Andrew Roger for founding the *Nosema* Library Club and to Patrick Keeling for showing me that there is almost always a molecular biology short-cut. Thanks to John Logsdon for a great deal of help along the way, including "business calls" from busy airports. I thank John Archibald for his honest and catching enthusiasm about work and for eating oatmeal every morning for breakfast while camping in Arizona.

Thank you to the Gray lab next door - particularly John Norman, Murray Schnare and Dave Price - for listening to gripes, helping with RNA work and for the very extended loan of a certain hyb tube!

Eve Teh and I made a pact to finish our PhDs before 2000, and I'm really thankful for all her help along the way. From coffee breaks to gourmet meals while exchanging work-gripes, life-gripes and laughter, she's been an excellent friend. I look forward to more fun in Vancouver!

Completing work at a distance would not have been possible without the help of Margaret Schenk, Wanda Danilchuk and Róisín MacKenzie. All helped me above and beyond their job descriptions and I'm extremely grateful.

My parents and sister have made several trips to visit me, and I appreciate all of their love and support. I thank my parents for buying a computer and learning about e-mail in order to stay in touch—especially handy while I was in Australia.

Finally, I thank Patrick for being enthusiastic about my work when I wasn't, for reading draft after draft of thesis chapters, and for the e-mails, phone calls and letters that kept us connected when we weren't physically in the same place. Thank you for loving me and providing me with a constant source of perspective.

INTRODUCTION

I. The discovery of spliceosomal introns and elucidation of the splicing pathway

In 1977 two research groups independently discovered that the human tumour virus Adenovirus late genes differed between their genomic organization and the mRNA produced, such that regions of the coding sequence contiguous in the message are not adjacent in the genome, but in fact are separated by many hundred bases (Berget, Moore and Sharp 1977; Chow et al. 1977). Evidence for the split nature of the Adenovirus genes came from two related findings: 1) the fact that all 'late' viral messages were found to have the same 5' terminal sequence, even though the more distal 3' sequences differed among genes, and 2) annealed complementary mRNA and genomic sequence viewed by electron microscopy revealed that large regions of the genomic sequence "looped out" with no complement in the mRNA. Further mapping studies linked the two findings by indicating that "looped out" intervening sequences in the DNA were flanked by sequences found to be adjacent in the RNA. In fact, the 200-nucleotide 5' leader sequence is itself composed in a tripartite manner from three non-contiguous segments in the genome! Quickly after the discovery of intervening sequences (or introns) in Adenovirus, other viral, plant, and human genes were found to be similarly split and sequence comparisons amongst introns revealed that all introns began with the dinucleotide GT and ended with the dinucleotide AG.

Early *in vivo* studies (mainly in yeast and viral systems) indicated that the removal of introns is an RNA processing event in the nucleus (Nevins 1979), and was given the name "splicing," referring to the removal of introns and ligation of exons. Further characterization of the splicing mechanism required the development of *in vitro* systems - an innovation which came about following two experimental advances. The first was the production of efficient run-off transcription systems so that accurate high-copy unspliced transcripts could be produced (Kassavetis et al. 1982; Green, Maniatis and Melton 1983). The second advance was the successful isolation of whole-cell (Kole and Weissman 1982; Padgett et al. 1983) and nuclear extracts (Hernandez and Keller 1983; Krainer et al. 1984) that support RNA splicing reactions. Whole-cell yeast extracts and nuclear HeLa extracts were found to be most active for splicing, and both systems require the addition of monovalent and divalent cations, along with ATP.

Employing these systems, and using a simple substrate composed of two exons flanking a single intron, the mRNA splicing pathway (see Figure I-1) was deduced. First, the splicing intermediates were characterized and S1 mapping indicated that, although the co-existence of products corresponding to the 5' exon and the intron-3'exon could be documented (identifying 5' cleavage), the 3' splice site was never cleaved in the absence of exon ligation, indicating the order of splice-site cleavage (Krainer et al. 1984). Another splicing intermediate is the excised intron. This RNA molecule is not linear, as was evident early-on from its mobility under electrophoresis, and was confirmed by cleavage studies indicating that it was indeed circular at its 5', but not its 3' end (Ruskin et al. 1984). An in-depth biochemical analysis of the intron "lariat" structure revealed that the 5' end of the intron was joined to an internal (and specific) adenosine in a 2'-5' phosphodiester bond, and labelling studies provided corroboratory evidence indicating that the phosphate at the branch site originates from the 5' splice-site terminal G (Ruskin et al. 1984; Padgett et al. 1984; Konarska et al. 1985).

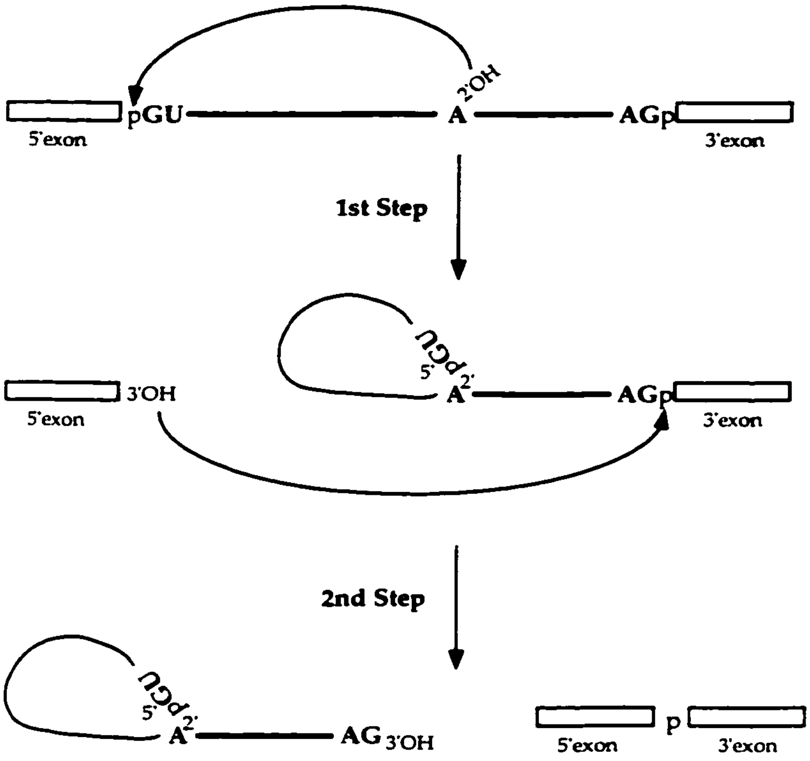


Figure I-1 mRNA splicing pathway
The two transesterification reactions of splicing.

These results predicted that splicing takes the form of a two step pathway, where the first step is 5' cleavage producing free exon1 and intron-exon2, and the subsequent second step involves 3' cleavage, release of the intron lariat and ligation of exon1 and exon2 (where the phosphate in the phosphodiester bond connecting the exons is that from the 5' end of exon2) (Ruskin et al. 1984; Padgett et al. 1984). As a linear intron-exon2 product was never identified, it was predicted that 5' cleavage and branch formation are coupled (Padgett et al. 1984; Ruskin et al. 1984). Likewise, since 3' splice-site cleavage is also never seen without exon ligation, these events were also predicted to be coupled. Such coupling, along with conservation in the number of phosphodiester bonds in splicing (5' splice-site phosphodiester bond cleavage coupled with 2'-5' phosphodiester bond creation, followed by 3' splice site phosphodiester bond cleavage coupled with phophodiester bond formation in the linking of the exons) strongly indicates that two transesterification reactions make up the splicing pathway (Padgett et al. 1984; Ruskin et al. 1984; Konarska et al. 1985).

II. Discovery of the splicing body or spliceosome

Since a transitory stage in the splicing reaction involves the free 5' exon and the intron-3' exon which must stay closely juxtaposed for the second step reaction, it was predicted that splicing might take place in a molecular complex to facilitate positioning of substrates for reaction efficiency (Hardy et al. 1984). This prediction was consistent with the observation that during *in vitro* splicing assays there is a lag time before splicing proceeds, which could indicate the formation of such a complex (Hernandez and Keller 1983; Krainer et al. 1984; Hardy et al. 1984). In 1980 two research groups made the suggestion that one such component of the predicted complex could be U1 small nuclear RNA (snRNA) (along with the proteins of its associated snRNP - ribonucleoprotein particle) based on the fact that the snRNA has the potential to base pair with the 5' splice site of introns (Lerner et al. 1980; Rogers and Wall 1980). Five years later, conclusive evidence for "splicing bodies" or "spliceosomes" came from analyses of *in vitro* HeLa and yeast systems indicating that a 60S (in human) and 40S (in yeast) complex must associate with the pre-mRNA for splicing to proceed (Grabowski, Seiler and Sharp 1985; Brody and Abelson 1985). Immunoprecipitation experiments revealed that the U1 snRNP was indeed a component of these complexes (Grabowski, Seiler and Sharp 1985). It quickly became clear that the other snRNPs known to be present in the nucleus (U2, U4/U6, and U5) were also part of the spliceosome and that conserved regions of each snRNA could pair with different conserved regions of the intron and/or another snRNA. The pathway of spliceosome assembly and predicted interaction of each snRNA is outlined and summarized in Figure I-2 and the accompanying figure legend.

III. Origins of spliceosomal introns and spliceosomal machinery

Upon the elucidation of the splicing pathway, and discovery of the spliceosomal components, immediate comparisons were drawn between the excision of Group II and spliceosomal introns. In 1986 it was found that organellar Group II introns underwent self-splicing with two transesterification reactions via a 2'-5' branched intermediate, with a hydroxyl at the branch site acting in the first transesterification reaction—identical at the biochemical level to the reactions of spliceosomal intron removal (see Michel and Ferat 1995). Potential similarities can also be seen between the structures within the Group II intron and structures of some of the snRNAs. Most snRNA comparisons have

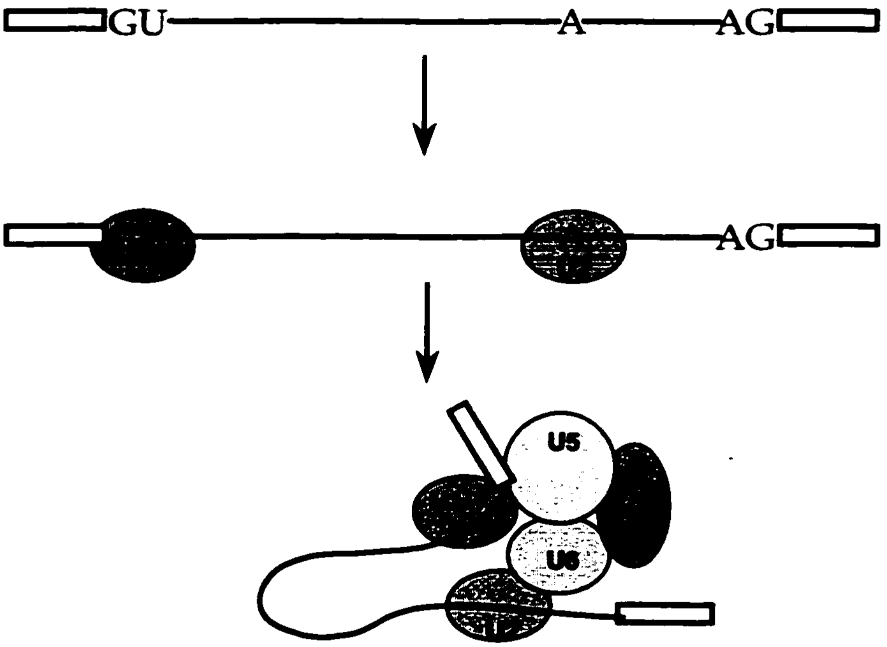


Figure I-2 Spliceosome assembly
Each oval represents an snRNP composed of one snRNA and its associated proteins, both common and snRNP-specific. The first step in spliceosome assembly is binding of the U1 snRNP to the 5' splice site by the pairing of U1 snRNA. Next, the U2 snRNP binds to the branch site by pairing with intronic regions around the nucleophilic adenosine (hence causing the adenosine to "bulge-out"). The U4, U5 and U6 snRNPs are thought to enter as a complex, with the U4 and U6 snRNAs base-paired together. This completes assembly of the spliceosome, although U4 snRNP is quickly displaced from the spliceosome as U6 snRNA pairs instead with U2 snRNA (the U6 snRNA interactions with U2 and U4 snRNAs are mutually exclusive). Through protein and RNA interactions within the spliceosome, the reaction substrates are brought into close proximity for the two transesterification reactions to proceed.

focused on U6 snRNA, as it has long been postulated to play a critical role in the splicing reaction due to its extreme sequence conservation compared with that of other snRNAs. The most conserved region of the Group II intron is Domain V, a stem-loop proposed to be functionally analogous (or even homologous) to an intramolecular stem-loop of U6 snRNA. This proposal is based on a putative correspondence of an essential arrangement of backbone phosphates in U6 snRNA with a similar arrangement in Domain V of the Group II intron (Yu et al. 1995). Also, when conserved nucleotides in the U6 snRNA are mutated, the splicing phenotype is similar to that seen when the putatively homologous nucleotides in Domain V of the Group II intron are mutated (Peebles et al. 1995; Boulanger et al. 1995). Perhaps the best evidence for a relationship between spliceosomal snRNAs and the Group II intron comes from a more recent experiment where supplying U5 snRNA in *trans* rescues a splicing-defective mutant Group II intron (Hetzer et al. 1997).

Although specific similarities between snRNAs and Group II intron domains are a bit dodgy, the overall resemblance of the two splicing systems does suggest a potential common ancestry, an idea that has been taken so far as to refer to the splitting of a Group II intron into spliceosomal snRNAs as "Five Easy Pieces" corresponding to domains of the Group II intron (Sharp 1991). In addition, the similarities between spliceosomal intron excision and self-splicing Group II introns make up perhaps the biggest piece of circumstantial evidence supporting the hypothesis that spliceosomal intron excision is RNA catalyzed. (See, for example, discussion in Nilsen 1998.)

Even if one accepts that Group II introns and spliceosomal snRNAs are evolutionarily related, one still has to account for the extremely large protein component of the spliceosome. There are likely over 100 proteins in the human spliceosome, and it may be an oversimplification to attribute their presence as being a scaffold merely "presenting" the snRNAs. Although spliceosomal proteins continue to be characterized, there is hardly a complete catalog of those present. Many of the spliceosomal proteins are RNA helicases/ATPases, some with known specific functions (i.e. unwinding the U4/U6 helix), while the roles of other helicases remain to be identified. Other proteins in the spliceosome contain RNA recognition motifs and RS domains, implicating their activity in binding single-stranded RNA and also promoting annealing of complementary strands (Staley and Guthrie 1998). Still other proteins play essential roles within the spliceosome that have yet to be determined.

If the spliceosomal intron and snRNAs really are a Group II intron "in pieces," where did the large proteinaceous component of the spliceosome come from? And why is it necessary if spliceosomal intron excision really is an RNA-catalyzed reaction? Although there are no clear-cut and proven answers to either of these questions, they are interesting points of discussion to consider. First, it is important to note that Group II introns *in vivo* do not splice in the absence of protein factors (see Michel and Ferat 1995). Unfortunately, these factors are for the most part unknown and uncharacterized, so it remains to be seen whether they are at all related to any of the spliceosomal proteins. It could be envisioned that, although spliceosomal intron splicing may at one time have been completely RNA mediated, proteins became associated with the RNAs in a neutral way (Stoltzfus 1999). Over time, the RNA components could change to rely on the presence of these proteins, so that the proteins could come to be necessary for accurate and/or efficient splicing. As more proteins are added, protein-protein interactions would also come into play, and could also be "fixed" in the same way. It is a puzzle to determine where the proteins came from, although there is an intriguing case where a protein in the U5 snRNP (SNU114 in yeast; U5-116K in human) possesses a high degree of sequence similarity to the translation factor EF-2 (Fabrizio et al. 1997). It will be interesting to see whether the function of SNU114 within the spliceosome is at all similar to the role that EF-2 plays during translocation within the ribosome. There is also the possibility that proteins may have been added in groups. A recent report indicates that several proteins within the U5 snRNP are found together as a stable complex even in the absence of RNA (Achsel et al. 1998); however, it is questionable whether this association arose before or after co-option by the spliceosome. The genesis of the spliceosome could be a very good example of Constructive Neutral Evolution (as defined in Stoltzfus 1999).

IV. Archezoa and the absence of spliceosomal introns

The similarities of Group II introns to spliceosomal snRNAs and the shared splicing biochemistry of Group II and spliceosomal introns led Cavalier-Smith to propose (in 1991) that spliceosomal introns arose from a progenitor Group II intron imported with the genome of the alpha-proteobacterial symbiont that became the modern-day mitochondrion (Cavalier-Smith 1991). This hypothesis fit well with the picture of eukaryotic evolution present at that time. Spliceosomal introns were found to be present in those eukaryotes diverging after the acquisition of the mitochondrion, whereas those eukaryotes appearing to lack mitochondria and perceived to branch prior to organelle acquisition also appeared to lack spliceosomal introns (Cavalier-Smith 1991; Palmer and Logsdon 1991). Previously, Cavalier-Smith had named this group of mitchondrion-lacking protists the Archezoa (Cavalier-Smith 1983).

The Archezoa was created as a phylogenetic hypothesis to draw attention to a group of protists proposed to represent the earliest of eukaryotes. Archezoa exhibit relative morphological simplicity, and include Metamonada, Parabasalia, Archamoebae and Microspora. All members lack recognizable mitochondria and peroxisomes, while golgi and flagella are present in some groups but not others. Once molecular data (small subunit and large subunit rRNA sequences) were available from two members of the Archamoebae, both were found to branch well within the mitochondrial eukaryotes casting serious doubt on the validity of the Archamoebae as ancient (Sogin 1989; Hinkle et al. 1994; Morin and Mignot 1995). Evidence also came to light that *Entamoeba histolytica* possesses the gene for mitochondrial chaperonin 60 (*cpn*60), indicating that this organism is secondarily amitochondrial (Clark and Roger 1995). In contrast, phylogenetic analyses of small subunit rRNA sequences from the other archezoa were in agreement with their position as diverging prior to mitochondrial eukaryotes.

The Metamonads include diplomonads, retortamonads, and oxymonads; however, the most studied of the metamonads are the diplomonads, especially *Giardia lamblia*. Diplomonads have four flagella radiating from nuclear associated kinetids, with three flagella anterior and a recurrent flagellum generally used for feeding that is found within a cytostomal groove. There are both "monozoic" and "diplozoic" forms (for which the taxon bears its name), where the latter exhibit two-fold symmetry (like *Giardia*), appearing like a fusion of two "back-to-back" monozoic forms (see Vickerman 1990).

Evidence supporting the deep position of the diplomonads comes not only from phylogenetic analysis of small subunit rRNA, but also from the analyses of large subunit rRNA, EF-2, EF-1alpha, and the largest subunit of RNA polymerase II (Sogin et al. 1989; Hashimoto et al. 1994; Hashimoto et al. 1995; Klenk et al. 1995; Cavalier-Smith and Chao 1996). However, during the course of this thesis work, the amitochondrial nature of the diplomonads has been seriously called into question. Phylogenies of three genes believed to be of mitochondrial origin (GAPDH, TPI and valyl-tRNA synthetase) have also been analyzed from diplomonads and the sequences do not seem extremely different from those of mitochondrial-containing eukaryotes, implying that the diplomonad genes have not had a different origin (Rozario et al. 1996; Keeling and Doolittle 1997; Hashimoto et al. 1998). More conclusive evidence supporting the claim that diplomonads (or at least *Giardia*) at one time possessed mitochondria came from the discovery of the mitochondrial-type chaperonin 60 gene in *Giardia* (Roger et al. 1998). In sum, although diplomonads may represent early-branching eukaryotes, it is doubtful that they are ancestrally amitochondrial.

A similar situation exists for another archezoan group, the Parabasalia. The Parabasalia include trichomonads, monocercomonads, devescovinids, calonymphids, and hypermastigotes, and are united by the presence of a kinetosome-associated golgi apparatus known as the parabasal body (see Dexter Dyal 1990). Another parabasalid feature is the presence of a DNA-free organelle called the hydrogenosome, which produces hydrogen gas and acetate from pyruvate or malate. Recently, evidence has accumulated which indicates that the trichomonad hydrogenosome is related to, and is likely derived from, the mitochondrion. It has been found that the nuclear genome of *Trichomonas vaginalis* possesses genes for three chaperonins (*cpn*10, *cpn*60 and *cpn*70) known to function in the mitochondria, along with another gene found to be of mitochondrial origin in other eukaryotes, valyl tRNA synthetase (Germot, Philippe and Le Guyader 1998; Bui, Bradley and Johnson 1996; Horner et al. 1996; Roger, Clark and Doolittle 1996; Hashimoto et al. 1998). Indeed, phylogenetic analyses with the chaperonin sequences indicate their close relationship with the other mitochondrial chaperonins, and the valyl tRNA synthetase sequence branches with those from eukaryotes, indicating that it likely shares the same ancestry. Even stronger support for the relationship between the mitochondrion and the hydrogenosome comes from the determination that at least some of these chaperonin proteins are localized in trichomonad hydrogenosomes (Bui, Bradley and Johnson 1996). Furthermore, the *T. vaginalis cpn60* gene's inferred translation product includes an N-terminal extension that is also seen in hydrogenosome metabolic enzymes known to function in that organelle (Bui, Bradley and Johnson 1996). Certainly the primitive amitochondrial status of the parabasalia is unlikely; however, there is currently no molecular phylogenetic evidence to indicate that the parabasalia are not early-diverging eukaryotes.

In contrast, in the last three years the primitive nature of the microsporidia has been seriously questioned due to evidence that has come to light throughout the course of my thesis work, in addition to research described in my thesis. Microsporidia are obligate intracellular parasites that are always found as chitin-walled spores outside of a host cell. They exhibit an intriguing method of infecting potential hosts by everting a polar tubule that is otherwise tightly coiled within the spore. Upon piercing a host cell, the polar tube provides a means for the spore contents to enter the host cytoplasm, where the parasite grows and divides to eventually produce more spores (see Canning 1990; Keeling and McFadden 1998). With strategies similar to those employed to determine that diplomonads and trichomonads once harboured a mitochondrion (or a similar symbiont), microsporidia have also been found to possess genes of mitochondrial origin, even though no such organelle has been identified. Genes for chaperonin 70 have now been characterized from three microsporidia, and all branch within the mitchondrial clade, increasing doubt that microsporidia are ancestrally amitochondrial (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998).

At the onset of my thesis work, the only molecular phylogenetic evidence available for microsporidia supported their deep placement within the eukaryote tree, and came from analysis of multiple molecules: small subunit rRNA, large subunit rRNA, and elongation factors EF-2 and EF-1alpha (Vossbrinck et al. 1987; Kamaishi et al. 1996a; Kamaishi et al. 1996b). Although phylogenetic reconstruction with EF-1alpha sequences indicates an early divergence of the microsporidia, the microsporidian sequence contains an insertion shared by animals and fungi (Kamaishi et al. 1996). Another feature shared by microsporidia and animals/fungi is the presence of separate dihydrofolate reductase and thymidilate synthase genes, which are otherwise fused in plants, green algae, *Plasmodium*, and several other protists (Vivarès et al. 1996). In 1996, the first molecular phylogenetic evidence supporting a later divergence for the microsporidia came with the analysis of alpha- and beta-tubulin genes (Edlind et al. 1996; Keeling and Doolittle 1996). Other more recent analyses have supported the conclusion from the tubulin phylogenies, and these will be discussed in the context of the appropriate thesis chapters.

As mentioned above, parabasalia and microsporidia appeared to lack spliceosomal introns, based on the (admittedly limited) amount of sequence data available as I began my thesis research. Since looking for introns directly would be a nearly impossible task, I chose to search for the components of the spliceosomal machinery that would necessarily be present for intron removal, if such introns did exist. In the absence of complete sequence data, this may be the best, albeit indirect, method to predict intron presence.

In the parabasalid *T. vaginalis*, I found a gene encoding the essential spliceosomal protein PRP8 and in the microsporidian *Nosema locustae*, I found evidence for three snRNA components of the spliceosome. In both cases, the spliceosomal components were conserved to the extent that I predict that they could be parts of functioning spliceosomes, and that introns could be present in both parabasalia and microsporidia.

To more closely examine the intron status and phylogenetic position of the microsporidia, I also characterized the genes encoding triosephosphate isomerase, TATA box binding protein, and second largest subunit of RNA polymerase II from *N. locustae*. Results suggest that microsporidia are not ancient eukaryotes, but diverged much later, possibly as a relative of the fungi. In addition, all three microsporidian genes are intron-lacking despite the presence of conserved "old intron positions" in two of the genes examined, indicating that these spliceosomal introns have been lost throughout the course of microsporidian evolution.

MATERIALS AND METHODS

Nucleic acid sources and extraction procedures.

Nosema locustae spores were obtained from L. Mearril of the M&R Durango biocontrol company in Colorado. Approximately 10^{10} spores were pelleted and then resuspended in 500 µl of an extraction buffer composed of 50 mM Tris-HCl (pH 7.5), 50 mM EDTA, 3% SDS and 1% 2-mercaptoethanol. The suspension was added to chilled mortar and pestle, and spores were ground with liquid nitrogen for several hours. Whilst grinding, several millilitres of extraction buffer were added, and the degree to which spores were broken was tested by examining small samples under a light microscope. Once it was deemed that ~75% of the spores were broken, they were scraped into a eppendorf tube and resuspended in 1 ml of the extraction buffer. Proteinase K was added to a final concentration of 200 µg/ml and incubated for 1 hour at 50°C. Following standard phenol, phenol/chloroform/iso-amyl alcohol, and chloroform extractions, the DNA was ethanol-precipitated. To remove co-precipitating carbohydrates, the DNA was extracted with cetyltrimethylammonium bromide (CTAB) employing a protocol developed to purify DNA from carbohydrate-rich protist cells. This procedure involves resuspending the DNA in a 0.7 M NaCl and 1% CTAB solution, and subsequent incubation at 65°C for 30 minutes. Two chloroform extractions were performed, carefully removing the aqueous layer, and leaving behind the carbohydrate-CTAB layer that forms between aqueous and organic phases. DNA was then precipitated with ethanol.

LiCl-precipitated *N. locustae* RNA was a generous gift from Dr. A. Roger and Dr. J. Brown.

Trichomonas vaginalis (strain NIH-C1, ATCC# 30001) genomic DNA was kindly provided by Dr. M. Müller and J. Lee (Rockefeller University, New York) and *Naegleria andersoni* (strain PPFMB-6) genomic DNA was a gift from Dr. S. Kilvington (Bath Public Health Service, Bath).

Primer design and PCR conditions

Genes of interest were amplified by polymerase chain reaction (PCR) using degenerate oligonucleotides (or primers). PCR primers were designed to mirrir regions encoding conserved blocks of amino acids of the protein, as ascertained by an alignment of representative taxa. In general, primers were 17-21 nucleotides in length, with the 3′-most 10 nucleotides left fully degenerate, and the remaining 5′ nucleotides fixed to balance G+C content. Attempts were made to design primers for regions that do not include the residues S, L or R, since these amino acids exhibit 6-fold degeneracy.

Twelve different primers were designed to conserved regions of the PRP8 protein, to be used in all possible combinations. Taxon sampling was low for this protein; at the time that primers were designed the alignment only included two sequences in some regions. The *N. andersoni* PRP8 partial sequence (~1400 bp) was amplified with primers Spl8-2 (5′-CCCATGAARTTRTARTTCCA) and Spl8-3 (5′-GCTAAGTGGAARACNGCNGA). The *T. vaginalis* PRP8 partial coding sequence (~400 bp) was amplified with Spl8-1 (5′-ATCCATAAGACGTTYGARGG) and Spl8-6 (5′-AAGAGTACCATYTGNGGYTC). Using this PCR product as a probe to screen the *T. vaginalis* genomic library only retrieved genomic clones containing approximately the 3′ third of the entire gene. To make another probe to attempt to obtain the 5′ region of the gene from the genomic library, PCR was used with an exact match primer to the 5′-most part of the *T. vaginalis* PRP8 known (Spl8-exR: 5′-TGATGCGAGGAATTCCTCTG) and a degenerate primer designed to a further upstream conserved region (Spl8-g: 5′-GGACGAGCNGTNTTYTGG). To amplify nearly the entire alpha-tubulin gene adjacent to the *T. vaginalis* PRP8 gene, I designed exact-match primers for non-coding regions just downstream of the alpha-tubulin gene (to ensure that I amplified the same alpha-tubulin gene copy that is next to PRP8). TvaTubby1: 5′-GAGGAAATGATAGTCACGCC; TvaTubby2: 5′-TATGAATGAAATTGGATCTG. Both were used in conjunction with the standard AtubA oligo previously described (Keeling and Doolittle 1996) (5′-TCCGAATTCARGTNGGAAYGCNTGYTGGGA).

The *N. locustae* TBP partial gene sequence (~320 bp) was amplified with TBP-F1 (5′-AACGCAGARTAYAAYCC) and TBP-R1 (5′-CACGATCTTACCNSWNACRAA), and also with TBP-F1: TBP-R2 (5′-GAACAGYTCNGGYTCRTA). The PCR product from the TBP-F1:TBP-R1 reaction was used as a probe to screen the *N. locustae* genomic library.

The *N. locustae* TPI gene fragment (~200 bp) was amplified with TPI-F2 (5′-TGGGCTATHGGNACNGG) and TPI-R1 (5′-AGCTCCACCGACNARRAANCC).

The *N. locustae* RNA polymerases I, II, and III second-largest subunit partial sequences (~1100 bp) were amplified with primers designed and kindly provided by T.M. Embley and R. Hirt: RPB2-F2 (5′-CCNTTYCCNGAYCAYAAYCA) and RPB2-R2Q (5′-TGRCARTCNCKYTCCATYTC).

Standard PCR reactions were 50 µl in volume and contained 1xPCR buffer (20 mM Tris-HCl, 50 mM KCl), 1.5 mM $MgCl_2$, 10 mM dNTP (each), 1 µM each primer, 2 U *Taq* DNA polymerase, and 0.5 U *Pfu* DNA polymerase. To this mixture, 50-100 ng of template DNA was added. Standard cycling conditions were 2 minutes at 94°C; 1 minute at annealing temperature; 1 minute at 72°C, followed by 34 cycles of 1 minute at 92°C; 1 minute at annealing temperature; 1 minute at 72°C, followed by 1.5 minutes at 72°C. Annealing temperatures varied from 45-55°C depending on the degree of stringency attempted. *Taq* polymerase was purchase from Gibco BRL and *Pfu* polymerase was purchased from Stratagene. PCR products were analyzed by gel electrophoresis.

To quickly screen *E. coli* transformants to ascertain whether they harboured a vector with the appropriate sized insert, I used a PCR screening approach. This involved picking white colonies off the transformation plate with a sterile pipet tip. First the tip was touched to a "master plate" and then was used to "inoculate" a 10 µl PCR reaction composed of 0.2 U Taq polymerase, 1.5 mM $MgCl_2$, 1xPCR buffer (as above), 5 mM dNTPs and 0.5 µM M13 -20 and reverse primers. The reactions were then cycled 30 times for 1 minute at 94°C; 1 minute at 57°C; 1 minute at 72°C. A polishing step of 5 minutes at 72°C ended the program. The products were then separated by agarose electrophoresis, so the "true" clones could be ascertained and then picked from the master plate for growth of overnight cultures and subsequent plasmid preparation and analysis.

Gel isolation and cloning.

DNA fragments were separated by standard agarose electrophoresis (0.8%-1.5%) buffered with TAE (See Appendix). Excised agarose slices were crushed by a sterile glass rod in 100-150 µl of TE (See Appendix). An equal volume of 1:1 phenol:chloroform was added and the mixture was vortexed vigourously for at least 1 minute, and immediately frozen at -70°C for 10 minutes. After thawing at room temperature, and centrifugation at maximum speed for 15 minutes, the DNA-containing supernatant was removed from above the agarose layer. DNA was precipitated by the standard ethanol procedure, employing 1 µg of yeast tRNA as a carrier.

PCR products were cloned into the T-vector pCR2.1 (Invitrogen) with T4 ligase, as specified by the manufacturer. The vector pBlueScript SK+ was used for the subcloning of genomic clones isolated from library screening.

E. coli strains, maintenance and manipulations

Transformation of plasmids into *E. coli* cells was achieved by one of two methods. Competent *E. coli* strain DH5α was electrotransformed with the BioRad Gene Pulser at 1.8 kV and 200 ohms, across an electrode gap of 0.1 cm. After recovery in SOC medium for 30 minutes at 37°C, cells were plated on a selective LB medium and incubated overnight at 37°C. Alternatively, transformation was achieved by heat shock with chemically competent *E. coli* strain INVαF' (Invitrogen), following the Original TA Cloning Kit directions (Invitrogen) exactly. In both methods, LB solid medium was overlaid with 80 µl of 2% X-Gal in dimethyl formamide to allow blue-white screening.

Positive clones containing the desired insert size were identified by PCR screening, as described above.

Clones were standardly grown overnight in selective LB broth for plasmid preparation. Standard alkaline lysis preparation was employed with commercial kits using ion-exchange spin columns: initially the NucleoSpin Mini-prep Kit (Machery-Nagel) and later the Eclipse Mini Plasmid Prep Kit (Eclipse Molecular Biologicals, distributed by Gordon Technologies). Good quality plasmid was produced in sufficient concentration for sequencing or restriction enzyme digestion.

Genomic library construction

Both *N. locustae* and *T. vaginalis* genomic libraries were constructed in the calf alkaline phosphatase dephosphorylated and *Bam*HI pre-cut phage vector Lambda Zap Express (Stratagene). Genomic DNA was isolated from *N. locustae* spores as described above and *T. vaginalis* genomic DNA was a gift from M. Müller and J. Lee (Rockefeller University, New York). In constructing both libraries, the goal was to partially digest genomic DNA with *Sau*3AI and incorporate 6-10 kb fragments into the vector arms. Test *Sau*3AI (New England BioLabs, with supplied 10x buffer) digests were carried out on the genomic DNA for different time periods in an attempt to maximize the amount of 6-10 kb fragments produced. Once determined, these same conditions were used to digest 6 µg of *N. locustae* DNA and 10 µg of *T. vaginalis* DNA. Fragments were size-selected by agarose gel isolation and DNA was isolated from the agarose slice by the Prep-A-Gene kit (BIO-RAD). Test ligations and packaging experiments were carried out with a small amount of the *N. locustae* or *T. vaginalis* genomic DNA along with test insert and arms-only controls as specified by the manufacturer. Comparison of the controls with the *N. locustae* or *T. vaginalis* results allowed the insert amount to be scaled up accordingly and used in the ligation reaction. Ligation took place with 1.0 µg of lambda arms and 200 U of T4 DNA ligase (New England Biolabs), and was incubated overnight at 14°C followed by further incubation for 24 hours at 4°C. The ligation was packaged into phage heads with the Gigapack II packaging kit (Stratagene) in three extracts and plated, following the manufacturer's instructions exactly. Titering of the *N. locustae* and *T. vaginalis* primary libraries revealed that they contained 8.3 x 10^4 and 4 x 10^5 recombinants, respectively. In both cases, taking the estimated genome size into account (6 Mb for *N. locustae* and 10 Mb for *T. vaginalis*), and assuming an average insert size of 6 kb, statistically there should be a greater than 99% chance that any given DNA segment is represented in each library. Additional blue-white screening (by the addition of X-gal and IPTG) showed that less than 1% of the recombinants lacked inserts. Following the manufacturer's directions, all of the *N. locustae* and two-thirds of the *T. vaginalis* libraries were amplified. Titers of the newly amplified *N. locustae* and *T. vaginalis* libraries were 3 x 10^8 pfu/ml and 6 x 10^8 pfu/ml, respectively.

Genomic library screening

In the case of the *N. locustae* TPI, TBP and *T. vaginalis* PRP8, probes were generated from PCR products. Plasmid clones were digested with restriction enzymes to release the PCR product DNA fragment that was purified by gel isolation (described above). Random-prime labelling (Prime-It II, Stratagene) with α-^{32}P-dATP produced labelled probes.

For the *N. locustae* U6 and U2 snRNA probes, oligonucleotides were designed to highly conserved regions of the snRNA genes. The U6 probe is complementary to the "catalytic core" of U6: U6-36, 5'-GCAGGGGCCATGCTAATCTTCTCTGTATAATTCCAA. The U2 probe is complementary to the branch point recognition region of U2 snRNAs: U2-L15, 5'-CAGATACTACACTTG. Both oligomer probes were labelled by 3'-tailing using terminal deoxynucleotidyl transferase (TdT, Promega). In the presence of α-^{32}P-dATP and 20 U TdT, 2 pmol of oligomer was incubated at 37°C for thirty minutes.

Both the *N. locustae* and the *T. vaginalis* genomic libraries were screened in the same manner (as briefly described below), with the only exception being that the number of primary plates screened differed. Standardly, four 150 mm *N. locustae* plates were screened, while twelve *T. vaginalis* plates were screened.

Cultures of XL-1 Blue MRF′ were grown to mid-log stage in LB broth supplemented with 0.2% maltose and 10 mM $MgSO_4$. Cells were pelleted and resuspended at OD_{600}=0.5. For 150 mm plates, an appropriate amount of genomic library stock (determined to give ~15,000 pfu/plate) was added to 600 µl of culture, and incubated at 37°C for 15 minutes. Following incubation, 9 ml of melted NZY top agar was added, mixed and poured onto a dried, pre-warmed 150 mm NZY plate. This protocol was repeated as needed and plates were incubated overnight at 37°C.

Plates were chilled at 4°C prior to performing plaque lifts. Charged nylon membranes (Hybond N^+, Amersham) were employed as follows. A dry membrane was placed on the plate for 2-3 minutes, and orientation of the membrane was marked by india ink with a needle. To denature the phage DNA, the membrane was then placed plaque-side-up in a pool of 0.5 N NaOH for 2 minutes, blotted and then repeated with fresh NaOH. Neutralization was achieved by the same method with 1.0 M tris-HCl, pH 7.5. After air-drying, DNA was further fixed to the membrane by UV crosslinking with the UV Stratalinker (Stratagene) at the standard setting.

Hybridization took place in a rotary hybridization oven (Hybaid). All hybridizations took place overnight, and temperature varied depending on the probe. Hybridization with PCR product probes took place at 65°C, whereas hybridization with the U6-36 and U2-L15 oligomer probes took place at 60°C and 37°C, respectively. For all PCR product probes, and the U6-36 oligomer probe, the hybridization solution contained 0.5% nonfat dry milk powder, 4xSSC and 1% SDS. Hybridization with the U2-L15 probe occurred in a solution with a higher salt concentration: 0.5% nonfat dry milk powder, 5xSSC, 5 mM EDTA, and

1% SDS. All washes were conducted in 2xSSC, 0.5% SDS. For all PCR product probes, and the U6-36 oligomer probe, washes included one room temperature wash, and two 15 minute washes at the restrictive temperature. Membranes probed with the U2-L15 oligo were washed once at room temperature for 5 minutes and twice at 32°C for 15 minutes.

Following washing, membranes were exposed to film overnight at -70°C in the presence of intensifying screens, and the film was then developed.

Regions of the plates that corresponded to signals on the autoradiographs were cored into 1 ml of SM buffer and 20 µl of chloroform and left at room temperature for ~5 hours and vortexed frequently to allow the phage to be released into the buffer. Serial dilutions of these "primary" cores were used in "secondary" infections and screenings as described above, with the only exception being that volumes differed to accommodate the smaller 100 mm plates.

Cores from secondary plates that contained only a single plaque forming unit were incubated in 500 µl of SM buffer and 20 µl of chloroform to release the phage. To isolate plasmid DNA with the insert of interest, in vivo excision was performed with the ExAssist helper phage and the XLOLR strain of *E. coli*, following the Stratagene Lambda Zap Express "Single-Clone Excision Protocol" exactly. After selection of clones containing the pBK-CMV phagemid on LB-kanamycin (50 µg/ml) solid medium, colonies were grown and plasmids isolated. Following restriction digestion of the plasmid, and separation by agarose electrophoresis, the restriction fragments were blotted to a nylon membrane and Southern blots were performed with the same hybridization and washing conditions as employed in the library screening. This identified positive clones, and also identified fragments for subcloning into pBlueScript SK+, as was done for *N. locustae* U6 and U2 snRNA genes, along with the *T. vaginalis* PRP8 gene.

Southern blotting

Southern blotting was performed in order to confirm the origin and copy number of genes of interest, and also to identify positive genomic clones and further identify restriction fragments for subcloning.

For genomic Southerns, genomic DNA of interest was digested by a number of restriction endonucleases and then separated by agarose gel electrophoresis. The gel was treated first in 0.25 M HCl to depurinate and nick the DNA, followed by 0.4 M NaOH to denature the DNA. DNA was transferred to a charged nylon membrane (GeneScreen Plus Hybridization Transfer Membrane, Dupont) as follows. Capillary transfer was employed with a transfer buffer of 10xSSC. A 3 MM filter paper wick was constructed on a solid support over the reservoir of transfer buffer. The gel was placed face-down on top of this, and surrounded by plastic wrap to prevent "short-circuiting." The membrane, cut to the size of the gel, was first immersed in 2xSSC and then placed on top of the gel, ensuring that all air bubbles were removed. Two sheets of 2xSSC-soaked pieces of gel-sized 3 MM filter paper were placed on top of the membrane. A large stack of paper towels or Kim-Tuffs was placed on top and weighted. Transfer was allowed to proceed overnight.

For the transfer of digested plasmid DNA, dry blots were used. Following agarose electrophoresis of the samples, the gels were treated prior to transfer as described above. Transfer was achieved by laying the gel on a piece of plastic wrap on the bench-top, and overlaying the 2xSSC soaked membrane and two pieces of 3mm filter paper. As above, a stack of paper towels or Kim-Tuffs was added and weighted. Transfer was allowed to proceed from 5 hours to overnight.

DNA was fixed to the membrane by UV-crosslinking at the standard setting on the Stratalinker (Stratagene). Hybridization and washing conditions were as described for library screening.

Northern blotting and hybridization

Northern analysis of *N. locustae* RNA was employed to assay the expression of three snRNA genes. Approximately 1 µg of LiCl-isolated RNA was separated on a 6% polyacrylamide gel. The fragmented *Euglena gracilis* ribosomal RNA was used as a size standard, and was kindly provided by Dr. Y. Watanabe. Transfer was achieved by capillary transfer as described above for genomic Southern blots; however, in this case the transfer buffer was 5 mM NaOH. Transfer was allowed to proceed overnight and the RNA was then fixed to the blot by baking overnight at 65°C.

Both the U6 and U2 snRNA probes were made by incorporating α-^{32}P-dCTP in PCR reactions, using exact-match primers to internal regions of the gene. PCR primer sequences: U6ampF: 5′-TTAGTTTGGAACAACACTGAG, U6ampR: 5′-CACCTCTCAAAGAAAGATG, U2ampF: 5′-AGCCCTCCACCTCTCAAAGC, U2ampR: 5′-CCAGCATATTCTAGCTCCAAG. The U4 snRNA probe was a 17-nucleotide oligomer, U4-comp17: 5′-GGCCCCTGCGCAAGGAC, designed to be complementary to a predicted region of U4 snRNA. U4 snRNA sequence could be predicted based upon the known U6 snRNA sequence and its characterized regions of base-pairing with U4 from organisms in which splicing is known to occur. This oligomer was labelled by 3'-tailing, as described for the labelling of the probe used for library screening, U2-L15.

Prehybridization and hybridization of the blot took place in Church's Buffer (see Appendix), at 65°C with the U6 and U2 snRNA probes, and at 38°C for the U4 snRNA probe. Following overnight hybridization, the blots were washed. Blots hybridized with the U6 and U2 snRNA probes were washed in 1xSSC, 0.5% SDS once at room temperature for 5 minutes and once at 65°C for 30 minutes. The U4 northern blot was washed in 2xSSC, 0.5% SDS once at room temperature for 5 minutes, once at 38°C for 20 minutes and once in 1xSSC, 0.5% SDS at room temperature for 30 minutes. All blots were exposed to film in the presence of intensifying screens for variable amounts of time (48 hours to 1 week) depending on signal intensity.

DNA Sequencing

To determine whether cloned PCR products were the sequence of interest, positive clones identified from PCR screening were grown, and the plasmids sequenced manually with α-^{35}S-dATP. Dideoxy sequencing reactions were performed with T7 DNA polymerase using the ^{T7}Sequencing Kit (Pharmacia) and the instructions were followed exactly, with the exception being that the reactions were performed in microtiter dishes on top of the hot block, instead of using individual tubes. Reactions were resolved by electrophoresis on 6% polyacrylamide gels, which were dried onto 3 MM Whatman paper and exposed overnight prior to filmdevelopment.

Polishing of sequences was done by automated sequencing (ABI or LiCor) and carried out by the Joint Lab Sequencing Service (NRC). Walking primers used for achieving polished double stranded sequence are listed in the Appendix.

RNA secondary-structure prediction and primer extension analysis

For predicting *N. locustae* RNA secondary structures, computer-assisted free energy minimalization was carried out with the mfold server (http://www.ibc.wustl.edu/~zuker/rna/form1.cgi), followed by additional folding by eye. Secondary structures were considered improbable (and discarded) unless they were in agreement with the alignment.

The *N. locustae* U6 snRNA 5′ end was assigned by primer extension. A primer was designed based on the genomic sequence (U6pext-1: 5′-GAAAGATGCCGTCCTTGCGC) and was 5′-end labelled with polynucleotide kinase. A 13 µl cocktail of 10 U of PNK (Pharmacia), 1x supplied buffer, 5 mM dithiothreitol (DTT), 1 mM spermidine, 1 µl gamma-^{32}P-dATP was incubated at 37°C for 30 minutes. The reaction was then heated at 65°C for 5 minutes, ethanol-precipitated and resuspended in 3.5 µl of water. The labelled primer was mixed with 10 µl of *N. locustae* RNA (approximately 10 µg) and 1.5 µl of 10x hybridization buffer (1.5 M KCl, 0.1 M Tris-HCl, pH 8.3, 10 mM EDTA). Following a 90 minute incubation at 65°C, the mixture was cooled at room temperature. A 30.3 µl mixture of the following reagents was prepared and added to the RNA-primer mix: 30 mM Tris-Cl, pH 8.3, 15 mM $MgCl_2$, 8 mM DTT, 0.2 mM (each) dNTP mix, 5 U AMV reverse transcriptase (Promega). Primer extension was carried out for 1 hour at 42°C. Following extension, 105 µl of RNase mix (100 µg/ml salmon sperm DNA, 20 µg/ml RNaseA in TEN 100) was added and the mixture incubated at 37°C for 15 minutes. 15 µl of 3M NaOAc was added and extraction with 150 µl of phenol/chloroform/isoamyl alcohol was performed. The aqueous layer was precipitated and washed with ethanol, and the pellet dried at 42°C. Loading dye (provided with the ^{T7}Sequencing Kit [Pharmacia]) was used to resuspend the pellet.

For comparison, the U6pext-1 primer was used to sequence from the genomic clone employing the ^{T7}Sequencing Kit, using α-^{32}P-dATP. This reaction was run alongside the primer-extension product on a 9% denaturing polyacrylamide gel. The gel was transferred to a solid support and exposed to film overnight at -70°C with intensifying screens.

Sequence Management

Nucleotide sequences were entered into EditSeq (LaserGene) and compared to the database using BLAST (Basic Local Alignment Search Tool) to identify potential genes. DNA Strider 3.1 was used to translate nucleotide sequences and identify open reading frames. For contig assembly, Seqman (LaserGene) was originally employed, while Sequencher was later adopted as the program of choice.

Phylogenetic analysis

Conceptual translations of the genes of interest were aligned with other representative taxa using PIMA (Pattern Induced Multiple Alignment) with the BCM Search Launcher (http://www.hgsc.bcm.tmc.edu/SearchLauncher/). Alignments were further adjusted by eye in PAUP version 3.1 or 4.0, which was also used to perform heuristic parsimony searches with the TBP dataset. PUZZLE 3.1 was used to calculate maximum likelihood distances (Strimmer and von Haeseler 1996). Distances were corrected with the JTT substitution matrix estimating amino acid usage from the data, and site-to-site rate variation was accounted for with a gamma distribution of 8 variable categories plus invariant sites. Distance trees were constructed with neighbor-joining and Fitch-Margoliash algorithms using BioNJ and FITCH, respectively (Gascuel 1997; PHYLIP package, Felsenstein 1993). In trees constructed with FITCH, the input sequences were jumbled 10 times and trees were searched with global rearrangements. Statistical support values for these trees were determined with 100 bootstrap replicates created by SEQBOOT (PHYLIP). Bootstrap distances were calculated based on the Dayhoff PAM 250 substitution matrix with PROTDIST, and trees constructed with NEIGHBOR and FITCH (the latter with global rearrangement and input sequence order jumbled five times) (PHYLIP). Consensus trees were generated with CONSENSE (PHYLIP). Exhaustive protein maximum likelihood searches using partially constrained trees were conducted with ProtML 2.3 from the MOLPHY package (Adachi and Hasegawa 1996). Resampling estimated log-likelihood bootstraps (RELL) were collated with mol2con (from Dr. A. Stoltzfus), and relative likelihood support values (RLS) were calculated with TreeCons using a Class V (exponential) weighting scheme, and alpha parameters of 0.05 to 0.001 (Jermiin et al. 1997).

Statistical support for tree topologies was also determined by Kishino-Hasegawa tests. Alternative topologies were manually constructed in MacClade, and tested at a significance level of 0.05 using PUZZLE 4.0 with six gamma categories to take into account site-to-site rate variation.

CHAPTER I

Trichomonas vaginalis Possesses Spliceosomal Protein PRP8

This chapter includes work published in Fast, N.M. and W.F. Doolittle. 1999. *Trichomonas vaginalis* possesses a gene encoding essential spliceosomal component, PRP8. *Mol Biochem Parasitol* **99**:275-278.

New sequence has been deposited in Genbank under accession no. AF115849.

INTRODUCTION

The removal of introns from pre-messenger RNA is mediated by the spliceosome: a large macromolecular complex composed of five snRNAs and over one hundred proteins. The RNA component of the spliceosome has been fairly well characterized, as the interactions of the snRNAs with each other and with the pre-mRNA substrate have been elucidated in yeast and mammalian systems (see Nilsen 1998). In contrast, much less is understood about the nature and roles of the substantial proteinaceous component of the spliceosome.

Much of the body of information regarding the proteins within the spliceosome has been generated from the study of temperature sensitive mutants in the yeast *Saccharomyces cerevisiae*. These mutations, originally called *rna* mutations (to reflect the accumulation of unspliced RNA), and later renamed *prp* mutations (for pre-mRNA processing), include approximately twenty-five different mutations, a handful of which have had the wild-type gene cloned and protein product characterized biochemically. The majority of *prp* mutations reside in genes encoding proteins that possess DEAD or DEAH motifs - putative

RNA helicases. That RNA helicases are found in spliceosomes is not terribly surprising due to the multitude of RNA-RNA interactions that take place during pre-mRNA splicing, as these proteins likely help mediate the dynamic interactions between snRNAs and also between snRNAs and the pre-mRNA. However, not all *prp* mutants involve helicases.

One splicing-defect mutation that does not appear to implicate a putative helicase is *prp8*. Exhibiting a phenotype that includes the accumulation of RNA products, but no DNA synthesis impairment, this mutation was found to be lethal in a temperature sensitivity screen (Hartwell, McLaughlin and Warner 1970; Jackson, Lossky and Beggs 1988). A decade ago, the yeast gene responsible for the defect was cloned, and was found to include an extremely large ORF coding for a protein (PRP8) with a predicted molecular mass of 280 kDa (Jackson, Lossky and Beggs 1988). The predicted protein sequence lacked any of the putative helicase motifs; in fact the sequence surprisingly contained no recognizable protein motifs of any sort. Antibodies were raised against four different regions of the protein, and immunoprecipitation revealed that PRP8 is a component of the U5 snRNP (Lossky et al. 1987).

Shortly after this discovery, the same anti-yeast PRP8 antibodies were used in immunoprecipitation experiments with HeLa cell nuclear extracts, and a cross-reacting >200 kDa protein was identified as a component of the human U5 snRNP, and predicted to be a PRP8 homolog (Anderson et al. 1989). Since then, cross-reactive proteins of 200-300 kDa have also been identified in other mammals, several plant species, *Drosophila melanogaster*, and *Caenorhabditis elegans* (Paterson et al. 1991; Kulesza et al. 1993; Hodges et al. 1995). The large conserved size, coupled with the epitope conservation evident from the cross-reacting antibodies, suggests that the PRP8 protein homologs exhibit a high degree of structural conservation. The remarkable nature of the conservation of

PRP8 truly became apparent when sequence data began to accumulate. Currently, full-length PRP8 sequences are known from human, *S. cerevisiae*, *Schizosaccharomyces pombe*, *C. elegans*, *Trypanosoma brucei*, and *Arabidopsis thaliana*, and partial sequences are known from several other plants, along with *Plasmodium falciparum*. Across all taxa, the sequence and size of PRP8 is extremely highly conserved, and no known protein structural motifs can be recognized (Hodges et al. 1995).

The presence of this essential splicing factor in a broad range of taxa raises the question of what potential role it might play in the splicing reaction. Like all snRNPs, the U5 snRNP is composed of several core Sm proteins (common to all snRNPs, and named for their recognition by autoantibodies as part of a human disease) and also specific U5 snRNP proteins, like PRP8. In general, the U5 snRNP is thought to "aid" U5 snRNA, and together bring the two exons in close proximity for the transesterification reactions of splicing (Newman 1997). A more refined proposal is that the U5 snRNP tethers the free exon 1 (released after the first transesterification reaction of splicing) and keeps it in place within the spliceosome for the second step of the reaction to proceed (Newman 1997). Although the U5 snRNA secondary structure is well conserved, the sequence is not, with the exception of U5 loop I (Hinz, Moore and Bindereif 1996). This lack of sequence conservation of U5 snRNA, and more importantly the lack of conservation in exonic sequences, indicates that the interaction of U5 snRNA with the splice site is most likely mediated by protein factors. One such mediator could be the largest protein within the U5 snRNP, PRP8 (Beggs, Teigelkamp and Newman 1995; Newman 1997).

The immense size of PRP8 led to early speculation that it could act as a scaffold within the spliceosome (Whittaker, Lossky and Beggs 1990), and genetic studies in yeast along with biochemical studies in yeast and HeLa cells offer further clues as to the function of PRP8 within the spliceosome. The extreme conservation of the PRP8 sequence could reflect common recognition of conserved regions of the pre-mRNA substrate, and one method to assess such potential interactions is by conducting photo-crosslinking assays. Efficient crosslinks indicate that two components are separated by the length of a covalent bond, and predict that the two components are interacting with one another. In both yeast and human systems, crosslinking experiments have implicated PRP8 in abundant interactions with the pre-mRNA substrate including the 5' splice site, the 3' splice site, the branch region, and the polypyrimidine tract (Garcia-Blanco et al. 1990; Whittaker and Beggs 1991; Teigelkamp, Newman and Beggs 1995; Wyatt, Sontheimer and Steitz 1992; MacMillan et al. 1994; Reyes et al. 1996; Umen and Guthrie 1995; Reyes et al. 1999). PRP8 is associated with substrate RNA throughout both steps of splicing, and has distinct interactions at each step as shown by yeast *in vitro* analyses in the presence and absence of the splicing factor PRP2 (which allows the first transesterification reaction to proceed) (Teigelkamp, Whittaker and Beggs 1995). In addition to interactions with conserved regions of the intron, PRP8 interactions with exonic sequences are also indicated by crosslinking studies (Teigelkamp, Newman and Beggs 1995).

Further elucidation of the interaction of PRP8 with substrate RNA came from mutagenesis of the yeast gene. Isolation of second step mutants defined regions of the PRP8 protein responsible for 3' splice-site fidelity and polypyrimidine tract recognition (Umen and Guthrie 1996). An *in vitro* HeLa cell system further identified a particular set of five contiguous amino acids near the region responsible for 3' splice site fidelity of PRP8 that interacts specifically with the 5' splice site (Reyes et al. 1996; Reyes et al. 1999). When these three putative functional regions of PRP8 are examined in an alignment of homologs, they are amongst the most highly conserved areas of the protein, providing further evidence of their functional significance (see results below).

Taking the essential nature of the yeast PRP8 mutation together with the abundance of interactions that the protein makes with critical regions of the pre-mRNA substrate, it is clear that PRP8 plays a crucial role in the spliceosome. The sheer number of interactions involving PRP8 suggests that the protein could be at the centre of the spliceosome (Reyes et al 1996), a position in line with the proposal that PRP8 is responsible for the U5 snRNP activity of bringing the exons into place for the reactions of splicing (Newman 1997).

Further evidence of the critical nature of PRP8 comes from *in vivo* and *in vitro* experimental results of the effect of depleting PRP8 in yeast spliceosomes (Brown and Beggs 1992). In both cases there is a lack of spliceosome assembly, with the *in vitro* results indicating that the triple U4/U6.U5 snRNP does not form, leaving spliceosome assembly stalled after the addition of U2 snRNP. *In vitro* depletion of PRP8 also resulted in a decrease in the levels of U4, U5 and U6 snRNAs, implying that PRP8-lacking U5 snRNPs are more accessible to nuclease attack.

In addition to interacting with the RNA substrate, PRP8 is known to interact with a number of other protein splicing factors that are helicases, and may play a role in stabilizing RNA-RNA interactions. One example is the splicing helicase PRP28, where a mutant PRP8 allele can rescue a mutant PRP28, linking the activities of the helicase and PRP8 (Strauss and Guthrie 1991). Very recent evidence indicates that PRP8 is involved in (and possibly "governs") the stabilization of the U4-U6 snRNA pairing that is later unwound by PRP28 so that the first step of splicing can proceed (Kuhn, Li and Brow 1999).

As I have presented above, there is a large amount of evidence supporting a critical role for PRP8 within the spliceosome. The extreme conservation and known widespread distribution of this protein also lends credence to it being a necessary spliceosomal component, which taken together indicate that PRP8 is likely a necessary component of spliceosomes in every organism known to splice. Based on this premise, I chose to use PRP8 as a "marker" for splicing, and examined an organism where no introns had been found. Since it is a nearly impossible task to search for introns directly, looking for the machinery that would necessarily be present to remove introns may be the best method to assay their presence.

The parabasalia are believed to be among the earliest known eukaryotes, and it has been proposed that they diverged from the eukaryotic lineage prior to the acquisition of spliceosomal introns (Cavalier-Smith 1991; Cavalier-Smith 1998). Information from sequenced protein-coding genes does not dispute this claim, as not one of the over one hundred parabasalid genes sequenced to date possesses an intron. Here I report that the parabasalid *Trichomonas vaginalis* possesses a gene encoding the essential spliceosomal component PRP8. The size and sequence of the inferred trichomonad protein are highly conserved, particularly in regions predicted to be functional based on studies in yeast and HeLa cells. The presence of PRP8 in *T. vaginalis* leads me to predict that, although none has yet been found, introns are present in parabasalia, albeit at a low density.

RESULTS

Amplification of PRP8 sequence from *Naegleria* as a "positive control"

At the outset of this project, there were very few PRP8 sequences available in the database and full-length sequences were only known from *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. I aligned these sequences, along with all

```
N.an EEVTSLIRSLPVEEQPKQIIVTRKGLLDPLEVHLLDFPNIVIKGSELNLPFSALLKIDKFNDLVIKA
C.el EEVAALIRSLPVEEQPRQIIVTRKAMLDPLEVHLLDFPNIVIKGSELMLPFQAIMKVEKFGDLILKA
S.ce EEVSALVRSLPKEEQPKQIIVTRKAMLDPLEVHMLDFPNIAIRPTELRLPFSAAMSIDKLSDVVMKA

N.an NEPKMLVFNLFDDWRETTNSYTAFLRLILILRAFNVNAEKTKMILKPNKTTIT-QPHHVWPTLDDQT
C.el TEPQMVLFNLYDDWLKTISSYTAFSRVVLIMRGMHINPDKTKVILKPDKTTIT-EPHHIWPTLSDDD
S.ce TEPQMVLFNIYDDWLDRISSYTAFSRLTLLLRALKTNEESAKMILLSD-PTITIKSYHLWPSFTDEQ

N.an WMKVELALKDLILSDYGKKNNVKVTSLTQSEVRDIILGMEIAPPSIQQQQIAEI-----EKEAKEAT
C.el WIKVELALKDMILADYGKKNNVNVASLTQSEVRDIILGMEISAPSQQRQQIADI-----EKQTKEQS
S.ce WITIESQMRDLILTEYGRKYNVNISALTQTEIKDIILGQNIKAPSVKRQKMAELEAARSEKQNDEEA

N.an QLTTI--TTKKTNAYGEEIITVTSTPYEQETFSSKTDWRMRAISAPNLQIRTKRIFIEQEEISENAY
C.el QVTAT--TTRTVNKHGDEIITATTSNYETASFASRTEWRVRAISSTNLHLRTQHIYVNSDDVKDTGY
S.ce AGASTVMKTKTINAQGEEIVVVASADYESQTFSSKNEWRKSAIANTLLYLRLKNIYVSADDFVEEQN

N.an DYVLPMNILKKFIAISDLRTQVMGYVYGKSPEDNPKVKEIHCIVLVPQLGTHKTISVP-----KQLP
C.el TYILPKNILKKFITISDLRTQIAGFMYGVSPPDNPQVKEIRCIVLVPQTGSHQQVNLP-----TQLP
S.ce VYVLPKNLLKKFIEISDVKIQVAAFIYGMSAKDHPKVKEIKTVVLVPQLGHVGSVQISNIPDIGDLP

N.an DHEILNSLEPLGWIHTQPSSVSHTTLTPQDVTFHSKVMSQHKSWDGEKTVTITCQI--TPGSCNLTA
C.el DHELLRDFEPLGWMHTQPNEL--PQLSPQDVTTHAKLLTDNISWDGEKTVMITCSF--TPGSVSLTA
S.ce DTEGL---ELLGWIHTQTEEL--KFMAASEVATHSKLFAD------KKRDCIDISIFSTPGSVSLSA

N.an HKLTPEGYDWGRKNKDIFS-DEPQGFNHGAFYEGVQLILSESFMGFFMVPEEGSWNYNFM
C.el YKLTPSGYEWGKANTDKGN--NPKGY-MPTHYEKVQMLLSDRFLGYFMVPSNGVWNYNFQ
S.ce YNLTDEGYQWGEENKDIMNV-LSEGF-EPTFSTHAQLLLSDRITGNFIIPSGNVWNYTFM
```

Figure 1-1 *Naegleria andersoni* PRP8 partial coding sequence
The inferred amino acid sequence of the *N. andersoni* PRP8 partial sequence (*N.an*) is shown aligned with the homologous region from *Caenorhabditis elegans* (*C.el*) and *Saccharomyces cerevisiae* (*S.ce.*).

available others (short ESTs from rice, human and *Plasmodium*) and designed a suite of six degenerate PCR primers. As the sequence sampling was so limited, I was concerned that degenerate primers based on such a sparse alignment would have little chance of amplifying the *Trichomonas vaginalis* sequence, which I expected could be much more divergent than those in the alignment (if the sequence were present in *T. vaginalis* at all). To this end, I tested the PCR primers by amplifying PRP8 from a protist expected to have more divergent sequences than the representation at that time, and where introns were known to be present. I chose the heterolobosean *Naegleria andersoni*.

Of the possible primer combinations, only one pair (Spl8-2 and Spl8-3) produced a ~1400 bp PCR product that, once cloned and sequenced, was truly PRP8. Multiple clones were sequenced on both strands, showing that the *Naegleria* inferred coding sequence is highly conserved compared to the *C. elegans* and yeast PRP8 homologs as shown in Figure 1-1.

PRP8 from *T. vaginalis*

More confident in the PCR primers after the successful amplification of part of the PRP8 gene from *N. andersoni*, I used the same suite in PCR reactions with genomic *T. vaginalis* DNA as a template. Almost all primer combinations failed to produce a product of predicted size (including the Spl-2 and Spl-3 combination that had successfully amplified PRP8 sequence from *Naegleria*); however, one pair (Spl8-1 and Spl8-6) produced an amplification product of the predicted size (~400 bp). This short product was cloned and multiple clones were sequenced. Comparing the sequences against those in Genbank (with BLAST) revealed that several clones contained the same sequence and that this sequence was, in fact, highly similar to known PRP8 sequences. The PCR product was amplified from a region of the 3' portion of the gene, the most highly conserved part of known PRP8 sequences.

As the PCR product was only 400 bp in length and full-length PRP8 genes are ~6-7 kb in length, it was necessary to construct a genomic library to determine the entire sequence of the *T. vaginalis* gene. This library was then screened with one of the positive PCR product clones (Tva816-6) which was selected arbitrarily. Screening with this probe only resulted in the retrieval of genomic clones (T1 set) containing ~4 kb of insert genomic DNA, containing the 3′ third of the PRP8 gene. As the library had been constructed with size-selected fragments of 6-10 kb, this insert was disconcertingly small, implying that the number of pfu screened should be increased for subsequent screening attempts. Repeating the screen with the same Tva816-6 probe and doubling the number of pfu screened resulted in the retrieval of another set of clones (T1a1 set) that only differed only from the T1 set by having an additional ~1 kb of 5′ sequence, still leaving the 5′ half of the *T. vaginalis* PRP8 gene unknown. One final screen with the same probe over another increased number of pfu did not retrieve clones different than those in the T1 and T1a1 sets.

In an attempt to retrieve clones containing more 5′ sequence an anchored PCR approach was used. The PCR was anchored with an exact match primer to the most 5′ sequence known from the genomic clones and a degenerate primer was designed to match a conserved block of sequence 700 bp farther upstream in sequences not present in the T1a1 set of clones. This amplification reaction produced a band of the expected size that subsequent cloning and sequencing indicated was indeed PRP8 sequence overlapping with known clones. I arbitrarily chose clone TvagexR-B1 as a probe, and screened the genomic library again. Screening results indicated that most clones were very small, but fortunately one genomic clone (TvagexR10) was ~10 kb in length, and sequencing the ends indicated that TvagexR10 contained the remainder of the *T. vaginalis* PRP8 gene. The large genomic clone was further subcloned into five *Hin*dIII digested fragments in order to expedite sequencing of the full-length gene. In total, finishing the almost 7 kb of full-length, double-stranded, polished sequence of the *T. vaginalis* PRP8 gene required ~40 sequencing reactions (including both ABI walking and LiCor reactions).

Evidence indicating the parabasalid origin of PRP8

Sequencing the ends of the large TvagexR10 genomic clone revealed the presence of an alpha-tubulin gene adjacent to PRP8. The alpha-tubulin gene was truncated on the genomic clone with only approximately the 3′ third of the gene present. As there is a wide sampling of alpha-tubulin gene sequences available (including representative parabasalid sequences other than *T. vaginalis*), I sought to determine the full-length sequence of the adjacent alpha-tubulin gene and test its phylogenetic position.

Employing PCR, I attempted to retrieve the 5′ end of the gene, while taking into consideration the multi-copy nature of alpha-tubulin genes, and making every effort to amplify the same gene as that adjacent to PRP8 in the *T. vaginalis* genome. To this end, I designed exact-match PCR primers (Tvatubby-1 and -2) to downstream non-coding sequence and used these primers in conjunction with a standard alpha-tubulin degenerate PCR primer (AtubA, Keeling and Doolittle 1996) that is complementary to the 5′ end of the gene. Both primer combinations produced amplification products of the predicted size, and once cloned and sequenced proved to be the alpha-tubulin gene in question, overlapping perfectly with the genomic clone sequence. Two clones were sequenced completely on both strands. Sequence identity alone (evident from comparing the sequence with those in the database using BLAST) indicates that

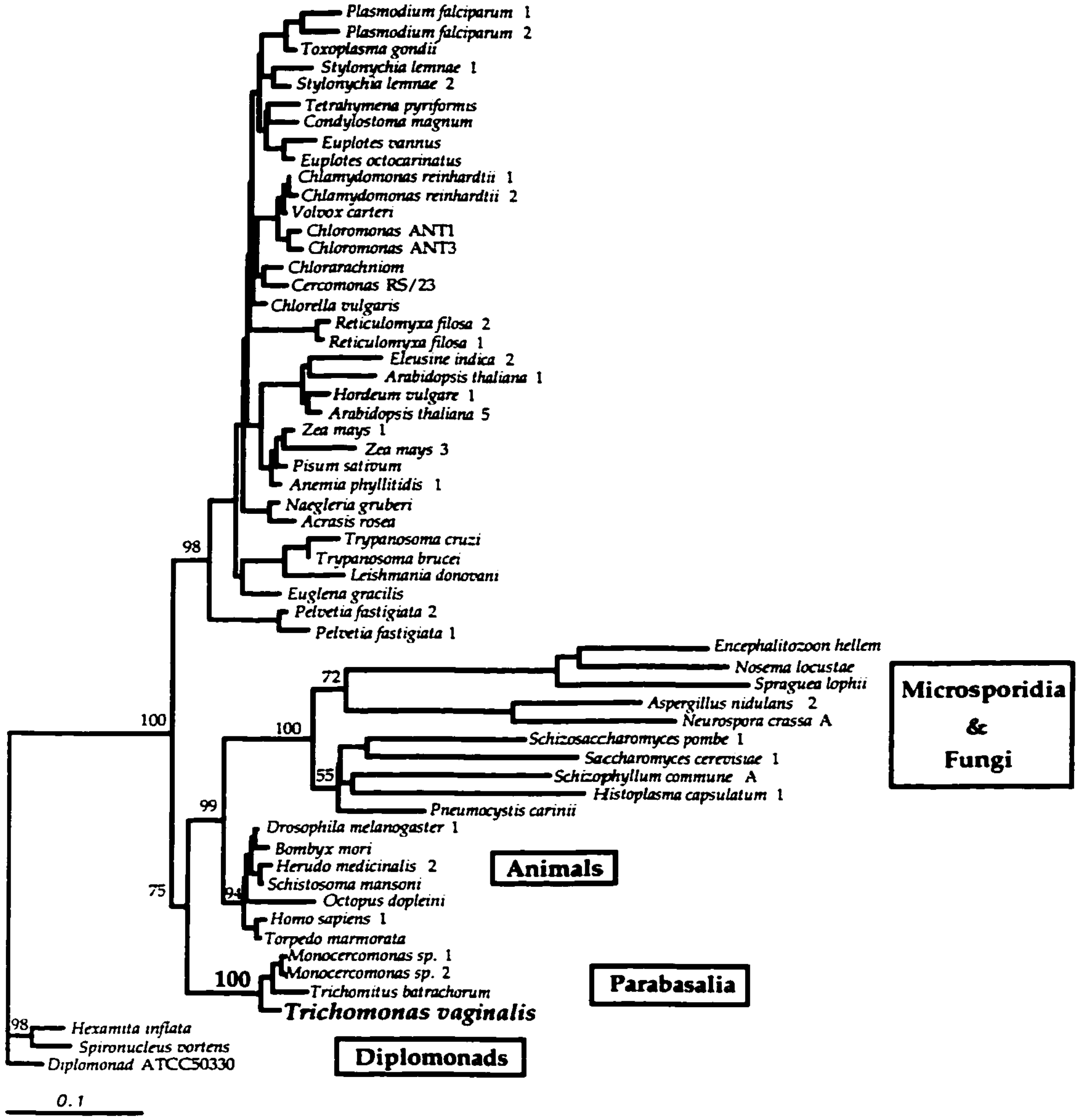


Figure 1-2 Phylogenetic tree of alpha-tubulin sequences
The neighbor-joining tree is shown with neighbor-joining bootstrap values. The putative trichomonad alpha-tubulin sequence next to PRP8 branches with 100% support with the other parabasalid sequences, indicating that the alpha-tubulin gene (and the adjacent PRP8 gene) are indeed of *Trichomonas vaginalis* origin.

the alpha-tubulin sequence is most similar to those from parabasalia (data not shown).

To test the phylogenetic position of the alpha-tubulin gene adjacent to PRP8, the sequence was aligned with 58 other eukaryotic alpha-tubulin sequences and phylogenetic analyses were carried out. Distance methods were used with both neighbor-joining and Fitch-Margoliash algorithms, outgrouping with diplomonad sequences. The neighbor-joining tree is shown in Figure 1-2, and the Fitch-Margoliash tree possesses the same overall topology of parabasalid, animal, fungi-microsporidia, and "large heterogeneous" (a.k.a. "everything-else") clades that is typical of alpha-tubulin phylogeny (for example, see Keeling and Doolittle 1996; Keeling, Deane and McFadden 1998). Differences between the two trees lie only in the weakly supported relationships within these clades. The specific focus here is the position of the putative alpha-tubulin sequence of *T. vaginalis*. It branches with 100% bootstrap support in both distance analyses with the other parabasalid sequences, providing strong, indirect evidence that both the alpha-tubulin gene and the adjacent PRP8 gene are both of parabasalid origin.

Direct evidence for the PRP8 gene originating from *T. vaginalis* came from a genomic Southern blot (Figure 1-3). Restriction enzymes were chosen based on the complete sequence of the gene. The probe detected bands of the sizes predicted from the sequence of the genomic clone, indicating that the clones are authentic representations of their genomic loci in *T. vaginalis* and that the gene is present in single copy in the *T. vaginalis* genome.

The *T. vaginalis* PRP8 gene is highly conserved

The *T. vaginalis* PRP8 open reading frame is 6963 bp in length, comparable in size to other known PRP8 homologs. This gene is almost certainly expressed,

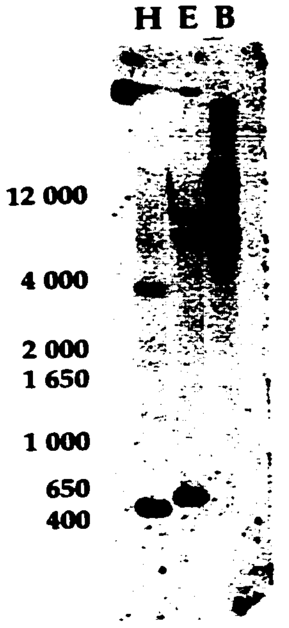


Figure 1-3 Confirming the provenance of the *T. vaginalis* PRP8 gene
A genomic Southern blot probed with a PCR product probe (which was also used to screen the genomic library). Genomic DNA was digested with *Hin*dIII, *Eco*RI and *Bam*HI (lanes labelled as H, E, and B respectively in the figure). The enzymes were chosen based upon the genomic sequence: *Hin*dIII was expected to cut once within the probe sequence and relieve 560 bp and 3724 bp products; *Eco*R1 sites outside of the probe predicted the presence of a 660 bp fragment. These results (along with the fact that multiple independent clones from the library always contained the same sequence) indicate that the *T. vaginalis* gene exists as a single copy.

based on the presence of an upstream sequence that resembles an initiator (Inr) element. Characterized protein-coding genes from *T. vaginalis* have so far lacked upstream consensus TATA sequences, instead possessing conserved Inr sequence elements flanking the transcription start sites (Liston and Johnson 1998). Directly upstream of the *T. vaginalis* PRP8 coding region is the sequence TCAAAATTTTAGATG, where the start codon is underlined. The length and position of this sequence relative to the start codon, and the presence of the highly conserved TCA within the sequence element, lend credence to this being the Inr element, responsible for expression of PRP8. Further indicative evidence for expression is the very length of the PRP8 coding region, since it is extremely unlikely that almost 7 kb of in-frame contiguous coding sequence would be maintained if the gene were not expressed.

The inferred protein sequence of the *T. vaginalis* PRP8 homolog is 2320 amino acids in length, with a predicted molecular mass of 271 kDa. Since the beginning of this work, the number of PRP8 sequences available for comparison has increased. In addition to complete gene sequences from *S. cerevisiae* and *C. elegans*, full-length sequences from *Schizosaccharomyces pombe*, human, *Arabidopsis thaliana* and *Trypanosoma brucei* are now available. Several partial coding sequences or ESTs from genome sequencing projects have been added to the alignment, as have unpublished nucleomorph sequences from *Chlorarachnion* CCMP 621 and *Guillardia theta*, kindly provided by Dr. G. McFadden and Dr. S. Douglas respectively. The larger dataset allows for a more meaningful comparison of the *T. vaginalis* PRP8 homolog to known functional protein sequences. The *T. vaginalis* PRP8 sequence, aligned with all full-length sequences, along with the partial nucleomorph sequences, is shown in Figure 1-4.

Over its entire length, the *T. vaginalis* PRP8 homolog exhibits a high degree of sequence identity, 56% with both the human and *C. elegans* sequences.

Figure 1-4 PRP8 sequence alignment
The *Trichomonas vaginalis* PRP8 conceptually translated amino acid sequence is shown aligned with all other available full-length PRP8 protein sequences in addition to partial nucleomorph sequences from *Guillardia theta* and *Chlorarachnion* species CCMP 621. (The N-terminal region of the *T. vaginalis* sequence is shown unaligned, as this regions varies greatly in both length and sequence among homologs.) Specific yeast mutations are noted with asterisks (*) and labelled. The five amino acid block involved in recognizing the 5' splice site in human cells is also indicated with asterisks and labelled as 5' SS. Abbreviations: *T.va, Trichomonas vaginalis; C.el, Caenorhabditis elegans; H.sa, Homo sapiens; A.th, Arabidopsis thaliana; S.po, Schizosaccharomyces pombe; S.ce, Saccharomyces cerevisiae; T.br, Trypanosoma brucei; C.sp, Chlorachnion* species CCMP 621 nucleomorph; *G.th, Guillardia theta* nucleomorph.

```
T.va    METGDTGVLQERIAQWKHLRKKHFKLEKIKTTSKGPSAKELPPGHIRQIMTSHGNMSHDKFAGQKRLYIGALKYAPHAVLKF  [82]
C.el                             SEKKKFGMSDTQKEEMPPEHVRKVIRDHGDMTSRKYRHDKRVYLGALKYMPHAVLKL  [92]
H.sa                             AEKRKFGFVDAQKEDMPPEHVRKIIRDHGDMTNRKFRHDKRVYLGALKYMPHAVLKL  [102]
A.th                             GDKRKFGFVETQKEDMPPEHVRKIIR-------RKHRLDKRVYLGALKFVPHAVFKL  [117]
S.po                             GVKRKQGYVQTEKADLPPEHLRKIMKDRGDMSSRKFRADKRSYLGALKYLPHAVLKL  [123]
S.ce                             TKKAKRSNLYTPKAEMPPEHLRKIINTHSDMASKMYNTDKKAFLGALKYLPHAILKL  [175]
T.br                             GYRATYQEAVAQKDEVPPEYLRKLVKDNGDLSGKRFNAERKLCVALLRYMPLALYKL  [103]

T.va    LENMPMPWEELRKVRVLYHTAGALTFVNEVPRVVPPQYLAQWAETWIAMRREKRD-RHQIRRLR--FPPFDDEEQPLDFVTNIEGVEPPE  [169]
C.el    LENMPMPWEQIRDVKVLYHITGAITFVNDIPRVIEPVYMAQWGTMWIMMRREKRD-RRHFKRMR--FPPFDDEEPPLDYADNILDVEPLE  [179]
H.sa    LENMPMPWEQIRDVPVLYHITGAISFVNEIPWVIEPVYISQWGSMWIMMRREKRD-RRHFKRMR--FPPFDDEEPPLDYADNILDVEPLE  [189]
A.th    LENMPMPWEQVIIHLVLYHITGAITFVNEVRWVVEPIYMAQWGSMWIMMRREKRD-RRHFKRMR--FPPFDDEEPPLDYADNLLDVDPLE  [204]
S.po    LENMPMPWEEYREVKVLYHVTGAITFVNESPRVIEPHFIAQWGTMWMMMRREKRD-RKNFKRLR--FPPFDDEEPPFSI-DQLLDLEPLE  [209]
S.ce    LENMPHPWEQAKEVKVLYHTSGAITFVNETPRVIEPVYTAQWSATWIAMRREKRD-RTHFKRMR--FPPFDDDEPPLSYEQHIENIEPLD  [262]
T.br    LENMPMPWEEARYVNVVYHMRGVLTLVEDTPTAAEPLYLAQWGSIWTKMRSHKVELQQECGTFRRVISKGNENEPPIDFSDYIMDREPPP  [193]

T.va    AILMELDEDEDAAVYEWFYDYLGLPSQY--ING-LSYRKWKLPLTVMSTLYRLARPLVDQYEDPNSKYLIDLPSLFTSKALNEVIPGGPR  [256]
C.el    PIQMELDPEEDGAVAEWFYDHKPLATTR-FVNG-PTYRKWAFSIPQMSTLYRLANQLLTDLVDDNYFYLFDMKSFFTAKALNVAIPGGPK  [267]
H.sa    AIQLELDPEEDAPVLDWFYDHQPLRDSRKYVNG-STYQRWQFTLPMMSTLYRLANQLLTDLVDDNYFYLFDLKAFFTSKALNMAIPGGPK  [278]
A.th    AIQLELDEEEDSAVYSWFYDHKPLVKTK-MING-PSYQTWNLSLPIMSTLHRLAAQLLSDLVDRNYFYLFDMPSFFTAKALNMCIPGGPK  [292]
S.po    AIRMDLDEEDDAPVMDWFYENKALEDTPH-VNG-PTYRRWKLNLPQMANLHRLGYQLLSDLRDDNYFYLFNDNSFFTAKALNVAIPGGPK  [297]
S.ce    PINLPLDSQDDEYVKDWLYDSRPLEEDSKKVNG-TSYKKWSFDLPEMSNLYRLSTPLRDEVTDKNYYYLFDKKSFFNGKALNNAIPGGPK  [351]
T.br    ALYDDLDEEDAAAVLDWFYDPFPRLVHPNQIRGSRRPNGYYFTIDVIETLFRNAIPILPNLDDRNYYYLWDLKSFYAAKAMHIAIPRAPK  [283]

T.va    FEPLFTDV-DPNQEWTEFNDINKII-------IRTPISTEWKIAYPNLYNNRPRK--ISIAPYHYPLSCFA-KYNTIITPVFQLAPNLSS  [335]
C.el    FEPLVKDLH-TDEDWNEFNDINKVI-------IRAPIRTEYRIAFPFMYNNLISSLPVQVSWYHTPSVVFI-KTEDPDLPAFYYDPLINP  [348]
H.sa    FEPLVRDINLQDEDWNEFNDINKII-------IRQPIRTEYKIAFPYLYNNLPHH--VHLTWYHTPNVVFI-KTEDPDLPAFYFDPLINP  [358]
A.th    FEPLHRDMEKGDEDWNEFNDINKLI-------IRSPLRTEYKVAFPHLYNNRPRK--VKLCVYHTPMVMYI-KTEDPDLPAFYYDPLIHP  [372]
S.po    FEPLYKDEAPEMEDWNEFNDIYKLI-------IRHPIKTEYRIAFPYLYNSRARS--VALSEYHQPSNVFV-PPEDPDLPAFFWDPIINP  [377]
S.ce    FEPLYPRE--EEEDYNEFNSIDRVI-------FRVPIRSEYKVAFPHLYNSRPRS--VRIPWYNNPVSCIIQNDEEYDTPALFFDPSLNP  [430]
T.br    FEAPSTIQ-EEEGEWTEFNDLRRVIHRDDPRKPRFTMLTERQIAFPFLYSDVVD--GVTVAPYRYPAQIRV-ENEDPAVPCFSWNPSLNP  [369]
```

Figure 1-4 PRP8 sequence alignment

```
T.va   I-------SRPRANNNPEQAS---------EEDLQPFTV-------NCDPLFNDLDHESEEKQFKIFDTLSLFWAPSPFNCRTGNTQRAQ   [402]
C.el   IV-----LSNLKATEENLPEG----------EEEDEWELP-----EDVRPIF----EDVPLYTDNTANGLALLWAPRPFNLRSGRTRRAV   [414]
H.sa   I-------SHRHSVKSQEPLP----------DDDEEFELP-----EFVEPFL----KDTPLYTDNTANGIALLWAPRPFNLRSGRTRRAL   [422]
A.th   IS-----NSNNTNKEQRKSNGY--------DDDGDDFVLP-----EGLEPLL----NNSPLYTDTTAPGISLLFAPRPFNMRSGRTRRAE   [440]
S.po   I------TSRQLTLHELDTSPE-----DSAIEEDPNFEIP-------FDPFF----HSEDIEFEHTASALILLWAPHPFNKRSGATKRAQ   [445]
S.ce   I-------PHFIDNNSSLNVS--------NTKENGDFTLP-----EDFAPLLA---EEEELILPNTKDAMSLYHSPFPFNRTKGKMVRAQ   [497]
T.br   IKAIQKRHSDPVGSSSVALCSAALRKSQWLGDEEPEDGCQPMSLMENFSPLF----QELPLENVDTKSAMLLAFAPGPFNEFEGGMKRRV   [455]
C.sp                                                  FKFLF----DFLKLYNRNTSKALNLSYAPLPFSNQHPEIKRIL    [39]

T.va   DVPLIRPWYTQ--RAPWKQP-VKVRVSYQKLLKNYVLNR---------SHHRKQYVVRRRKTITKIFKTT--PYFQSTTLDWVEAGLQVV   [478]
C.el   DVPLVKSWYRE--HCPAGMP-VKVRVSYQKLLKVFVLN----------ALKHRPPKPQKRRYLFRSFKAT--KFFQTTTLDWVEAGLQVL   [489]
H.sa   DIPLVKNWYRE--HCPAGQP-VKVRVSYQKLLKYYVLN----------ALKHRPPKAQKKRYLFRSFKAT--KFFQSTKLDWVEVGLQVC   [497]
A.th   DIPLVAEWFKE--HCPPAYP-VKVRVSYQKLLKCYLLN----------ELHHRPPKAQKKKHLFRSLAAT--KFFQSTELDWVEVGLQVC   [515]
S.po   DVPLIKHWYLE--HCPPNQP-VKVRVSYQKLLKSHVMN----------KLHMAHPKSHTNRSLLRQLKNT--KFFQSTSIDWVEAGLQVC   [520]
S.ce   DVALAKKWFLQ--HPDEEYP-VKVKVSYQKLLKNYVLN----------ELHPTLPTNHNKTKLLKSLKNT--KYFQQTTIDWVEAGLQLC   [572]
T.br   DIPVAEHWCRDPPSLLTNDTRDKILRSYTQLLKHHVAKNLRRDRQKERPKEEGGNQDEGGQPVRRLDELANLDFFHKTKIDWLEAGLQVM   [545]
C.sp   DITITRSWYLV--HCPLEFP-KKVRVSYQKLLKKYIMN----------KITKKSIQKNSSSSLLTLLKKN--EFFQSTKLDWVEAGIHLC   [129]

T.va   QQGYNMCNLLIKRHRLVFLHLDYNFNLKPIKTLNTKERRKSRFGNAYHLMREFFRFTKLLLDCHIQYRLGQIDAYVLADALQYVFSHAGH   [568]
C.el   RQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREILRLTKLVVDAHVQYRLNNVDAYQLADGLQYIFAHVGQ   [579]
H.sa   RQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREVLRLTKLVVDSHVQYRLGNVDAFQLADGLQYIFAHVGQ   [587]
A.th   RQGYNMLNLLIHRKNLNYLHLDYNFNLKPVKTLTTKERKKSRFGNAFHLCREILRLTKLVVDANVQFRLGNVDAFQLADGLQYIFSHVGQ   [605]
S.po   RQGYNMLQLLIHRKGLTYLHLDYNCNLKPTKTLTTKERKKSRFGNAFHLMREILRLTKLIVDSHVQYRLGNIDAYQLADGLHYIFNHVGQ   [610]
S.ce   RQGHNMLNLLIHRKGLTYLHLDYNFNLKPTKTLTTKERKKSRLGNSFHLMRELLKMMKLIVDTHVQFRLGNVDAFQLADGIHYILNHIGQ   [662]
T.br   RQGHNMLVQLINVKSLPYVHINYNFEAKPTRTLTTKEIKKSRLGPAFHLIRELLGFMKQLIDMHTMYRLGKNDSIQLADAIQYLFSHLGR   [635]
C.sp   KQGYNMLNLLIHKKGLNFLHLDFNFNLKPIKTLTTKERKKSRFGNAFHLTREILRLTKLIIDANIQFRLGNVDAYQLADCLFYIFSHVGH   [219]
```

Figure 1-4 PRP8 sequence alignment

```
T.va   LTGMYRYKYKLMHQVRTCKDLKHVLYSRFNTGEVGKGPGVGFWGPMWRVWVFFMRGSIPLMERWLGSRVAREYEGRFSKRLP--STVTKQ    [656]
C.el   LTGMYRYKYKLMRQVRMCKDLKHLIYYRFNTGPVGKGPGCGFWAPGWRVWLFFLRGITPLLERWLGNLLSRQFEGRHSKGVA--KTVTKQ    [667]
H.sa   LTGMYRYKYKLMRQIRMCKDLKHLIYYRFNTGPVGKGPGCGFWAAGWRVWLFFMRGITPLLERWLGNLLARQFEGRHSKGVA--KTVTKQ    [675]
A.th   LTGMYRYKYRLMRQIRMCKDLKHLIYYRFNTGPVGKGPGCGFWAPMWRVWLFFLRGIVPLLERWLGNLLARQFEGRHSKGVA--KTVTKQ    [693]
S.po   LTGMYRYKYRLMRQIRACKDFKHLIYYRFNTGPVGKGPGCGFWAPSWRVWLFFLRGIVPLLERWLGNLLARQFEGRHSTGVA--KQITKQ    [698]
S.ce   LTGIYRYKYKVMHQIRACKDLKHIIYYKFNK-NLGKGPGCGFWQPAWRVWLNFLRGTIPLLERYIGNLITRQFEGRSNEIV---KTTTKQ    [748]
T.br   LTGVYRYKLRAMRQIKRSRDLKHVLYSKFNVGEVLRGPGCGFWAPSWRVWVFFLRGMTPLLQRYLGNLTDRVLRGREAKGKHDGKRITRQ    [725]
C.sp   LTGIYRYKYRVMRQIRICKDFKHLIHYRFNTGEIGKGPGVGFWFPLWRVWVFFLRGIIPLLEKWLFNLLSRQFFGRTENKIA--KNVTKQ    [309]

T.va   RVESNYDIELRASVLHDITDTMPEGIRNAKAHTVLAHMSEAWRCWKANIPWKVPGLPPPLEAIILRYVKAKADWWTKNAHYARERIARDG    [746]
C.el   RVESHFDLELRAAVMHDILDMMPDGIKQNKARVILQHLSEAWRCWKANIPWKVPGLPTPVENMILRYVKAKADWWTNSAHYNRERVRRGA    [757]
H.sa   RVESHFDLELRAAVMHDILDMMPEGIKQNKARTILQHLSEAWRCWKANIPWKVPGLPTPIENMILRYVKAKADWWTNTAHYNRERIRRGA    [765]
A.th   RVESHFDLELRAAVMHDVVDAMPEGIKQNKARTILQHLSEAWRCWKANIPWKVPGLPVAIENMILRYVKSKADWWTNVAHYNRERIRRGA    [783]
S.po   RVDSHQDLELRAAVMNDILDMIPEGIRQGKSKTILQHLSEAWRCWKANIPWKVPGLPAPIENMILRYVKSKADWWTSVAHFNRERIRRGA    [788]
S.ce   RLDAYYDLELRNSVMDDILEMMPESIRQKKARTILQHLSEAWRCWKANIPWDVPGMPAPIKKIIERYIKSKADAWVSAAHYNRERIKRGA    [838]
T.br   RVETDKDVNIKEAFRRELREMLPPDVRTEVIRTMDQHMNEAFRHWRAGLRWSVPGLAKPLTDLVNKYVKLRAEEYVRVTQYQRKRINEGD    [815]
G.th                                                             IQKVLLKQYIKQKAKWWIMDTFIVRKKIIKNK     [32]
C.sp   RVESHYDLELRASVIHEILSLLPDSTKLNKANLILSHLSEAWRCWKANISWIVPNMPPKIEKIILKYIKQKADWWTNIAHYNRERIKKGA    [399]

T.va   TVDKAITRKNTGRLTRLYLKQQSDYQANYLK--EGPYITPEQGVAMLTTMQNWLEMRQFTPIPFPPMQYKHDTKMLILALENMRFGH       [831]
C.el   TVDKTVCKKNLGRLTRLYLKSEQERQHNYLK--DGPYISAEEAVAIYTTTVHWLESRRFSPIPFPPLSYKHDTKLLILALERLKESY       [842]
H.sa   TVDKTVCKKNLGRLTRLYLKAEQERQHNYLK--DGPYITAEEAVAVYTTTVHWLESRRFSPIPFPPLSYKHDTKLLILALERLKEAY       [850]
A.th   TVDKTVCRKNLGRLTRLWLKAEQERQHNFQK--DGPYVTADEGIAIYSTTVNWLESRKFSAIPFPPLSYKHDTKLLILALERLKESY       [868]
S.po   TVDKTVAKKNLGRLTRLWLKAEQERQHNYLK--DGPYVTADEAVAIYTTTFVHWLESRRFQPIPFPPLSYKHDTKLLVLALERLKEAY      [873]
S.ce   HVEKTMVKKNLGRLTRLWIKNEQERQRQIQK--NGPEITPEEATTIFSVMVEWLESRSFSPIPFPPLTYKNDTKILVLALEDLKDVY       [923]
T.br   TVDKQAFMKNLGRLTRLKLMDEQNRQRSYMEGTDTDIITPEQATEIYRMMANWLSDRGFKKISFPKASRPAELRLLELALNRLRDQH       [902]
G.th   IIDKKIFRKNLSRLSRLWMKSEIDRQKKSFS-HESSFLSDR-----FDFFNIWLNIIDFKIINFPKFYSRLENK---LALITISGI-       [109]
C.sp   TVDKVVCKKNLGRLTRLYLKSEQEKQHKYLK--EGPYIKVREVITIYRILIHWLDINNFNLISFPSNEYKFSSK                    [476]
```

Figure 1-4 PRP8 sequence alignment

```
T.va   DVSMRMNQTLREELGLIENAHDNPHEALIRIKRDFMTARAFREVKFTFLEHYTRVIPNYEIYALEKMTDAYLDQYLWYEA-DRRHLFPPW   [920]
C.el   SVKNRLNQSQREELALIEQAYDNPHEALSRIKRHMLTQRAFKEVGIEFMDLYTHLIPVYDIEPLEKVTDAYLDQYLWYEA-DKRRLFPAW   [931]
H.sa   SVKSRLNQSQREELGLIEQAYDNPHEALSRIKRHLLTQRAFKEVGIEFMDLYSHLVPVYDVEPLEKITDAYLDQYLWYEA-DKRRLFPPW   [939]
A.th   SAAVKLNQQQREELGLIEQAYDNPHEALMRIKRHLLTQHSFKEVGIEFMDLYSHLIPVYQIDPLEKITDAYLDQYLWYEG-DKRHLFPNW   [957]
S.po   SVKGRLNQSQREELALVEQAYDNPHEMLSQIKRRLLTMRTFKEVGIEFMDMYSHLIPVYSVDPMEKICDAYLDQYLWFEA-DRRHLFPSW   [962]
S.ce   ASKVRLNASEREELALIEEAYDNPHDTLNRIKKYLLTQRVFKPVDITMMENYQNISPVYSVDPLEKITDAYLDQYLWYEA-DQRKLFPNW  [1012]
T.br   NIANRLTQAQREEQARIEEAFNSPHETLSKIVDCLARVRRFKNVEVEYMDTFSSLYPIYNVVPSEKLVDSFLDQYLWYEAMDQQRLFPNW   [992]
G.th   NNFVEYSENYRLKEFKIYNLIN----FFDTIKSILLTQKNFIQVSISFSENFFFLRPIYHFDLYEKLTCLFLDKYFWYLG-SKQFFFPKF   [194]

T.va   VQPSDLIPPPVLVHKWCERINSLVDAWNTEDGQTMVLVETSL-EKFYEQIDLTFLNYMLRLVVDHNLADYMTSKNNVKISFKDMSYLNGV  [1009]
C.el   VKPGDTEPPPLLTYKWCQGLNNLQDVWETSEGECNVIMETKL-EKIAEKMDLTLLNRLLRLIVDHNIADYMTSKNNVLINYKDMNHTNSF  [1020]
H.sa   IKPADTEPPPLLVYKWCQGINNLQDVWETSEGECNVMLESRF-EKMYEKIDLTLLNRLLRLIVDHNIADYMTAKNNVVINYKDMNHTNSY  [1028]
A.th   IKPADSEPPPLLVYKWCQGINNLQGIWDTSDGQCVVMLQTKF-EKLFEKIDLTVLNSLLRLVLDPKLANYVTGKNNVVLSYKDMSYTNTY  [1046]
S.po   VKPSDSEPPPLLVYKWCQGINNLTDVWETSNGECNVLMETRL-SKVFEKVDLTLLNRLMSLLMDTNLASYASAKNNVVLSYKDMSHTNSY  [1051]
S.ce   IKPSDSEIPPLLVYKWTQGINNLSEIWDVSRGQSAVLLETTL-GEMAEKIDFTLLNRLLRLIVDPNIADYITAKNNVVINFKDMSHVNKY  [1101]
T.br   VKPSDVEPVPILVYKWCQGINDSPGIWDFDRDESVVLLHAKLEDDFYGNIDWNLFRPLLELIMDKSLAEYIVSRHDVVVEFKDMAYHCRK  [1082]
G.th   VKPCDTQLCQLSIFNYFKYIEKIIDEKKYANI-LNCFIKIKI-NNYLENLRLINHKSQLSKVFDDNLLVFIFSRNNCFINYKDMNFFNAF   [282]

T.va   GLIHGLQFTSFIAQYMGLLVDLLILGLRRANEMCGPPSMPNSLFQF-ASIEDEIRHPIRMYQRYATRIHILYKFNAEQARDLIRDYCDVN  [1098]
C.el   GIIRGLQFASFIVQFYGLVLDLLVLGLRRASEIAGPPQCPNEFLQF-QDVATEIGHPIRLYCRYIDRVWIMFRFSADEARDLIQRYLTEH  [1109]
H.sa   GIIRGLQFASFIVQYYGLVMDLLVLGLHRASEMAGPPQMPNDFLSF-QDIATEAAHPIRLFCRYIDRIHIFFRFTADEARDLIQRYLTEH  [1117]
A.th   GLIRGLQFASFVVQFYGLVLDLLLLGLTRASEIAGPPQRPNEFMTY-WDTKVETRHPIRLYSRYIDKVHIMFKFTHEEARDLIQRHLTER  [1135]
S.po   GLVRGLQFSSFIWQFYGLVLDLLILGLQRATEIAGPADAPNDFLHF-KDQATETSHPIRLYTRYIDKVYIMFRFTDEESRDLIQRFLNEN  [1140]
S.ce   GLIRGLKFASFIFQYYGLVIDLLLLGQERATDLAGPANNPNEFMQF-KSKEVEKAHPIRLYTRYLDRIYMLFHFEEDEGEELTDEYLAEN  [1190]
T.br   GMLRGFMFSSFLAQYWGLVIDVLLLGTQRSQEIAGPARRPNPFMSWMRDPLLATSHPIRGYCRYKNEVYVLLKYTKVEADDVRHRYLEET  [1172]
G.th   GIIKGFALNNFIIQLFYFIVDISSLGIKNIIYVISSKKTENSFFEI------KSKEKIIIYMRYIEQIYI---FQIKEGKN---------   [354]
```

Figure 1-4 PRP8 sequence alignment

```
T.va    SN----NNDEMLGYN--NKTCWPKDARMRLIKHDVNLGRAVFWDLQNRLPRSLCEVNWNSSENSASGFHQSFASVYSKDNPNLLFYMCGF  [1182]
C.el    PDP---NNENIVGYN--NKKCWPRDARMRLMKHDVNLGRAVFWDIKNRLPRSITTVEWE----------NSFVSVYSKDNPNMLFDMSGF  [1184]
H.sa    PDP---NNENIVGYN--NKKCWPRDARMRLMKHDVNLGRAVFWDIKNRLPRSVTTVQWE----------NSFVSVYSKDNPNLLFNMCGF  [1192]
A.th    PDP---NNENMVGYN--NKKCWPRDARMRLMKHDVNLGRSVFWDMKNRLPRSITTLEWE----------NGFVSVYSKDNPNLLFSMCGF  [1210]
S.po    PDP---TNSNVVNYSKGKKNCWPRDARMRLMKHDVNLGRAVFWEIRNRLPRSLTTLEWE----------DTFPSVYSKDNPNLLFSMTGF  [1217]
S.ce    PDP---NFENSIGYN--NRKCWPKDSRMRLIRQDVNLGRAVFWEIQSRVPTSLTSIKWE----------NAFVSVYSKNNPNLLFSMCGF  [1265]
T.br    KNDPQKRAENASVYGFKNFKQWPRDARMRLFLNDVNLARAVIWEFRGRLPPGIADINES----------NALASVYSKDNPNLLFDMGGF  [1252]
G.th    ---------TKIDYDM--YDHLKKNFRAKLEKSILK----FNLNFKLEISIICRHL------------------VDDRND----LTLIGF   [407]

T.va    EVRILPKIRLEREDF-TPQEGTWVLQNEITHETTAFVFLRVSEKSITYFRNRVRTILLSSQATTFMKVSNKWNTAIIALVVYFREALVAT  [1271]
C.el    ECRILPKCRTANEEF-VHRDGVWNLQNEVTKERTAQCFLKVDEESLSKFHNRIRQILMSSGSTTFTKIVNKWNTALIGLMTYFREAVVNT  [1273]
H.sa    ECRILPKCRTSYEEF-THKDGVWNLQNEVTKERTAQCFLRVDDESMQRFHNRVRQILMASGSTTFTKIVNKWNTALIGLMTYFREAVVNT  [1281]
A.th    EVRVLPKIRMGQEAFSSTRDGVWNLQNEQTKERTAVAFLRADDEHMKVFENRVRQILMSSGSTTFTKIVNKWNTALIGLMTYFREATVHT  [1300]
S.po    EVRILPKIRQ-NEEF-SLKDGVWNLTDNRTKQRTAQAFIRVTEDGINQFGNRIRQILMSSGSTTFTKIANKWNTALIALMTYYREAAIST  [1305]
S.ce    EVRILPRQRM-EEVV-SNDEGVWDLVDERTKQRTAKAYLKVSEEEIKKFDSRIRGILMASGSTTFTKVAAKWNTSLISLFTYFREAIVAT  [1353]
T.br    SVRILPVVRTEDEVL--ENESTWNLQNTTTRDVTARAFLQVSPDDVNNIRNKARRAIMMVGSSTFQSIAAKWNALVTEIVPYYREAILGT  [1340]
G.th    SFQNLLYS---YES---------TLTKKVKKNFIFNNF-KICNISIKLFQLKVKNFLLTSGSSSFSKLIKKWNGLLLGYFCFFRKALVSS   [484]

T.va    PELIDEIVKCENRVQTRVKLGLNSKMPNRFPPVVFYAPKEFGGLGLISMGHVLIPQSDLRYATQTM-AETTHFRDGMDHPEENFIP-ALY  [1359]
C.el    QELLDLLVKCENKIQTRIKIGLNSKMPSRFPPVVFYTPKEIGGLGMLSMGHVLIPQSDLRWMQQTEAGGVTHFRSGMSHDEDQLIP-NLY  [1362]
H.sa    QELLDLLVKCENKIQTRIKIGLNSKMPSRFPPVVFYTPKELGGLGMLSMGHVLIPQSDLRWSKQTD-VGITHFRSGMSHEEDQLIP-NLY  [1369]
A.th    QELLDLLVKCENKIQTRVKIGLNSKMPSRFPPVIFYTPKEIGGLGMLSMGHILIPQSDLRYSNQTD-VGVSHFRSGMSHEEDQLIP-NLY  [1388]
S.po    PELLDLLVKCESKIQTRVKISLNSKMPSRFPPAVFYSPKELGGLGMLSMGHVLIPQSDLRWSKQTD-TGITHFRSGMTTNGEHLIP-NLY  [1393]
S.ce    EPLLDILVKGETRIQNRVKLGLNSKMPTRFPPAVFYTPKELGGLGMISASHILIPASDLSWSKQTD-TGITHFRAGMTHEDEKLIP-TIF  [1441]
T.br    DSLQQVLARAEHRMQSRIMMALNSRAKARFPPVIFYAPTDLGGLGMLSVGHSLIPARDLVYSKSTS-TGVQFFYSGLTNADNIPIP-NIL  [1428]
G.th    QNFTTRLKKYEKEIIANIKASLSSKMPSRFPPVLFFSPKEFGGLGMLSLFNYYIPENDL----------KSDLKMVISKSNSLNYTNSLT   [564]
                                                     *
                                                 prp8-124
```

Figure 1-4 PRP8 sequence alignment

```
T.va    RYVQSWEGEIEDSKRVWQHYTNMRKEAGALNKKITIEDLDSLWDRGIPRINVLFQRDRHTLAYDKGWRTRLYFKKYSLFKTNPYAWTHHH  [1449]
C.el    RYIQPWEAEFVDSVRVWAEYALKRQEANAQNRRLTLEDLDDSWDRGIPRINTLFQKDRHTLAYDKGWRVRTEFKAYQILKQNPFWWTHQR  [1452]
H.sa    RYIQPWESEFIDSQRVWAEYALKRQEAIAQNRRLTLEDLEDSWDRGIPRINTLFQKDRHTLAYDKGWRVRTDFKQYQVLKQNPFWWTHQR  [1459]
A.th    RYIQPWESEFIDSQRVWAEYALKRQEAQAQNRRLTLEDLEDSWDRGIPRINTLFQKDRHTLAYDKGWRVRTDFKQYQALKQNPFWWTHQR  [1478]
S.po    RYIQPWESEFIDSQRVWAEYAMKRQEALQQNRRLTLEDLEDSWDRGIPRINTLFQKDRHTLAYDKGWRVRTEFKQYQLLKNNPFWWTSQR  [1483]
S.ce    RYITTWENEFLDSQRVWAEYATKRQEAIQQNRRLAFEELEGSWDRGIPRISTLFQRDRHTLAYDRGHRIRREFKQYSLERNSPFWWTNSH  [1531]
T.br    QYYTPWETEVRESVKAWTEFNMRDREAKAAGTRLSIDDIEHIINKGVPRIRVLFSRHAKLFQFDKGFRCRMEFQRYLAGKYLKNWWFHQE  [1518]
G.th    KFIKDWTNEFKKSNIAWKKLLILKKNFKKRRIKIYYQKISNLFGKGIPRIETIFSKYRLFLPYDYGWRLNLDLSRYILSTNNSFWWTSPK   [654]

T.va    HDGKLW--NLKDYRADVIQALGGVEGILSHSIFKATGYKHWEGLFWDNTTGFEEALKYRKLTNAQRQGMSQVPNRRYTLWWSPTINRANV  [1537]
C.el    HDGKLW--NLNNYRTDMIQALGGVEGILEHTLFRGTYFPTWEGLFWERASGFEESMKFKKLTNAQRSGLNQIPNRRFTLWWSPTINRANV  [1540]
H.sa    HDGKLW--NLNNYRTDMIQALGGVEGILEHTLFKGTYFPTWEGLFWEKASGFEESMKWKKLTNAQRSGLNQIPNRRFTLWWSPTINRANV  [1547]
A.th    HDGKLW--NLNNYRTDVIQALGGVEGILEHTLFKGTYFPTWEGLFWEKASGFEESMKYKKLTNAQRSGLNQIPNRRFTLWWSPTINRANV  [1566]
S.po    HDGKLW--QLNNYRVDVIQALGGVEGILEHTMFKATGFPSWEGLFWEKASGFEESMKFKKLTNAQRSGLNQIPNRRFTLWWSPTINRANV  [1571]
S.ce    HDGKLW--NLNAYRTDVIQALGGIETILEHTLFKGTGFNSWEGLFWEKASGFEDSMQFKKLTHAQRTGLSQIPNRRFTLWWSPTINRANV  [1619]
T.br    HDGNICGGVLERYRVDTNIALGGVEAILEHSLFRGTGFPSWEGIEFNRAGGFENSKKDSKLAKQQRAGLANVPNRRFALWWCPTINRSDV  [1608]
G.th    HEGKLY--NLSSYNNQMIFRLGGIKNILEHTLFKATFYSNWEGLFWEKRSKFENLIRNKKLTNAQKLGLNQIPNRRFTLWWSPTINRNNV   [742]
                                           *         **                                *
                                      prp8-125  prp8-122&-123                         prp8-121

T.va    YVGFQVQLDLTGIFMHGKIPSLKVSLIQLFRGHMWQKTHESIVMDVMQVLDSHLSQLQIDYITKETIHPRKSYKMNSSCADLIMISQNKW  [1627]
C.el    YVGFQVQLDLTGIFMHGKIPTLKISLIQIFRAHLWQKIHESVVMDLCQVFDQELDALEIQTVQKETIHPRKSYKMNSSCADVLLFAQYKW  [1630]
H.sa    YVGFQVQLDLTGIFMHGKIPTLKISLIQIFRAHLWQKIHESIVMDLCQVFDQELDALEIETVQKETIHPRKSYKMNSSCADILLFASYKW  [1637]
A.th    YVGFQVQLDLTGIYMHGKIPTLKISLIQIFRAHLWQKIHESVVMDLCQVLDQELEPLEIETVQKETIHPRKSYKMNSSCADVLLFAAHKW  [1656]
S.po    YVGFQVQLDLTGIMMHGKIPTLKISLIQIFRSHLWQKIHESVVWDLCQVLDQELESLQIETVQKETIHPRKSYKMNSSCADILLLAAYKW  [1661]
S.ce    YVGFLVQLDLTGIFLHGKIPTLKISLIQIFRAHLWQKIHESIVFDICQILDGELDVLQIESVTKETVHPRKSYKMNSSAADITMESVHEW  [1709]
T.br    QAGFETKIDTTGVFMCGKLETIKKSLIKIFSGSLWEKCHGAVVNDIASKLKDMMVELDAASVTLQQQHPQKSYTYTSSAPDIVMASTSRW  [1698]
G.th    YIGYQTQIDLTGIFMHGKIPTLKISIIQIFRSHLWQKIHESLVILLLKRLDIEKQNLNIGILEKKVNHPKKSYKFESSSADLILYPKINF   [832]
```

Figure 1-4 PRP8 sequence alignment

Figure 1-4 PRP8 sequence alignment

```
T.va    -NVSRPSLMADSK---DVMDNTTTQKYWLDVQLRWGDYDSHDVERYARAKFLDYTTDNMSIYPSPTGVLIAIDLAYNLYSAYGNWFPGMK  [1716]
C.el    -NVSRPSLLADSK---DVMDSTTTQKYWIDIQLRWGDYDSHDIERYARAKFLDYTTDNMSIYPSPTGVLIAIDLAYNLHSAYGNWFPGSK  [1723]
H.sa    -PMSKPSLIAESK---DVFDQKASNKYWIDVQLRWGDYDSHDIERYTKAKFMDYTTDNMSIYPSPTGVIIGLDLAYNLHSAFGNWFPGSK  [1742]
A.th    -NVSRPSLLNDNR---DVLDNTTTNKYWIDVQLRFGDYDSHDIERYTRAKFLDYSTDAQSMYPSPTGVLIGIDLCYNMHSAYGNWIPGMK  [1747]
S.po    -EVSKPSLLHETN---DSFKGLITNKMWFDVQLRYGDYDSHDISRYVRAKFLDYTTDNVSMYPSPTGVMIGIDLAYNMYDAYGNWFNGLK  [1795]
S.ce    -PSTEPCFVNETK---TFHGEFLTTRFWVDIQLRWGDYDMHDIERYTRSLFYAYTSGTQSMYPSSTGIIIGVDLCYNEWTAFGTWIPGLQ  [1713]
T.br    PVTSKPTVLSDETG--DEYRAHTTSKYWIDVQLRWGNYDSHNIAEYTRSRFYEYSS--AKMYPFPAGIVVAIDLAYNCHSAFGYWVPRLK  [1784]
G.th    -LITYPILLGIKKILTHGDFLKVSQVYWIDVQLRWGDFDSHDIERYVRMKYYEYNDVKKKLFPSGHGILIAYDLCYNVYSSYGNWILGLS   [921]


T.va    ELIDKAMKHVLQFNPSISVLRERVKKSLQLYTSEIPEPALNSTNFGELFGNKI-TWIVEDKHVVYRVKIQKTFEGNYTTSPVNGGVFIMNP  [1802]
C.el    PLIRQAMAKIIKANPAFYVLRERIRKGLQLYSSEPTEPYLTSQNYGELFSNQI-IWFVDDTNVYRVTIHKTFEGNLTTKPINGAIFIFNP  [1805]
H.sa    PLIQQAMAKIMKANPALYVLRERIRKGLQLYSSEPTEPYLSSQNYGELFSNQI-IWFVDDTNVYRVTIHKTFEGNLTTKPINGAIFIFNP  [1812]
A.th    PLLAQAMNKIMKSNPALYVLRERIRKGLQLYSSEPTEPYLSSQNYGEIFSNQI-IWFVDDTNVYRVTIHKTFEGNLTTKPINGVIFIFNP  [1831]
S.po    PLIQQSMNKIMKANPALYVLRERIRKGLQLYASEPQEQYLSSSNYAELFSNQI-QLFVDDTNVYRVTIHKTFEGNLTTKPINGAIFIFNP  [1836]
S.ce    PLIQNSMRTIMKANPALYVLRERIRKGLQIYQSSVQEPFLNSSNYAELFNNDI-KLFVDDTNVYRVTVHKTFEGNVATKAINGCIFTLNP  [1884]
T.br    PLMMKLMTAIIRHNIALNTLRERMKRDLQLFSSAPTEAGLSVTNIAELFSEGMRTWIVDDSATYVTSEQPTAEGGRKFRSENGAVLIFEP  [1874]
G.th    NFIKNELFSFHKNSAILNILRSRIRKSLQIYQKNNIESNESILNIDDFFKKKC--LIVDDSCLSNHLELQNLQKNKVINYHSGFLFIFNP  [1009]
                                              *
                                        prp8-103/107


T.va    ATGQLFLKIITTKAWQQQKRLQQLAKWKAAEETCALVRTLPPEEQPKQVICTSELLLLDPVQSYL-SEFPNTVVKGSDMDLPLPAFMKIPK  [1891]
C.el    RTGQLFLKIIHTSVWAGQKRLSQLAKWKTAEEVAALIRSLPVEEQPRQIIVTRKAMLDPLEVHL-LDFPNIVIKGSELMLPFQAIMKVEK  [1894]
H.sa    RTGQLFLKIIHTSVWAGQKRLGQLAKWKTAEEVAALIRSLPVEEQPKQIIVTRKGMLDPLEVHL-LDFPNIVIKGSELQLPFQACLKVEK  [1901]
A.th    RTGQLFLKIIHTSVWAGQKRLGQLAKWKTAEEVAALVRSLPVEEQPKQVIVTRKGMLDPLEVHL-LDFPNIVIKGSELQLPFQACLKIEK  [1920]
S.po    RTGQLFLKVIHTSVWAGQKRLGQLAKWKTAEEVAALIRSLPVEEQPRQIIVTRKGMLDPLEVHL-LDFPNITIKGSELQLPFQAIIKLDK  [1925]
S.ce    KTGHLFLKIIHTSVWAGQKRLSQLAKWKTAEEVSALVRSLPKEEQPKQIIVTRKAMLDPLEVHM-LDFPNIAIRPTELRLPFSAAMSIDK  [1973]
T.br    ȦTGNLKLSIVHKSVFAGQKRRTKLAREKAAEEIASWLRSVPASQRPGKLIVTRSRFRQTLHNMLILDYPNIIIGQSDLNLAVPMVLṘHSR  [1964]
G.th    INGLIYIKNLNCSKFNNKKNFKDFSKIFLGNELVNFLKTCPKFELPLNIITLKKGLKEFLHLKL-IEYPDINVYQAEIDIMFRNLLKLNI  [1098]
                                                                                    *     *****
                                                                           prp8-101/102   5' SS
```

```
T.va    IAEEVIHAPEPKMVLYNLYDDWLDTVSPHAAFRRLMLILRALLMERMKAWDILRPSANVVTQQNHLWPTHSADEWAEVEIRLKDLVIDIY   [1981]
C.el    FGDLILKATEPQMVLFNLYDDWLKTISSYTAFSRVVLIMRGMHINPDKTKVILKPDKTTITEPHHIWPTLSDDDWIKVELALKDMILADY   [1984]
H.sa    FGDLILKATEPQMVLFNLYDDWLKTISSYTAFSRLILILRALHVNNDRAKVILKPDKTTITEPHHIWPTLTDEEWIKVEVQLKDLILADY   [1991]
A.th    FGDLILKATEPQMALFNIYDDWLMTVSSYTAFQRLILILRALHVNNEKAKMLLKPDMSVVTEPNHIWPSLTDDQWMKVEVALRDLILSDY   [2010]
S.po    INDLILRATEPQMVLFNLYDDWLQSVSSYTAFSRLILILRALNVNTEKTKLILRPDKSIITKENHVWPNLDDQQWLDVEPKLRDLILADY   [2015]
S.ce    LSDVVMKATEPQMVLFNIYDDWLDRISSYTAFSRLTLLLRALKTNEESAKMILLSDPTITIKSYHLWPSFTDEQWITIESQMRDLILTEY   [2063]
T.br    LADLRISATESKGWEFCLYDDWLRQFQPATCFNLLNLILRGYHVNLSRTRQTLEPDLHVEVHHSHFWPTYTREEWEAVSVRLQEMIIADA   [2054]
G.th    FKDKIHSKNKDDFFIIELYDNWLDSISPITAFTRLILILKSIELDNKK---VINAVQNKELKNEKLWNSLSNLEWINTEIFLKNLILKHN   [1185]

T.va    CQRNNVSANSLTQSEIRDIILGVKIAAPSEERQQMAK------EVEEEDKAALKT--VTTAT-RDADGNQHIIQTFSQYEQQQFKSKSDW   [2062]
C.el    GKKNNVNVASLTQSEVRDIILGMEISAPSQQRQQIADI-----EKQTKEQSQVTA--TTTRT-VNKHGDEIITATTSNYETASFASRTEW   [2066]
H.sa    GKKNNVNVASLTQSEIRDIILGMEISAPSQQRQQIAEI-----EKQTKEQSQLTA--TQTRT-VNKHGDEIITSTTSNYETQTFSSKTEW   [2073]
A.th    AKKNKVNTSALTQSEIRDIILGAEITPPSQQRQQIAEI-----EKQAKEASQLTA--VTTRT-TNVHGDELISTTISPYEQSAFGSKTDW   [2092]
S.po    AKKNNINVASLTNSEVRDIILGMTITAPSLQRQQIAEI-----EKQGRENAQVTA--VTTKT-TNVHGDEMVVTTTSAYENEKFSSKTEW   [2097]
S.ce    GRKYNVNISALTQTEIKDIILGQNIKAPSVKRQKMAELEAARSEKQNDEEAAGASTVMKTKT-INAQGEEIVVVASADYESQTFSSKNEW   [2152]
T.br    ARRMNVSPNQFTEMEKKDILLGKKMTTVEIQEEEM-------KELEEMKRTKLV--QEHTIDVVTKSGETAKKRVKAAFDFGNSTSASNW   [2135]
G.th    FENGFYNLSSIDESTIKCLILGSKL----PENDSL---------KNFKRKSNFKI--LKSKD-SNSKSIQTYVKNKKFYNNIIEKN----   [1255]

T.va    RSRALTSRGLVMRANTLMIPPPVVKPK--LELIIPENIYRRF--VEISDPYMQICGFLFGVKMNDTLQVISI---VIPPQNGDRDEIDF-   [2144]
C.el    RVRAISSTNLHLRTQHIYVNSDDVKDT-GYTYILPKNILKKF--ITISDLRTQIAGFMYGVSPPDNPQVKEIRCIVLVPQTGSHQQVNLP   [2153]
H.sa    RVRAISAANLHLRTNHIYVSSDDIKET-GYTYILPKNVLKKF--ICISDLRAQIAGYLYGVSPPDNPQVKEIRCIVMVPQWGTHQTVHLP   [2160]
A.th    RVRAISATNLYLRVNHIYVNSDDIKET-GYTYIMPKNILKKF--ICIADLRTQIAGYLYGISPPDNPQVKEIRCVVMVPQCGNHQQVQLP   [2179]
S.po    RNRAISSISLPLRTKNIYVNSDNISETFPYTYILPQNLLRKF--VTISDLRTQVAGYMYGKSPSDNPQIKEIRCIALVPQLGSIRNVQLP   [2185]
S.ce    RKSAIANTLLYLRLKNIYVSADDFVEEQ-NVYVLPKNLLKKF--IEISDVKIQVAAFIYGMSAKDHPKVKEIKTVVLVPQLGHVGSVQIS   [2239]
T.br    RARSLANATVFGEGTTVEIDHSGVTGS-SDQLIFPQELLKIL--FPCFDVQAQFCAYLFGQTLPDSPNVKEVLCIMVPPQKSSAVEYT--   [2220]
G.th    ----LSKRIILNEFKNDNVNQAFINIG-TRFFIIPYNFIKIYTKFLIKQNEIEYYCYILGKFCKHNTKCKLF--IILKFQLGKTAT----   [1334]
```

Figure 1-4 PRP8 sequence alignment

```
T.va    -KQI---LPNH---DFLDGASPIGFIHTRVGENSSLEPRDAKVLASLCKKNPKIDSDNFANVVISFPVGGCIMAASTLSREGFEWAETNI  [2227]
C.el    -----TQLPDH---ELLRDFEPLGWMHTQPNELPQLSPQDVTTHAKLLTDNISWDGEKTVMITCSFTPGSVSLTAYKLTPSGYEWGKANT  [2235]
H.sa    -----GQLPQH---EYLKEMEPLGWIHTQPNESPQLSPQDVTTHAKIMADNPSWDGEKTIIITCSFTPGSCTLTAYKLTPSGYEWGRQNT  [2242]
A.th    -----SSLPEH---QFLDDLEPLGWIHTQPNELPQLSPQDVTFHTRVLENNKQWDAEKCIILTCSFTPGSCSLTSYKLTQAGYEWGRLNK  [2261]
S.po    -----SKLPHDLQPSILEDLEPLGWIHTQSSELPYLSSVDVTTHAKILSSHPEWDT-KAVTLTVSYIPGSISLAAYTVSKEGIEWGSKNM  [2269]
S.ce    NIPDIGDLPDT------EGLELLGWIHTQTEELKFMAASEVATHSKLFADKKR----DCIDISIFSTPGSVSLSAYNLTDEGYQWGEENK  [2319]
T.br    ---TPSCIPHDHPILTENHLSLLGVLRCSGGE-PSIHSRDVAIHGRLLACNEGLQTEGLTTVVVGVSQDGIGIRCYTTTREGISWALEEY  [2306]
G.th    ----NPLIYNN-MNLFIKRFSFLGFFTEKTLFESNMKNILNNNQGIFCIFKKKYISWVTITKNEKFIIEGRFLYISKNKIESITTSI---  [1416]


T.va    ---GMDNPKDFDDNFAKVLGISITNEINGWMMAPENGIWNYSFNSLRLQSVPDNYPISVQNPKTFFDMYHRVQHFTSFKREMN-GEELSI  [2313]
C.el    D--KGNNPKGYMPTHYEKVQMLLSDRFLGYFMVPSNGVWNYNFQG-QRWSPAMKFDVCLSNPKEYYHEDHRPVHFHNFKAFDDPLGTGSA  [2322]
H.sa    D--KGNNPKGYLPSHYERVQMLLSDRFLGFFMVPAQSSWNYNFMG-VRHDPNMKYELQLANPKEFYHEVHRPSHFLNFALLQEG-EVYSA  [2328]
A.th    D--TGSNPHGYLPTHYEKVQMLLSDRFFGFYMVPENGPWNYNFMG-ANHTVSINYSLTLGTPKEYYHQVHRPTHFLQFSKMEE---DGDL  [2345]
S.po    D-INSDEAIGYEPSMAEKCQLLLSDRIQGFFLVPEEGVWNYNFNG-ASFSPKMTYSLKLDVPLPFFALEHRPTHVISYTELET-NDRLEE  [2356]
S.ce    D-IMNVLSEGFEPTFSTHAQLLLSDRITGNFIIPSGNVWNYTFMG-TAFNQEGDYNFKYGIPLEFYNEMHRPVHFLQFSELAG-DEELEA  [2406]
T.br    SHALQREPTEVPPLHVIPARVTLSTELQGFFLVPTDNGWNHTFRG-ATWREDTTFDVRVDTPQFFFFATHRPDHFLNFARLTE--EEATI  [2393]
G.th    ---FFLISKILIGFLISLSDIH                                                                      [1436]


T.va    DVDNNFI      [2320]
C.el    DREDAFA      [2329]
H.sa    DREDLYA      [2335]
A.th    DRDDSFA      [2352]
S.po    DMPDAFA      [2363]
S.ce    EQIDVFS      [2413]
T.br    DMADLENLMA   [2403]
```

Figure 1-4 PRP8 sequence alignment

Compared with *S. cerevisiae* and *T. brucei*, the *T. vaginalis* inferred protein sequence is 50% and 37% identical, respectively. For comparison, the human sequence is 86% identical to *C. elegans* and 63% identical to *S. cerevisiae*. Isolated regions of the protein share even greater identity. Comparing 100 amino acid aligned character blocks of the *T. vaginalis* and human proteins, some regions are >70% identical (Figure 1-5). Amino acid character blocks of fifty were also analyzed for percent identity. The overall distribution was identical to that shown in Figure 1-5, but narrower amino acid regions of PRP8 highly conserved between *T. vaginalis* and human were determined.

Those regions exhibiting over 70% sequence identity can correspond to regions of the PRP8 protein predicted to be functionally significant based on mutational and biochemical analyses. Such is the case for the region of PRP8 involved in 3' splice-site selection as identified by mutational analyses in yeast (Umen and Guthrie 1996). Based on percent identity scores, one of the most highly conserved regions of the protein (>70% identical between human and *T. vaginalis*) includes the positions of the five yeast mutations (*prp8-121* though *prp8-125*) that suppress 3' splice-site selection mutants (mutations are marked on the alignment in Figure 1-4). In contrast, the specific regions surrounding yeast mutations *prp8-103/107* and *prp8-101/102* exhibit less drastic conservation (marked in Figure 1-4), although the amino acid in question for mutants 101 and 102 (a glutamate residue in yeast, human and *C. elegans*) is conservatively substituted with aspartate in *T. vaginalis*. Likewise, nonpolar amino acids are found in all taxa (including *T. vaginalis*) at the amino acid position implicated in mutants *prp8-103/107*. A similar situation exists for the region surrounding and including the five amino acids predicted from crosslinking studies in HeLa nuclear extracts to interact with the 5' splice site (Reyes et al. 1996; Reyes et al. 1999). As is evident from the alignment (Figure 1-4), the five amino acids are only

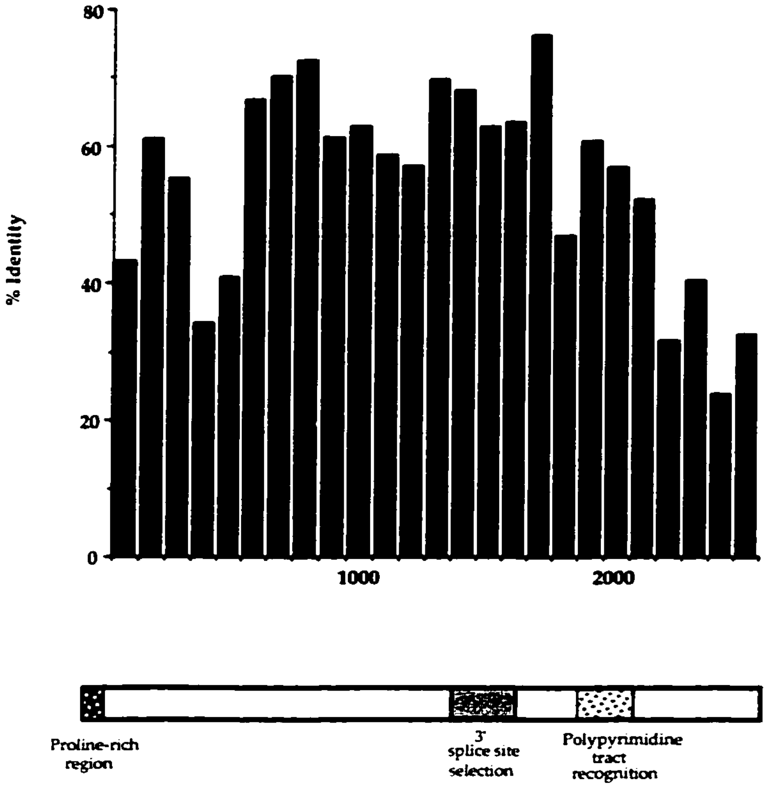


Figure 1-5 Identity of *T. vaginalis* and *H. sapiens* PRP8
Percent identity shared by the *T. vaginalis* and human PRP8 homologs within 100 amino acid consecutive blocks of their alignment. Below the histogram is a diagram of *S. cerevisiae* PRP8 identifying the functional regions proposed by mutational analysis. (Only the yeast PRP8 homologs are known to have a proline-rich N-terminal extension.)

moderately conserved among all taxa, and the *T. vaginalis* PRP8 sequence is not more divergent than any other homolog in this region.

There are other regions of the protein which are also highly conserved. These may correspond to yet-to-be-defined regions of functional significance: potential sites of activity or sites of interactions of PRP8 with other proteins within the spliceosome.

DISCUSSION

It is generally held that the catalytic activity of the spliceosome resides in the snRNAs of its RNA component (see, for example, Nilsen 1998). Much of this belief is based on sequence/secondary-structure similarity of the spliceosomal snRNAs to regions of self-splicing Group II introns (Sharp 1991; Palmer and Logsdon 1991; Coppertino and Hallick 1993). More recently, this theory has gained additional support from the discovery that a mutant Group II intron incapable of splicing can be rescued by the addition of U5 snRNA (Hetzer et al. 1997). Group II and spliceosomal introns also share the same splicing chemistry, further indicating a relationship. The substantial protein component of the spliceosome cannot be ignored, however, and it may be an oversimplification to assume that the proteins in the spliceosome solely act as a scaffold to present the snRNAs. In particular, the extremely large spliceosomal protein PRP8 is known to be essential for splicing, and could be a possible catalytic candidate based on the multitude of interactions it makes with critical regions of the pre-mRNA substrate. Although a catalytic role is questionable, the sheer number of interactions this essential protein makes with the substrate certainly indicates that it is crucial for splicing and could be located at the core of the spliceosome.

The protist *Trichomonas vaginalis* has no known spliceosomal introns in genes so far sequenced, and it has furthermore been proposed that the parabasalid lineage diverged prior to spliceosomal intron acquisition (Cavalier-Smith 1991; Cavalier-Smith 1998). Here I have presented evidence that seriously questions the intron-lacking status of *T. vaginalis* by showing that this protist possesses the essential spliceosomal component PRP8 - the first evidence for spliceosomes in the parabasalia.

A combination of degenerate PCR followed by genomic library screening revealed that *T. vaginalis* possesses an open reading frame that codes for a PRP8 homolog. The genomic library screening also showed that the gene adjacent to PRP8 encodes alpha-tubulin, and phylogenetic analysis of the alpha-tubulin sequence indicated that the adjacent PRP8 gene was indeed of parabasalid origin; the *T. vaginalis* origin of the gene was confirmed by Southern blotting.

Other results indicate that the PRP8 homolog of *T. vaginalis* is very likely expressed. Not only does the upstream region of the gene contain a sequence that can be identified as a putative Inr element, but also the sheer length of the open reading frame (6963 bp) argues against it being a pseudogene. It seems highly improbable that almost 7 kb of contiguous, in-frame coding sequence would be maintained if it were not being expressed.

The inferred translation product of the gene is 2320 amino acids in length and has a predicted molecular mass of 271 kDa. The large size of the predicted *T. vaginalis* gene product is consistent with that of known PRP8 proteins. In addition to size conservation, the coding sequence of the *T. vaginalis* gene shares a high percent identity with known PRP8 proteins indicating that this *T. vaginalis* gene is indeed a true PRP8 homolog. The sequence of the trichomonad PRP8 shares the highest overall percent identity (56%) with human and *C. elegans*. As the protein is so large, the extent of conservation fluctuates over the entire sequence, requiring a closer examination of blocks of aligned amino acid characters. A comparison of such aligned blocks from *T. vaginalis* and human reveals that there are specific regions of the protein that exhibit much higher percent identity; in places the *T. vaginalis* protein is over 70% identical to human PRP8.

Since there are currently no known introns in trichomonads, it is of interest to consider whether the *T. vaginalis* PRP8 protein has the potential to function in a spliceosome. The overall high identity with those homologs known to function in spliceosomes could indicate this, but stronger evidence comes from examining specific regions of the protein known to be critical for function. Although PRP8 has no recognizable protein motifs over its entire length (Hodges et al. 1995), studies in yeast and human (mutational analyses and crosslinking experiments) have identified particular amino acid regions critical for PRP8 function (Umen and Guthrie 1996; Reyes et al. 1999). Such regions define splicing activities implicated in 5' splice-site recognition, 3' splice-site fidelity, and polypyrimidine tract recognition. In an attempt to focus on specific regions of the *T. vaginalis* PRP8 homolog, blocks of amino acids were compared to the human sequence and percent identity within these blocks was determined (see Figure 1-5). The region of the *T. vaginalis* PRP8 protein that includes and surrounds yeast mutations affecting 3' splice-site fidelity is one of the most highly conserved regions of the protein. Not only is the general region highly conserved, but specific amino acid residues identified in the yeast study (*prp8-121, -122, -123, -125*) are either identical in *T. vaginalis* or are conservative substitutions (Figure 1-4), indicating that the *T. vaginalis* homolog may be able to function in 3' splice-site selection. The amino acid block of PRP8 found to be crosslinked to the 5' splice-site in HeLa studies is generally less conserved across all taxa (Figure 1-4) and the *T. vaginalis* homolog is not any more divergent than known functional homologs in this region. In the region of PRP8 identified in yeast studies to be involved in polypyrimidine tract recognition (which overlaps the 5' splice-site recognition block of amino acids), two 100 amino acid character blocks share 60 and 70% identity between *T. vaginalis* and human.

Although no introns have yet been found within the parabasalia, the presence of a highly conserved PRP8 homolog strongly indicates that functioning spliceosomes could be present in *T. vaginalis*, since it is unlikely that this crucial spliceosomal component would be present (and also highly conserved in regions of functional significance) otherwise. Predicting that an active spliceosome exists in *T. vaginalis* gains additional support from preliminary evidence from a collaborative project (with M. Konarska) indicating that *T. vaginalis* also possesses a critical RNA component of the spliceosome, U6 snRNA.

In addition to the defined functional regions of PRP8 that exhibit a high degree of sequence conservation as discussed above, there are regions shared by the *T. vaginalis* and human PRP8 homologs that have not yet been defined in the characterization of the PRP8 protein, but that could correspond to other sites of activity or protein recognition. Candidates for such interactions could include three proteins recently found to be tightly associated with human PRP8, even in the absence of RNA (Achsel et al. 1998). Indeed, regions of PRP8 conserved across the broad phylogenetic distance between, for example, human and trichomonads may be good candidates for future mutational analyses using existing splicing systems (i.e. in yeast and HeLa cells), to further determine the nature of PRP8's crucial role within the spliceosome.

Another protist known to possess a PRP8 homolog is *Trypanosoma brucei* (Lücke et al. 1997). The presence of this protein in trypanosomes, along with its sequence divergence, provide an interesting discussion point. All evidence collected to this point indicates that all trypanosomal messages are *trans*-spliced by the addition of a mini-exon (spliced leader sequence) to the 5' end of the mRNA and so far, of the over 1000 genes sequenced, none have been found to contain any spliceosomal introns, leading to speculation that trypanosomal messages undergo *trans*-splicing only. *Trans*-splicing is not unique to trypanosomes, as nematodes and trematodes also *trans*-splice, but in these organisms *cis*-splicing is also present (Bonen 1993). Evidence from *in vitro* experiments substituting intron elements in *trans*-splicing messages and splicing *trans*-spliced messages in a *cis*-splicing system indicates that *trans*-splicing can operate with *cis*-splicing machinery/components (Bruzik and Steitz 1990; Bruzik and Maniatis 1992; Metzenberg and Agabian 1996). It is therefore interesting to examine the trypanosomal PRP8 homolog with this in mind. The *Trypanosoma* PRP8 sequence is by far the most divergent in overall percent identity compared to other known homologs, including regions significant for function in yeast and human (see Figure 1-4). It is therefore possible that this lack of sequence conservation reflects diminished constraints on these regions of the PRP8 protein in the *trans*-splicing system. The possibility cannot be ruled out however, that a *cis*-intron could be found in trypanosomes, and the divergence in the trypanosomal PRP8 sequence is simply reflecting the divergent nature of trypanosomal genes in general.

If, however, the trypanosomal PRP8 sequence divergence is due to its role in *trans*-splicing, the greater sequence identity of the *T. vaginalis* PRP8 homolog with *cis*-splicing PRP8s than with the trypanosomal PRP8 could be significant. Currently there is no evidence for *trans*-splicing in trichomonads (all messages examined to date lack 5' mini-exons), and the presence of a PRP8 protein more similar to those involved in *cis*-splicing indicates that introns are very likely present, although the possibility that *trans*-splicing also takes place cannot be ruled out. If trichomonads did diverge earlier than trypanosomes (as seems the likeliest scenario based on current phylogenetic information) then the presence of *cis*-splicing in trichomonads would also contradicts the notion that *trans*-splicing in trypanosomes represents an intermediate stage in intron evolution that gave rise to *cis*-splicing spliceosomal introns (see Bruzik and Steitz 1990; Palmer and Logsdon 1991). In fact, it looks more and more likely that *trans*-splicing arose independently in trypanosomes (as in nematodes and trematodes) since the presence of both *cis*- and *trans*-splicing has recently been discovered in *Euglena,* a close relative of trypanosomes (Breckenridge et al. 1999).

Although recent evidence for the phylogenetic placement of putative early branching eukaryotes has left their position in some question (where the best example being the phylogenetic re-assignment of the microsporidia - see subsequent chapters and also Embley and Hirt 1998; Philippe and Adoutte 1998), there is currently no molecular evidence suggesting that parabasalia (and for that matter, diplomonads) are not the deepest branches in eukaryotic trees. Their previous status as primitively amitochondrial has been dismissed, however, as genes of mitochondrial origin have been found in both *T. vaginalis* and *G. lamblia* suggesting that at one time both harboured such symbionts/organelles (Horner et al. 1996; Bui, Bradley and Johnson 1996; Germot, Philippe and Le Guyader 1996; Roger, Clark and Doolittle 1996; Roger et al. 1998). Indeed, trichomonads possess hydrogenosomes which are probably derived from mitochondria (Müller 1997; Bui, Bradley and Johnson 1996). Here I have presented evidence that *T. vaginalis* possesses the spliceosomal component PRP8, and predicted that functional spliceosomes could therefore be present. I further predict that although introns have not yet been discovered in trichomonad genes, that they are present (albeit at a low density). The prevailing view of the origin of spliceosomal introns is that they arose from a progenitor Group II intron (Sharp 1991), possibly imported with the genome of the proto-mitochondrial endosymbiont (Cavalier-Smith 1991). Thus, the prediction that introns are present in parabasalia is not inconsistent with current understanding of the history of mitochondria in these organisms.

CHAPTER II

Spliceosomal snRNAs in *Nosema locustae*

This chapter includes work published in Fast, N.M., A.J. Roger, C.A. Richardson and W.F. Doolittle. 1998. U2 and U6 snRNAs in the microsporidian *Nosema locustae*: evidence for a functional spliceosome. *Nucl Acids Res.* **26**:3202-3207.

New sequences have been deposited in Genbank under accession numbers AF053588 and AF053589.

INTRODUCTION

The removal of intervening sequences, or introns, from RNA involves two transesterification reactions mediated by the spliceosome. An extremely large ribonucleoprotein complex (comparable in size to the bacterial ribosome), the spliceosome is composed of many proteins and five small nuclear RNAs (snRNAs): U1, U2, U4, U5 and U6 (Will and Luhrmann 1997). Each of these RNAs associates with its own group of specific proteins, along with the common core Sm proteins, to form a small nuclear ribonucleoprotein (snRNP) complex. The U1 snRNP complex recognizes the 5′ splice site, while the U2 complex interacts with the intron branch point, aiding in presenting the nucleophilic adenosine. U4, U5 and U6 enter as a tri-snRNP to complete assembly of the spliceosome, and hence, allow the two transesterification reactions to proceed (see introductory figures). The enzymatic activity responsible for the reactions of splicing is generally attributed to the RNA, due to the similarity of the chemical reactions involved in splicing both spliceosomal and autocatalytic Group II introns. However, a substantial catalytic role for the proteinaceous component of the spliceosome cannot be ruled out. These possibilities, along with a comprehensive review of the interactions within the spliceosome are presented in Nilsen 1998 and in Staley and Guthrie 1998.

Of the snRNA components of the spliceosome, U6 and U2 show the highest degree of sequence conservation (Madhani and Guthrie 1992), with U6 being the more highly conserved in sequence of the two, as it base-pairs extensively with U4, and also with U2 and the intron itself (Madhani and Guthrie 1994; Brow and Guthrie 1988). All known U6 snRNA sequences also contain the conserved U6 intramolecular helix, which further constrains the sequence (Wolff and Bindereif 1993). By comparison, the U2 snRNA sequence is less conserved. In addition to the regions of pairing with U6, U2 snRNAs (with the exception of those from the trans-splicing trypanosomes) contain the conserved GUAGUA which pairs with the intron branch site, along with several nucleotides in regions of conserved secondary structure, such as those recognized by proteins (Guthrie and Patterson 1988).

In 1996, DiMaria and coworkers described a highly unusual U2 snRNA homolog from the microsporidian parasite *Vairimorpha necatrix* (DiMaria et al. 1996). This RNA reportedly had no Sm binding site, and no conventional tri-methyl guanosine cap structure at its 5' end. In addition, the secondary structure they proposed showed many unusual features, including the absence and alteration of otherwise highly conserved structures, and the introduction of unique stem loops.

The relatively odd nature of this snRNA left its role in splicing in some doubt, and added to an already lengthy list of uniquely odd features that characterizes the microsporidia. As eukaryotic intracellular parasites, microsporidia survive outside a host cell as hardy spores surrounded by a thick layer of chitin and protein. Spores infect a host cell by an unusual mechanism: an organelle known as the polar tube everts from its tightly wound position within the spore to pierce the host's cell membrane and allow the parasite to inject itself into the host. Microsporidia appear to lack such typical eukaryotic features as mitochondria, stacked golgi, peroxisomes, and 80S ribosomes (microsporidian ribosomes are 70S) (Cavalier-Smith 1993). Another unique feature is the tiny size of microsporidian genomes. At the extreme, the 2.9 Mb genome of *Encephalitozoon cuniculi* is the smallest known nuclear genome, smaller than the genomes of many bacteria (Biderre et al. 1995). Other microsporidia possess genomes on the order of 5-6 Mb, still extremely small by eukaryotic standards and thus posing interesting questions with respect to genome organization and the possible reduction or loss of non-coding regions (Biderre et al. 1995; Keeling and McFadden 1998).

At the time of this work no introns had been found in microsporidian genomes, but fewer than 25 microsporidian gene sequences had so far been reported. For microsporidia and other protists with few sequenced genes, the presence of potentially functional components of the spliceosome may provide the best, albeit indirect, evidence for the occurrence of splicing. Given the unusual structure proposed for the *V. necatrix* U2 snRNA, I chose to search for genes for that snRNA and its partner, U6 snRNA, in a second parasitic microsporidian, *Nosema locustae*. Both snRNA genes were found, both are expressed, and both RNAs can form the intramolecular and intermolecular secondary structures typical for these RNAs in organisms known to have splicing. Furthermore, an additional small RNA that is likely U4 spliceosomal snRNA has been identified based on its complementarity to the *N. locustae* U6 snRNA sequence.

```
               -116                         -87
Nosema         TTATGGCACACCTGGGGTGCGATGATGCCATTTCTGCAAGCGTACCTACAACGACTT
               -58                          -29
               TCTGCTGCAACACGAGTGCAAAGCAACTAAAATAAAATTTGTACAGAAAACAACCCCT

               +1                           +30
Nosema         GCAGTTTGCTGCGCTATTAGTTTGGAACAACACTGAGAAGATTAGCATGGCCCCTGCG
Saccharomyces  CTTCG.G.ACATTTGG.C.A....A.....T..A....T...C....GTT.......A
Arabidopsis    CTTCGGG.ACATC.G..A.AA.......G.T..A........................
Homo           T.G.CAGCACATATAC.A.AA.......G.T..A........................
Caenorhabditis C.GA--GAACATATAC.A.AA.........T..A........................
Trypanosoma    TTC--GG.GACAT.C.CA.AC.G.A..TTCA..A.............CTCT.......
Entamoeba      TTC.G.G.AAATC.C..A.AA.........T..A........................
               +59                          +87
Nosema         CAAGGACGGCATCTTTCTT-TGAGAGGTGTGCTGGGCTCGCCCAGCTTTT
Saccharomyces  T.....T.A-.C.G..T.ACAA....A.T.AT.TC.T.TTTTTTTTA.C
Arabidopsis    ......T.A..CGCA.AAATC....AA..GT.CAAAT.T
Homo           ......T.A..CGCAAA..-C.T..A.C..T.CATAT.TTT
Caenorhabditis ......T.A..CGCAAA..-C.T..A.C..T.CAAAT.TTT
Trypanosoma    .....CT.A-TGTCAATC.TC.....A.A.AGCTTTT
Entamoeba      ......T.A..CGC.CA.A-C....A..TAAAAA.TT.TT

               +108                         +137
Nosema         GTGCTAACATTCTTTTTTACTTCGTAACGCAACCACTATGCACAGAGCTCCAGCTTTT
```

Figure 2-1 Alignment of a selection of U6 snRNA gene homologs with the *N. locustae* U6 snRNA coding sequence

Numbering of nucleotides is shown for the *N. locustae* sequence, with the predicted 5' start site labelled as +1. Nucleotides matching those of *N. locustae* are indicated with dots (.) and gaps introduced into the alignment are shown as hyphens (-). Upstream and downstream regions of the *N. locustae* sequence are also shown.

RESULTS

The *N. locustae* U6 snRNA gene sequence

Screening of the *N. locustae* genomic library with the probe designed to a highly conserved region of U6 produced several positive signals and subsequent independent genomic clones. Of these, two were chosen for further analysis. Restriction digestion and Southern hybridization with the U6 probe allowed ~500 bp subclones to be generated containing the DNA sequence of interest. Sequencing of the subclones revealed a single sequence corresponding to a clear homolog of the U6 snRNA gene, sharing extensive identity with others. The 5' end of the molecule was assigned based on secondary-structure predictions (discussed below) and confirmed by primer extension (results not shown). The 3' end assignment is based upon length estimation from northern blot analysis (discussed below).

Figure 2-1 shows the *N. locustae* U6 gene sequence aligned with known homologs from selected taxa to emphasize overall conservation, particularly in the central region or "catalytic core" of the molecule. One striking exception is nucleotide 34T, a unique base otherwise conserved as an A at that position in almost all other taxa. In fact, this nucleotide is part of the so-called "phylogenetically invariant" ACAGAGA heptad (the normally invariant nucleotide is underlined) and its functional significance has been shown by mutational analyses in both mammalian and yeast systems (Wolff et al. 1994; Datta and Weiner 1993; Fabrizio and Abelson 1990; Madhani, Borden and Guthrie 1990).

U6 snRNA genes are transcribed by RNA polymerase III. All have been found to have upstream promoter elements that consist of distal, proximal and TA-rich sequence elements (Eschenlauer et al. 1993; Brow and Guthrie 1990). An

```
                -91                   -69
Nosema          CATGCACATATTCTACACAGCTACTGAATCCCGAAATAAACCGCAAT
                -44                   -22                   +1
                ATTTACATCAAAAGGAGTGATATAATCCCAAATTCAAGCCCTCCACC

                +4                    +26
Nosema          TCTCAAAGCTCATAGCTTTGATCAA-GTGTAGTATCTGTTCTTGTCA
Vairimorpha     ........T.ATCG..........-................A...G
Saccharomyces   CTCTTTGC..TT.G....A.....-.................T...
Arabidopsis     .TCTCGGC..TT.G...AA.....-.................A...
Homo            ...-CGGC..TT.G...AA.....-.................A...
Caenorhabditis  ...TCGGCT.AT.....AA.....A.................A..G
Trypanosoma     CTA.GG.A..TT.G.ACAA.G-..CT.CA..TCT...CGG..AT.T.

                +50                   +72
Nosema          GTGTGACAGCTGACAAACTAGCTCCTTGGAGC-TAGAATATGCTGGT
Vairimorpha     .CT.A...A.....---T..T.CG.AA....--..T....CTTA.AA
Saccharomyces   ....A...A....-..TGACCTCAA.GA.GCTCATT.CCT.TTAAT.
Arabidopsis     ..T.A.T.T....T.TGTGG..CATCG-.CC.AC.CG....TAACTC
Homo            ..T.A.T.T....T.CGTCCT..ATCC-...GAC.AT....TAAATG
Caenorhabditis  TAT.A..CTAC.GT.T..ACT.GAA.--...TG..AT.A.G.T.ATA
Trypanosoma     .CTAAGATCAA.TT.TTAA.CTGTTC.TATCAGAGT..CTCCTGATA

                +96                   +118
Nosema          GTGTGTGTGGATGCTTTGACAGGCTTGCTTGTAGGGGCCATGCACAC
                +143                  +165
                ACCAGGCAGACTCCCGGAAGTTGTTCCGTCCGGAGCTGCACTTTTTA
                +190                  +212
                TTTAAGACAATCATAGAGTGCTACTTCCAGCAAATCTGGTGTCCTTC
                +237                  +259
                CAAAGATAAAATATGTAGTTGAGCAACAAACGCTGTTTACTTACTTG
```

Figure 2-2 Alignment of the 5' conserved regions of a selection of U2 snRNA gene homologs with that of *N. locustae*

Numbering and symbols shown are the same as those in Figure 2-1. Flanking 5' and 3' regions of the *N. locustae* U2 snRNA sequence are also presented.

intragenic A Block element and a downstream B Block element have been characterized in yeast, but appear to be absent from other well studied (i.e. mostly vertebrate) U6 genes (Eschenlauer et al. 1993). Examining the upstream sequence of the *N. locustae* U6 homolog reveals a prominent TA-rich element, TAAAATAAAA, found at -22 to -31 (Figure 2-1) which is consistent in both sequence and position with those previously identified. A possible A Block element is the highly conserved AGATTAGCATGG found at position +39 to +50 (Figure 2-1), identical in sequence to that recently predicted for *Entamoeba histolytica* (Miranda et al. 1996). Although sufficient upstream sequence was available, the proximal sequence element (PSE) (generally found at ~ -45) could not be identified, possibly because the PSE does not always have a high degree of phylogenetic conservation (Li, Haberman and Marzluff 1996; Thomas et al. 1990). The presence of the distal sequence element (located at ~-200 to -300) and the downstream B Block element (located at ~+250) were not addressed in this study, as they are not expected to be included in the examined clones.

The *N. locustae* U2 snRNA gene sequence

The sequence of the U2 snRNA gene does not show the same degree of overall conservation as does U6. However, a short region of the molecule corresponding to that which recognizes the intron branch site is extremely well conserved and can be used as a probe (Ares 1986). Screening of the *N. locustae* genomic library with such a probe (U2-L15) produced many signals and subsequent genomic clones that were analyzed by restriction digests and Southern blotting. Four different, independent clones were then selected for further subcloning and sequencing. Sequencing both strands of all clones revealed a single sequence with a high degree of sequence conservation in the region surrounding and including the branch point recognition sequence,

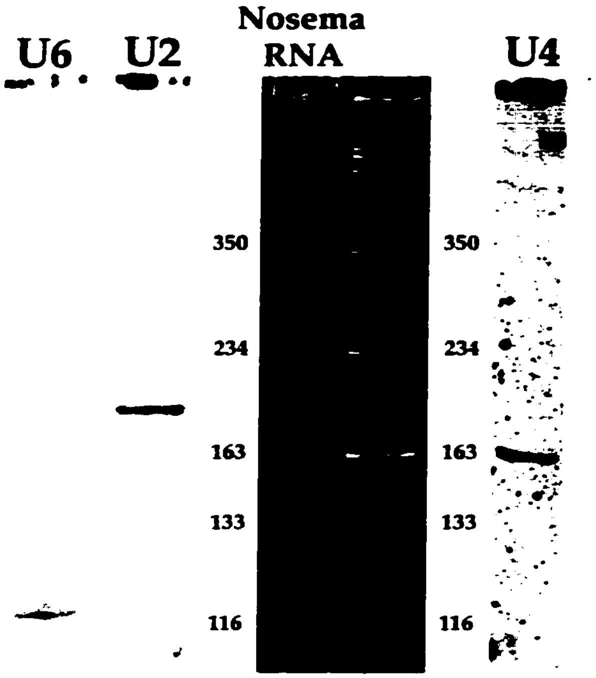


Figure 2-3 ***N. locustae*** **expresses U6 snRNA, U2 snRNA and (putative) U4 snRNA**
Northern blot analysis using probes based on the genomic sequences of U6 and U2 snRNA genes, in addition to a short probe predicted to be complementary to U4 snRNA based on the predicted U6-U4 snRNA interaction characterized in eukaryotes known to splice.

suggesting that this is a true U2 homolog. Additional evidence for this suggestion comes from the greater degree of overall similarity to the microsporidian *V. necatrix* U2 homolog compared to others. The conserved, 5' region of the *N. locustae* U2 sequence is aligned with that from *V. necatrix* and other representative taxa in Figure 2-2. Prediction of the *N. locustae* U2 snRNA 5' boundary is based upon the consensus position of the start site relative to the conserved secondary structure. This predicted 5' boundary (Figure 2-2) gains additional support from its close similarity (within one nucleotide) to that of the sister microsporidian, *V. necatrix* (DiMaria et al. 1996). The 3' end of the *N. locustae* U2 snRNA (+188) (Figure 2-2) is estimated based on a rough size calculation from the northern analysis (below).

U2 snRNAs characteristically have an Sm-binding site (general consensus: RR(U)$_{2\text{-}6}$RR). Such a possible site in the *N. locustae* U2 sequence is located between nucleotides 105 and 114 with the sequence GAUGCUUUGA (Figure 2-2). This site's position with respect to the predicted secondary structure of the *N. locustae* U2 snRNA is discussed later. U2 snRNA genes characterized to date also have an upstream TA-rich box, necessary for transcription of the gene by RNA Polymerase II. The TATAA element at -20 to -24 of the *N. locustae* sequence (Figure 2-2) is a predicted TATA Box.

Both the U6 and U2 snRNA gene sequences of *N. locustae* are clearly homologous to typical spliceosomal snRNAs, rather than to those involved in AT-AC intron splicing (Tam and Steitz 1997).

The *N. locustae* U6 and U2 snRNAs are expressed

To determine whether these snRNA genes are expressed by *N. locustae,* northern blot analysis was carried out using probes generated by PCR amplification of internal regions of the two genes. As shown in Figure 2-3, both

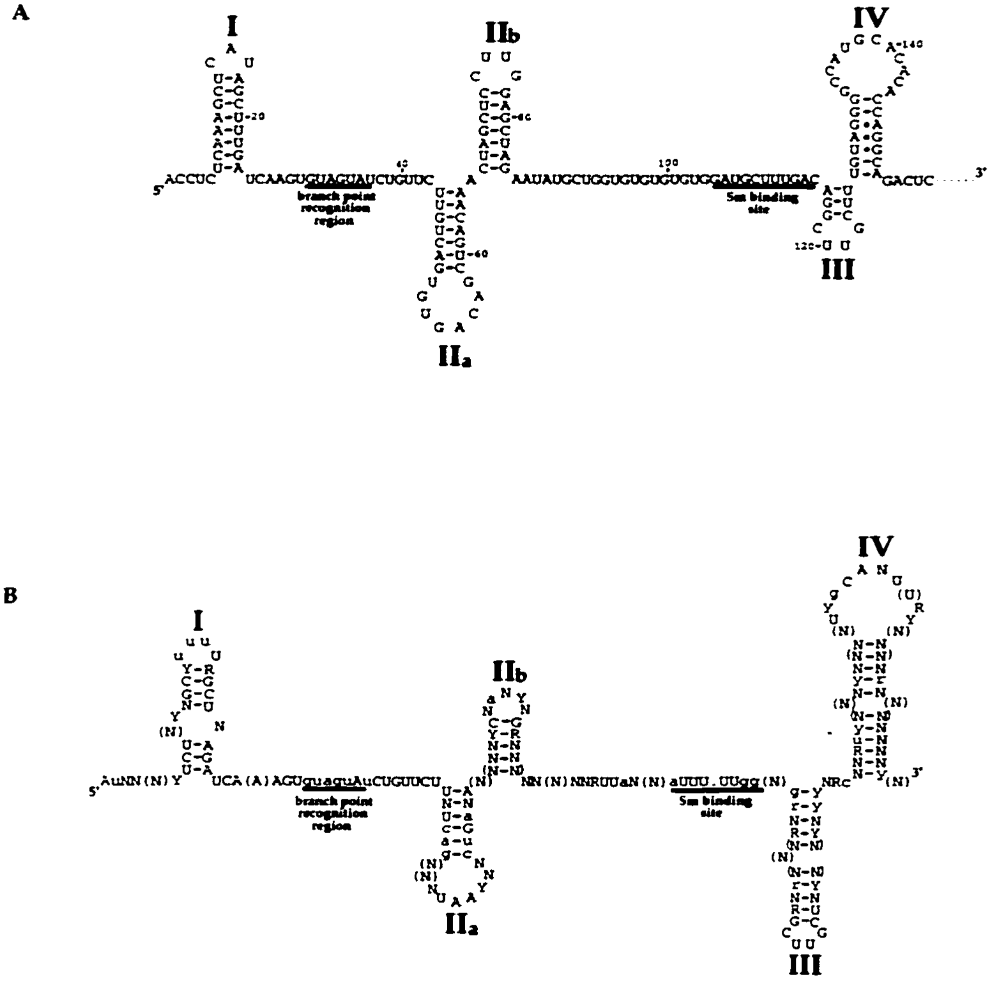


Figure 2-4 Secondary structure models for the *N. locustae* U2 snRNA and the consensus U2 snRNA

(A) The predicted *N. locustae* secondary structure was generated by the mfold software package and further folded by eye. Stem-loops are labelled based upon their corresponding structures in the consensus U2 folding.

(B) U2 consensus secondary structure redrawn from Guthrie and Patterson (1988) and based on an alignment of 12 U2 snRNA homologs. Invariant nucleotides are indicated in uppercase, while those identical in all sequences except one are shown in lowercase. Parantheses indicated length variability.

RNA species are expressed. The *N. locustae* U6 snRNA length is estimated as 110-115 nt, consistent with that of other U6 snRNAs. When the degree of migration of the *V. necatrix* U2 was calculated from the northern analysis shown by DiMaria et al. (1996) and compared to the *N. locustae* U2 in this study, it was found that both migrate with RNA species of approximately 180-185 nt in length. This indicates that the *N. locustae* and *V. necatrix* U2 snRNAs are very close in length, assuming that such closely related microsporidia share the same cap structure.

N. locustae possesses U4 snRNA

U6 snRNA enters the spliceosome closely base-paired with U4 snRNA, and this pairing has been proposed to sequester the "active sites" of the U6 snRNA prior to the activity of splicing. Pairing occurs in the region of U6 encompassing the U6-U2 intermolecular helix I and the U6 intramolecular helix, forming a U6-U4 interaction involving a pair of intermolecular helices separated by an intramolecular U4 helix. Bearing in mind the sequence and secondary structure of the *N. locustae* U6 snRNA, I predicted the sequence of the potential *N. locustae* U4 snRNA in the region of pairing with U6, presuming it would pair with U6 snRNA in the characteristic manner found in organisms known to splice. Using a short 17mer labelled oligonucleotide complementary to the predicted U4 sequence, I probed the northern blot described above. Autoradiography revealed a single band corresponding to a small RNA co-migrating with species of ~165 nt in length (Figure 2-3). This is on par with the sizes of known U4 snRNAs, and therefore this microsporidian RNA is likely the *N. locustae* U4 snRNA.

Secondary structure modelling

Conservation of the secondary structure of both U6 and U2 snRNAs, as well as the conserved conformation of their intermolecular interaction with one

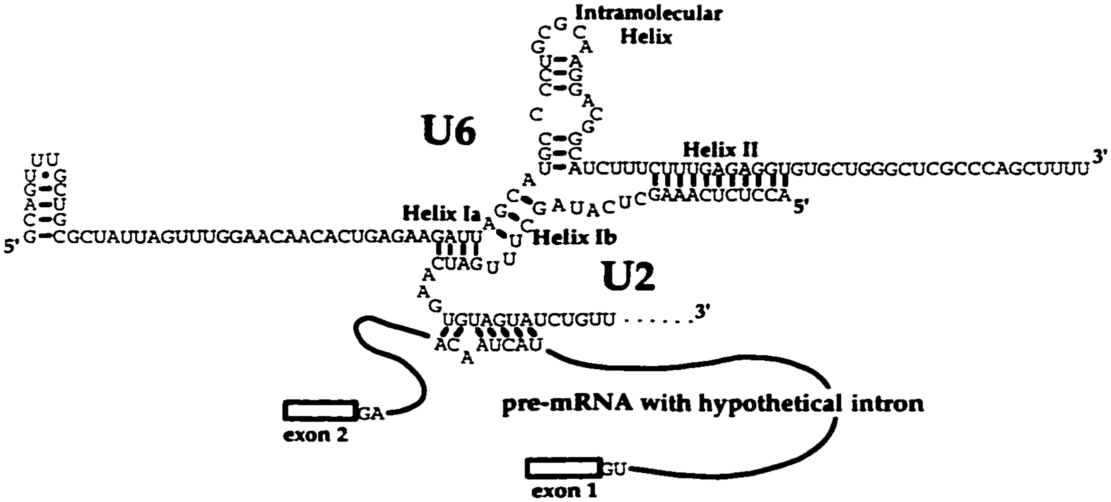


Figure 2-5 The potential intermolecular interactions between the *N. locustae* U6 and U2 snRNAs

Helices Ia, Ib and II are assigned based on homologous functional pairings characterized in other organisms. The conserved U6 intramolecular helix and 5' stem-loop, along with the potential interaction of U2 snRNA with a canonical intron branch-site are also shown.

another, is thought to reflect their catalytic significance in the spliceosome (Madhani and Guthrie 1992; Madhani and Guthrie 1994). Computer-assisted free energy minimalization using the mfold server (http://www.ibc.wustl.edu/~zuker/rna/form1.cgi), followed by additional folding by eye, generated the secondary structure for the *N. locustae* U2 homolog shown in Figure 2-4a. This structure is quite similar to that predicted for characterized functional U2s. When compared to a "consensus" structure shown in Figure 2-4b, support for the positions of stem-loops I, IIa, IIb, III and IV with respect to the branch point recognition region and the Sm binding site is evident. In addition, the *N. locustae* stem-loop I is supported by the first two conserved C-G and U-A pairings at the base along with the C-G and G-C pairings below the loop. Likewise, stem-loop IIa shares the conserved pairings along the stem and has a characteristic sized loop, while stem-loop IIb contains the conserved C-G pair nearest the loop. While an alternative structure for the 5' region of U2 snRNA has been suggested (Zavanelli et al. 1994), we could not find evidence for such a potential pairing in *N. locustae*. The 3' portion of the structure is less similar in sequence, but does maintain the typical stem-loops. The predicted Sm binding site gains support by its proximity to a stem-loop that, although shorter in stem length, shares the same loop sequence as stem-loop III, and is labelled as such in Figure 2-4a. Canonical stem-loop IV contains a YGCA sequence, a probable conserved protein-binding motif (Tang, Abovich and Rosbash 1996). The UGCA in the large, 12 nt loop of the *N. locustae* U2 secondary structure is identical, and along with the size of the loop, lends credence to this being a true stem-loop IV.

Taken alone, the conservation of the *N. locustae* U2 snRNA structure would imply that it could function in pre-mRNA splicing. Additional evidence is provided by examining the potential intermolecular interactions between the *N.*

locustae U2 and U6 snRNAs. Extensive mutational analyses have been carried out in yeast and mammalian systems, providing a framework of functional pairings between the two snRNAs against which the *N. locustae* interactions can be compared (Madhani and Guthrie 1992; Wolff et al. 1994). The predicted folding of the *N. locustae* U6 and U2 snRNAs is depicted in Figure 2-5.

The conservation of both the *N. locustae* U6 and U2 sequences allows them to form the characteristic Helix Ia and Helix Ib. Of more functional significance is the formation of a strong Helix II. The length and position of the *N. locustae* Helix II are consistent with that found in mammals and yeast (Datta and Weiner 1991; Wu and Manley 1991); however, the sequences in this region are not conserved among snRNAs. In fact, the sequences co-vary to maintain pairing, so identifying such an interaction indicates its potential functional value.

Aside from its interaction with U2, U6 also forms a typical intramolecular helix between Helix I and Helix II (Wolff and Bindereif 1993). This helix is evident in the folded *N. locustae* U6 and is conserved in both structure and many of the conserved nucleotides, including the 5′-most G and C and the 3′-most A within the loop. The 5' stem-loop of U6 snRNAs is also present.

DISCUSSION

We have characterized two integral spliceosomal snRNA components, U6 and U2, from the microsporidian *Nosema locustae*. Although the U2 snRNA is highly conserved in the 5' region surrounding the branch point recognition sequence (GUAGUA), the 3' region of the RNA is less well conserved and is not readily alignable with others. Nevertheless, the region can be folded into the characteristic stem-loops III and IV of known U2 snRNAs, although the stem lengths and loop sizes are different from those of mammals and yeast. The *N.*

locustae stem-loop IV is considerably shorter, but does contain the highly conserved YGCA sequence. This particular motif is found not only in yeast and human U2 stem-loop IV, but also in the U1 stem-loop II (Guthrie and Patterson 1988). Both of these loops are known to bind members of the U1A-U2B" protein family (Tang, Abovich and Rosbash 1996; Scherly et al. 1990), so such a motif in the *N. locustae* loop not only supports this structure, but also suggests that such a protein splicing factor could be present and functioning in *N. locustae*. The proposed stem-loop III has the characteristic loop sequence CUUG, but the stem length is much reduced compared to others. This alteration is probably not functionally significant, however, since stem-loop III can be removed from yeast without abolishing splicing activity (Shuster and Guthrie 1988). The position of the Sm binding site of the *N. locustae* U2 also supports the placement of this stem-loop.

In sequence, the *N. locustae* U2 snRNA is most like that of the related microsporidian *V. necatrix*. This similarity is most striking at the 5' end of the RNAs, with identical nucleotide sequences in the regions proposed here to interact with U6. Downstream of this region, however, the similarity between the two sequences is much less, although the *V. necatrix* U2 can fold similarly to the structure presented here. (We favour this alternative folding, including a putative Sm binding site, which is described (not drawn) by DiMaria et al. (1996).)

Assuming that the *N. locustae* U2 functions in the spliceosome, we predicted that we should also find a U6 snRNA, which could pair with it. This prediction was accurate: *N. locustae* possesses a clear U6 snRNA homolog that is highly conserved in sequence.

One exception to this conservation is U34, an A in other U6 snRNAs and part of the highly conserved (often called "phylogenetically invariant")

ACAGAGA sequence, where position 34 is underlined. Due to the conserved nature and supposed functional significance of this sequence, it has been the target of considerable mutagenesis, both *in vitro* and *in vivo* in yeast, as well as in human cell lines. In *in vivo* studies the A to U mutation is lethal in both yeast and human cell systems (Datta and Weiner 1993; Madhani, Bordonne and Guthrie 1990). The *in vitro* studies generally support this conclusion, although the mutant phenotypes do display variable degrees of splicing (50-100%) (Wolff et al. 1994; Fabrizio and Abelson 1990). In yeast it has been proposed that the ACA of the conserved heptad interacts with a UGU at the 5' splice site located one nucleotide downstream of the canonical GU (Lesser and Guthrie 1993; Kandels-Lewis and Seraphin 1993). As no introns have been characterized in the microsporidia, it could be hypothesized that a currently undiscovered intron contains a trinucleotide other than UGU (for example, AGU) that will satisfy the pairing. It is noteworthy that the intronic UGU sequence is not highly conserved outside of yeast, raising the possibility that this region of the snRNA has a different or possibly redundant function, not clearly understood.

Both the U6 and the U2 snRNA genes of *N. locustae* are expressed. Further support for their functional significance is their potential pairing with each other. There are three regions of defined pairing between functional U6 and U2 snRNAs that can also be postulated for the *N. locustae* U6 and U2, as illustrated in Figure 2-5. These pairings have been identified as functional units in yeast and mammals with mutational analyses (Madhani and Guthrie 1992; Wolff et al. 1994), although there is evidence for redundancy of the helices (Field and Friesen 1996).

In addition to the possession of U6 and U2 snRNAs, I have also presented evidence that *N. locustae* likely possesses U4 snRNA. As the small RNA was identified based on its sequence complementarity to the *N. locustae* U6 snRNA in a region that characteristically interacts with U4 snRNA, it is possible that the putative U4 snRNA and the U6 snRNA could interact together in microsporidian spliceosomes as their homologs do in organisms known to splice.

Furthermore, the folding of the *N. locustae* U6 and U2 snRNAs and their potential to interact with an intron (see Figure 2-5) implies that they could be part of a functional spliceosome, removing introns from pre-mRNA. At the time of this work no such spliceosomal introns have yet been found in microsporidia, and their existence in the genome of *N. locustae* would have to be at a very low density (introns per kb) based upon the absence of introns found in the ~10 kb of protein-coding DNA sequenced in the Doolittle laboratory. Considering the tiny nature of microsporidian genomes, a reduction in the number of non-coding elements such as introns may be a means to achieve such compaction. We think it unlikely, however, that the splicing machinery would be maintained if microsporidia have no introns to splice.

Another possibility is that the snRNAs have been maintained to act in spliced-leader *trans*-splicing, a rare phenomenon that, to date, has only been observed in euglenozoans, nematodes and platyhelminths (Bonen 1993; Nilsen 1993; Nilsen 1995). Currently, there is no supportive evidence from the admittedly limited data about microsporidian gene organization. However, the possibility cannot be ruled out and further data on microsporidian mRNAs may be helpful in testing for the presence of *trans*-splicing in these organisms.

For the past decade, microsporidia have been considered "ancient" (deeply diverging) eukaryotes (Vossbrinck et al. 1987). This notion was supported by their apparent lack of typical eukaryotic features such as mitochondria, stacked golgi, peroxisomes, and spliceosomal introns, features also lacking in other putatively deep-branching protists like the diplomonad *Giardia lamblia* (Cavalier-Smith 1993), where spliceosomal snRNAs have been searched for unsuccessfully

(Niu et al. 1994). More recently, the early divergence of the microsporidia has been called into question: genes of "mitochondrial origin" have been found in microsporidian genomes (Germot, Philippe and Le Guyader 1997; Hirt et al. 1997; Peyretaillade et al. 1998) and protein phylogenies inferred from tubulin (Keeling and Doolittle 1996; Edlind et al. 1996), HSP 70 (Germot, Philippe and Le Guyader 1997; Hirt et al. 1997; Peyretaillade et al. 1998) and the largest subunit of RNA polymerase II (Hirt et al. 1999) suggest a recent divergence, with the microsporidia branching from within, or as a sister-group to fungi. The fact that *N. locustae* possesses two core spliceosomal components, U6 and U2 snRNAs capable of interacting with each other in a functional manner (in addition to a third small RNA, likely U4 snRNA), is consistent with this phylogenetic reassignment. We suggest that microsporidia lost most (but probably not all) of their spliceosomal introns during radical genome size reduction, accompanying the adoption of an intracellular parasitic lifestyle.

ADDENDUM

The prediction of spliceosomal introns in microsporidia based on the presence of snRNA spliceosomal machinery has recently been borne out by the discovery of a putative spliceosomal intron in ribosomal protein L27a (L29 in yeast) of *Encephalitozoon cuniculi* (Biderre, Méténier and Vivarès 1998). Although there is no RNA evidence for the 28 bp intron, it does possess the canonical GT-AG boundaries. Such an intron is quite short, but small introns are in keeping with the extremely reduced genome sizes of microsporidia. Another prediction made in this chapter was based on a difference in the "phylogenetically invariant" sequence of U6 snRNA in *N. locustae*. As this region is known to interact with a trinucleotide just downstream of the 5' splice site in yeast introns

(UGU in yeast), we hypothesized based on the U34 in *N. locustae* U6 snRNA that the trinucleotide in a microsporidian intron could be AGU. This is, in fact, the case for the *E. cuniculi* putative intron. It is not unreasonable to assume that *E. cuniculi* also possesses a U6 snRNA with U34, in which case it could pair in the fashion shown in Figure 2-6. This potential interaction between the microsporidian U6 snRNA and the putative intron certainly suggests that the *E. cuniculi* intervening sequence is a *bona fide* spliceosomal intron.

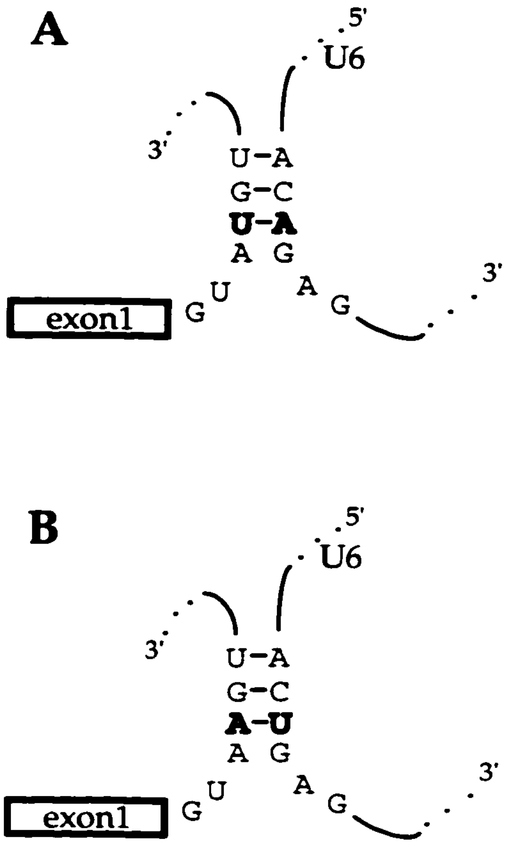


Figure 2-6 U6 snRNA-intron interaction in yeast and microsporidia
(A) In *S. cerevisiae* the bold-faced A in the U6 snRNA ACAGAGA element pairs with a conserved U found at the intron position 4. (Pair in question is shown in bold print.)
(B) Predicted microsporidian pairing based on the U6 snRNA sequence from *N. locustae* and the putative intron sequence from *E. cuniculi.* Note the compensatory changes in the U6 snRNA (ACUGAGA) and the A at position 4 of the intron.

CHAPTER III:

The TBP Gene of *Nosema locustae*

This chapter includes work published in Fast, N.M., J.M. Logsdon, and W.F. Doolittle. 1999. The TATA-box binding protein gene of *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Mol Biol Evol*. (in press).

New sequence has been deposited in Genbank under accession no. AF144035.

INTRODUCTION

The classification of the microsporidia as members of the Archezoa originally seemed accurate, as evidence from molecular phylogenetic analyses of EF-1 alpha, EF-2, and small subunit (SSU) rRNA are consistent with a deep-branching position at the base of the eukaryotic tree (Kamaishi et al. 1996a; Kamaishi et al. 1996b; Vossbrinck et al. 1987). Other characteristics of the microsporidia, such as the apparent lack of mitochondria, stacked golgi and peroxisomes, along with the possession of 70S ribosomes, also appeared to point to an ancient origin. However, during the course of this work a striking alternative has been offered by new phylogenetic evidence (See Keeling 1998; Keeling and McFadden 1998; Müller 1997; Embley and Hirt 1998).

Analyses of alpha- and beta-tubulin sequences were the first to suggest a relatively recent origin of microsporidia, as both gene trees show a specific relationship between the microsporidia and the fungi (Keeling and Doolittle 1996; Edlind et al. 1996)—a conclusion now supported by analyses of the largest subunit of RNA polymerase II (Hirt et al. 1999). The more recent finding that microsporidia possess mitochondrial-type HSP 70 suggests that microsporidia are secondarily amitochondriate, and phylogenetic analysis of the HSP 70 sequence supports a later origin of the microsporidia, with the sequence branching with weak support with those from fungi in some analyses (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998).

The conflicting evidence regarding the position of the microsporidia led me to seek other phylogenetic markers, as the database currently contains fewer than twenty-five protein coding gene sequences, hence limiting the potential for phylogenetic analysis. Here I describe the sequence of the main component of the transcription factor TFIID, the TATA box binding protein (TBP) from *Nosema locustae*.

TBP is a universal transcription factor, involved in transcription initiation with all three eukaryotic RNA polymerases. Its role has been best characterized for pol II transcription, where it is the primary subunit of the multimeric transcription factor TFIID. TFIID recognizes the upstream TATA element and promotes DNA bending, which in turn allows other factors of the pre-initiation complex to bind (Burley and Roeder 1996). With the exception of the N-terminal region of TBP (which varies greatly in sequence and length) the sequence of the protein is highly conserved.

In addition to containing phylogenetic information, TBP was also chosen as a good candidate sequence from which to learn about intron density within the microsporidia. Another feature of the microsporidia that originally classified them as Archezoa was their apparent lack of spliceosomal introns. As presented in the previous chapter, my discovery of snRNAs that have the potential to interact in a functional way strongly predicts that such introns should be present. However, they may be present at a low density, and simply are not represented since only relatively few microsporidian protein-coding genes have been sampled. Indeed, with the number of microsporidian protein-coding genes sampled thus far, the intron density in microsporidia could be on par with that of *Saccharomyces cerevisiae* (J. Logsdon pers. comm.; see also Logsdon 1998).

One strategy to search for microsporidian introns is to examine genes that would have a better-than-average chance of containing a spliceosomal intron. An example of such a gene could be one that contained an "old" intron position. These "old introns" (defined in Logsdon et al. 1995) are found at identical positions in diverse plants, animals, and fungi, and therefore likely predate the divergence of these groups. Such introns are rare, as there are very few introns conserved at this phylogenetic depth (Palmer and Logsdon 1991; Logsdon 1998). The TBP gene contains a pair of "old" intron positions, making it a good candidate for possession of an intron in the microsporidia. Therefore, the TBP gene appeared to be a good choice to study since the sequence was not only potentially informative phylogenetically, but also could provide information about intron evolution in the microsporidia.

To further both ends, I cloned and sequenced the *N. locustae* TBP gene. The phylogenetic position of the *N. locustae* TBP gene sequence favours a late origin of the microsporidia, and shows a weak sister-relationship with the fungi. The full-length sequence of the gene contains no introns, and, bearing in mind the phylogenetic results, this suggests that spliceosomal introns have been lost throughout the course of microsporidian evolution, possibly as a result of reduction in genome size.

RESULTS

N. locustae **possesses a TBP gene homolog that is present in single copy**

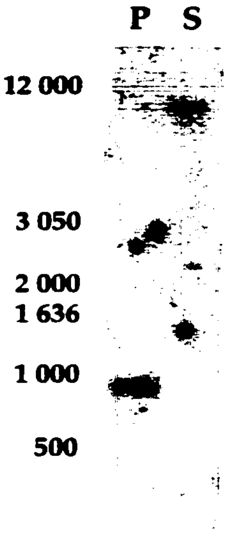


Figure 3-1 Confirming the provenance and single copy number of the *N. locustae* TBP gene

A genomic Southern blot probed with the TBPF1R1-5 probe. All four independent genomic clones contained the same TBP coding and flanking sequences, indicating that the gene is likely present in single copy. Based on the genomic sequencing there are no *Sac*I or *Pst*I restriction sites within the TBP coding region. A *Pst*I single digest is predicted to release a 976 bp fragment complementary to the probe if the gene is single copy; a DNA fragment of approximately that size is seen in lane 1. (P=*Pst*I; S=*Sac*I)

The extensive degree of conservation of known eukaryotic TBP sequences allowed the design of two forward direction and three reverse direction degenerate PCR primers that when used in all combinations produced products of ~300-400 bp in length. Two of these combinations (TBPF1-TBPR1 and TBPF1-TBPR2) produced PCR products that, when sequenced and compared to the database, showed a high similarity to TBP. Multiple clones possessing PCR products of both types were sequenced; a single sequence was obtained and clone TBPF1R1-5 was arbitrarily chosen to be used as a probe.

A genomic Southern blot was performed (shown in Figure 3-1) using the TBPF1R1-5 probe. The blot confirms that the PCR product is in fact from *N. locustae*, and also indicates that the TBP gene is present as a single copy within the genome.

The PCR products cover only approximately 49% of the gene, and do not include all old intron positions, so it was necessary to determine the entire gene sequence. Using the TBPF1R1-5 probe, I screened the *N. locustae* genomic library and four independent genomic clones were sequenced on both strands with ABI primer walking. The predicted open reading frame for the *N. locustae* TBP gene is 259 amino acids in length.

The conserved, central region of the *N. locustae* TBP sequence is shown in Figure 3-2, aligned with a diverse range of plant, animal, fungal, protist, and archaebacterial TBP sequences. It is clear that this is a true TBP homolog based on the high degree of identity. Indeed, the *N. locustae* TBP sequence in the aligned region shown is 75.4% identical to human TBP. A comprehensive alignment of all available eukaryotic TBP sequences and seven representative archaebacterial sequences was constructed with the 181 C-terminal, unambiguously aligned amino acid characters. Phylogenetic analysis was then carried out with this 40 taxon dataset.

Figure 3-2 Alignment of TBP homologs
The conserved, central region of the *N. locustae* TBP sequence is shown aligned with the homologous region from representative plants, animals, fungi, protists and archaebacteria. Numbering is based on the *N. locustae* inferred protein sequence and gaps are indicated with a dash (-). Taxon abbreviations: *N.lo, Nosema locustae; P.ca, Pneumocystis carinii; S.po, Schizosaccharomyces pombe; A.ni, Aspergillus nidulans; C.al, Candida albicans; S.ce, Saccharomyces cerevisiae; A.ca, Acanthamoeba castellanii, D.di, Dictyostelium discoideum; X.la, Xenopus laevis; D.me, Drosophila melanogaster; H.sa, Homo sapiens; M.mu, Mus musculus; C.el, Caenorhabditis elegans; A.th, Arabidopsis thaliana; T.ae, Triticum aestivum; N.ta, Nicotiana tabacum; A.cl, Acetabularia cliftonii, T.th, Tetrahymena thermophila; E.hi, Entamoeba histolytica; P.fa, Plasmodium falciparum; S.ac, Sulfolobus acidocaldarius; M.ja, Methanococcus jannaschii.*

```
      84                                                                     153
N.lo  PALQNVVATVNLNCKLDLKAIALRARNAEYNPKRFAAVIMRIRDPKTTALIFASGKMVVTGAKSEQTSKL
P.ca  PSLQNIVATVNLDCRLDLKTIALQRRNAEYNPKRFAAVIMRIREPKTTALIFASGKMVVTGAKSEDDSKL
S.po  PTLQNIVATVNLDCRLDLKTIALHARNAEYNPKRFAAVIMRIREPKSTALIFASGKMVVLGGKSEDDSKL
A.ni  PTLQNIVATVNLDCRLDLKTIALHARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVVTGAKSEDDSKL
C.al  PTLQNIVATVNLDCRLDLKTIALHARNAEYNPKRFAAVIMRIRDPKTTALIFASGKMVVTGAKSEDDSKL
S.ce  PTLQNIVATVTLGCRLDLKTVALHARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVVTGAKSEDDSKL
A.ca  PTLQNIVSTVNLGCKLDLKNIALHARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVCTGAKSEEASRL
D.di  PTLQNIVSTVNMATELYLKAIALGARNAEYNPKRFAAVIMRIREPKTTALIFKSGKMVCTGAKSEDASRF
X.la  PQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEQSRL
D.me  PQLQNIVSTVNLCCKLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEDDSRL
H.sa  PQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEQSRL
M.mu  PQLQNIVSTVNLGCKLDLKTIALRARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEQSRL
C.el  PALQNIVSTVNLGVQLDLKKIALHARNAEYNPKRFAAVIMRIREPRTTALIFSSGKMVCTGAKSEEASRL
A.th  PTLQNIVSTVNLDCKLDLKAIALQARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVCTGAKSEDFSKM
T.ae  PTLQNIVSTVNLDCKLDLKAIALQARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVCTGAKSEQQSKL
N.ta  PTLQNIVSTVNLDCKLDLKAIALQARNAEYNPKRFAAVIMRIREPKTTALIFASGKMVCTGAKSEQSSKL
A.cl  PELQNVVSTVNLGCTLELKEIAMQARNAEYNPKRFAAVIMRIRDPKTTALIFGSGKMVCTGAKSEQDSRT
T.th  PKLQNIVSTVNLSTKLDLKQIALRARNAEYNPKRFAAVIMRLRDPKTTALIFASGKMVCTGAKTEEDSNR
E.hi  PEIVNVVSRFQLGVKLELRKIVQKAINAIYNPKRFAGAIMRISSPKSTALIFQTGKIVCTGTRSIEESKI
P.fa  LNIHNIISSANLCIDINLRLVAVSIRNAEYNPSKINTLIIRLNKPQCTALIFKNGRIMLTGTRTKKDSIM
S.ac  VNIENIVATVTLDQTLDLYAMERSVPNVEYDPDQFPGLIFRLESPKITSLIFKSGKMVVTGAKSTDELIK
M.ja  IKIVNVVVSTKIGDNIDLEEVAMILENAEYEPEQFPGLVCRLSVPKVALLIFRSGKVNCTGAKSKEEAEI

                                                                     219
N.lo  AAQKFSRIIHKL-GFNTKFADFKIQNIVSSCDTQFSIRLEGLAFAHSNFCSYEPELFPGLIYRMVKP---
P.ca  ASRKYARIIQKL-GFNAKFTDFKIQNIVGSCDVKFPIRLEGLAYSHGTFSSYEPELFPGLIYRMVKP---
S.po  ASRKYARIIQKL-GFNAKFTDFKIQNIVGSCDVKFPIRLEGLAYSHGTFSSYEPELFPGLIYRMVKP---
A.ni  ASRKYARIIQKL-GFNAKFTDFKIQNIVGSCDIKFPIRLEGLASRHHNFSSYEPELFPGLIYRMMKP---
C.al  ASRKYARIIQKL-GFNAKFCDFKIQNIVGSTDVKFAIRLEGLAFAHGTFSSYEPELFPGLIYRMVKP---
S.ce  ASRKYARIIQKI-GFAAKFTDFKIQNIVGSCDVKFPIRLEGLAFSHGTFSSYEPELFPGLIYRMVKP---
A.ca  AARKYARIIQKL-GFAAKFLDFKIQNIVGSCDVRFPIRLEGLAFAHNHYCSYEPELFPGLIYRMVQP---
D.di  AARKYARIIQKL-DFPARFTDFKIQNIVGSCDVKFPIKLELLHNAHTSFTNYEPEIFPGLIYKMIQP---
X.la  AARKYARVVQKL-GFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKP---
D.me  AARKYARIIQKL-GFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHCNFSSYEPELFPGLIYRMVRP---
H.sa  AARKYARVVQKL-GFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKP---
M.mu  AARKYARVVQKL-GFPAKFLDFKIQNMVGSCDVKFPIRLEGLVLTHQQFSSYEPELFPGLIYRMIKP---
C.el  AARKYARIVQKL-GFQAKFTEFMVQNMVGSCDVRFPIQLEGLCITHSQFSTYEPELFPGLIYRMVKP---
A.th  AARKYARIVQKL-GFPAKFKDFKIQNIVGSCDVKFPIRLEGLAYSHAAFSSYEPELFPGLIYRMKVP---
T.ae  AARKYARIIQKL-GFPAKFKDFKIQNIVASCDVKFPIRLEGLAYSHGAFSSYEPELFPGLIYRMRQP---
N.ta  AARKYARIIQKL-GFDAKFKDFKIQNIVGSCDVKFPIRLEGLAYSHGAFSSYEPELFPGLIYRMKQP---
A.cl  AARKYAKIVQKL-GFPAKFTEFKIQNIVGSCDVKFPIRMEPLAYQHQQFCSYEPELFPGLIYRMLQP---
T.th  AARKYAKIIQKI-GFPVQFKDFKIQNIVGSTDVKFPINLDHLEQDHKKFVQYEPEIFPGKIYREFNT---
E.hi  ASKKYAKIIKKI-GYPIHYSNFNVQNIVGSCDVKFQIALRTLVDSDLAFCQYEPEVFPGLVYRMASP---
P.fa  GCKKIAKIIKIVTKDKVKFCNFKIENIIASANCNIPIRLEVLAHDHKEYCNYEPELFAGLVYRYKPTSNL
S.ac  AVKRIIKTLKKYGMQLTGKPKIQIQNIVASANLHVIVNLDKAA-FLLENNMYEPEQFPGLIYRMDEP---
M.ja  AIKKIIKELKDAGIDVIENPEIKIQNMVATADLGIEPNLDDIA-LMVEGTEYEPEQFPGLVYRLDDP---
```

Figure 3-2 Alignment of TBP homologs

	254
N.lo	KIVLLIFVSGKIVLTGAKMRDEIYEAFDNIYPVLTQYKK
P.ca	KIVLLIFVSGKIVLTGAKVREEIYQAFEAIYPVLNEFRK
S.po	KVVLLIFVSGKIVLTGAKVREEIYQAFEAIYPVLSEFRK
A.ni	KIVLLIFVSGKIVLTGAKVREEIYQAFELIYPVLSDFRK
C.al	KIVLLIFVSGKIVLTGAKKREEIYDAFESIYPVLNEFRK
S.ce	KIVLLIFVSGKIVLTGAKQREEIYQAFEAIYPVLSEFRK
A.ca	KIVLLIFVSGKIVLTGAKVREEIYEAFENIYPVLTEYKK
D.di	KVLLLIFVSGKIVLTGAKVREYIYEAFENIYPVLSAFKK
X.la	RIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRK
D.me	RIVLLIFVSGKVVLTGAKVRQEIYDAFDKIFPILKKFKK
H.sa	RIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRK
M.mu	RIVLLIFVSGKVVLTGAKVRAEIYEAFENIYPILKGFRK
C.el	RVVLLIFVSGKVVITGAKTKRDIDEAFGQIYPILKGFKK
A.th	KIVLLIFVSGKIVITGAKMRDETYKAFENIYPVLSEFRK
T.ae	KIVLLIFVSGKIVLTGAKVREETYSAFENIYPVLTEFRK
N.ta	KIVLLIFVSGKIVLTGAKVRDETYTAFENIYPVLTEFRK
A.cl	KIVLLIFVSGKVVLTGAKERTEIYRAFEQIYPVLTQFRK
T.th	KIVLLIFVSGKIVLTGAKTRENINKAFQKIYWVLYNYQK
E.hi	KVTLLVFSTGKVVLTGAKDEESLNLAYKNIYPILLANRK
P.fa	KSVILIFVSGKIIITGCKSVNKLYTVFQDIYNVLIQYKN
S.ac	RVVLLIFSSGKMVITGAKREDEVHKAVKKIFDKLVELDC
M.ja	KVVVLIFGSGKVVITGLKSEEDAKRALKKILDTIKEVQE

Figure 3-2 Alignment of TBP homologs

Phylogenetic analysis with the *N. locustae* TBP sequence

Phylogenetic analyses of TBP sequences were carried out with distance, parsimony, and maximum likelihood methods. Maximum likelihood distances of the alignment were computed using the JTT matrix, considering invariant sites along with eight variant gamma categories to account for site-to-site rate variation with PUZZLE 3.1 (Strimmer and Von Haeseler 1996). Tree construction was then carried out with both neighbor-joining and Fitch-Margoliash algorithms (Felsenstein 1993). Both methods gave topologies differing only in the (unsupported) position of the *Dictyostelium* sequence. The Fitch tree is shown in Figure 3-3. The major eukaryotic groups of plants, animals, and fungi are well supported in the distance trees; however, the sisterhood of plants and fungi is not well supported and is probably erroneous (based on a well established animal-fungi relationship (Baldauf and Palmer 1993)). Also unsupported in these trees are the positions of *Acanthamoeba* and *Acetabularia*. Several divergent protist sequences branch with high support at the base of the eukaryotic tree, but these are all relatively long branches and may be artificially drawn to their basal position. The position of the *N. locustae* sequence within this tree is not highly supported, but does show an affinity for branching with the fungal sequences (42 and 47% bootstrap support of the microsporidia-fungi clade for Fitch-Margoliash and neighbor-joining respectively).

A parsimony analysis was also carried out with the TBP dataset, although the extensive conservation of TBP resulted in the recovery of many (>500) equally parsimonious trees; plant, animal and fungal clades were all recovered, but the relationships among the taxa within these clades were not, due to their extensive sequence identity. Notably, a strict consensus of the equally parsimonious trees showed that all trees included the microsporidia-fungi clade.

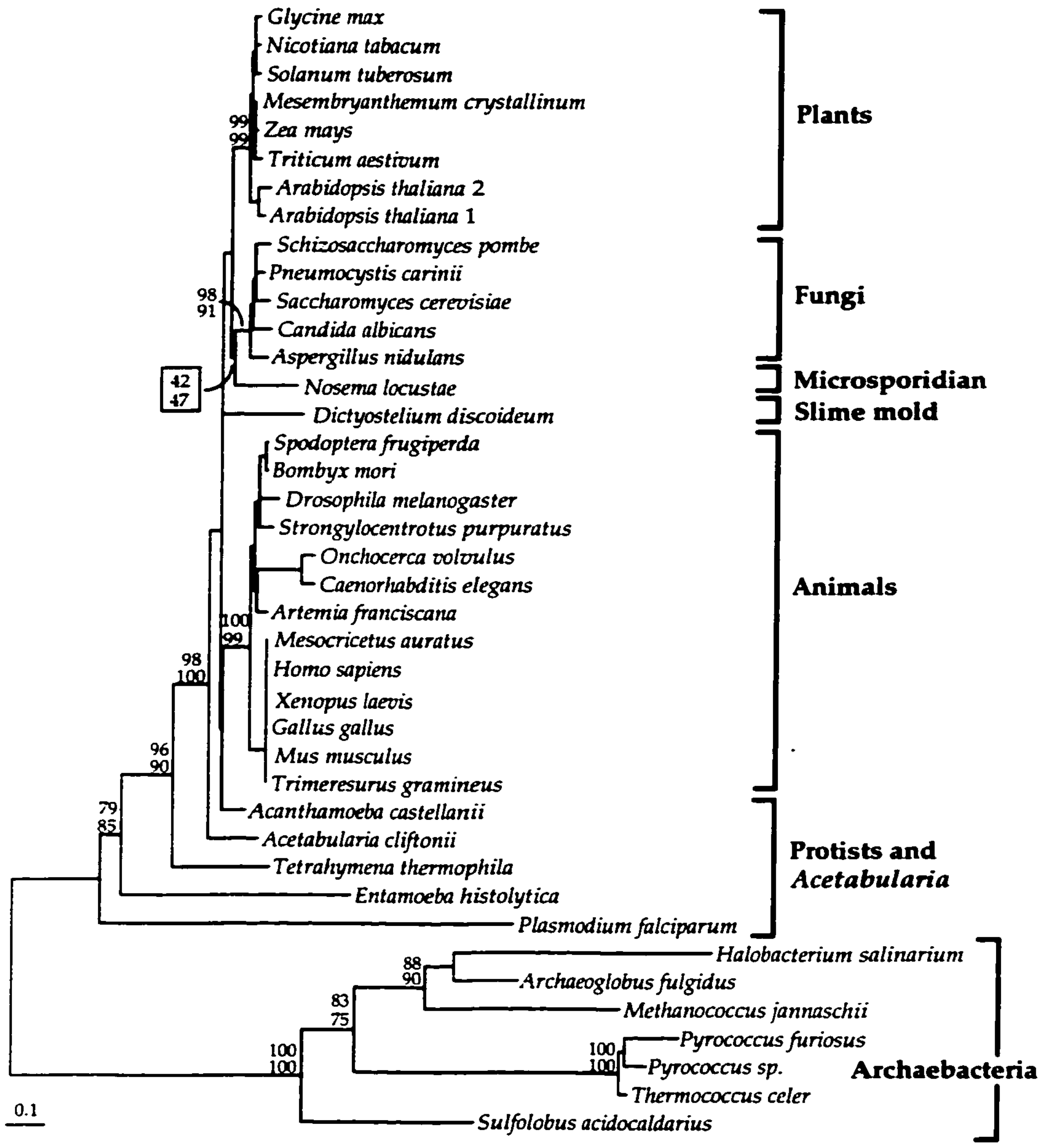
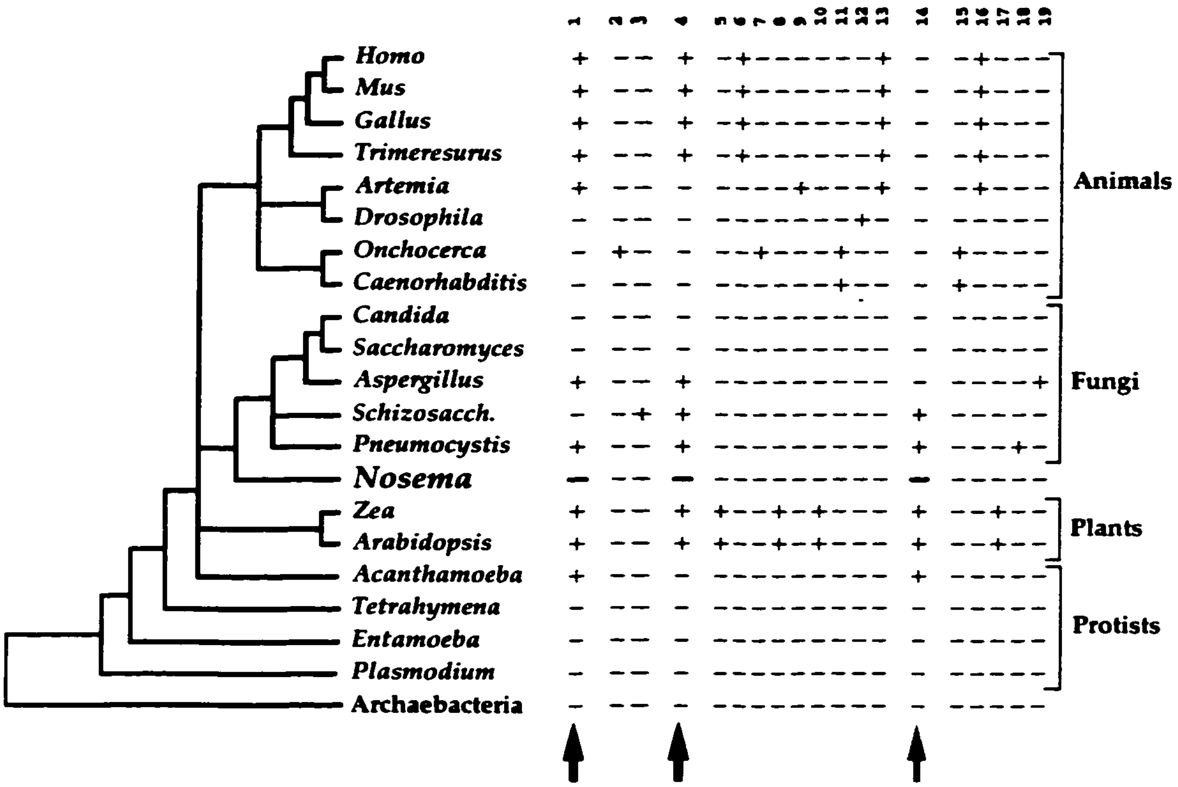


Figure 3-3 Fitch tree of TBP sequences
Bootstrap values based on 100 resampling replicates are indicated for selected groups (where over >50%), with Fitch values above those for neighbor-joining.

An exhaustive protein maximum likelihood (ProtML) search (Adachi and Hasegawa 1996) was also undertaken, constraining the well-defined plant, animal and fungal groups defined in both the distance and the parsimony analyses. The *Nosema* sequence, along with sequences from taxa possessing apparent long branches and/or uncertain positions in the tree (*Acetabularia, Dictyostelium, Tetrahymena, Entamoeba* and *Plasmodium*), were allowed to move freely. In order to further reduce the number of movable groups to make the dataset computationally manageable, the archaebacterial outgroup sequences were removed from the analysis. Although the most likely tree does not include the microsporidia-fungi clade (*Nosema* branches with *Dictyostelium, Acanthamoeba* and animals), many of the trees not significantly worse than the best tree (based on K-H values) do group the microsporidia and fungi together. Thus, TreeCons (Jermiin et al. 1997) was used to produce a majority-rule consensus tree by considering the maximum likely tree topology in addition to the topologies of those trees which are not significantly less likely than the ML tree. The trees that differ from the ML tree are weighted with an exponential distribution and the standard errors of the difference in the log likelihood estimates are taken into account, producing a standardized, exponentially weighted consensus tree. At all significance levels (α=0.05-0.001), the *N. locustae* sequence branches with the constrained fungal group. This indicates that the microsporidia-fungi relationship is present in a majority of those trees which were not significantly worse than the best tree and suggests that ML analysis does not exclude a microsporidia-fungi relationship. In sum, in all phylogenetic analyses the microsporidian TBP sequence does not branch at the base of the tree, but instead branches in the "crown" of the eukaryotic tree, with a weak but consistent fungal affinity.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
Homo	+	–	–	+	–	+	–	–	–	–	–	–	+	–	–	+	–	–	–	Animals
Mus	+	–	–	+	–	+	–	–	–	–	–	–	+	–	–	+	–	–	–	
Gallus	+	–	–	+	–	+	–	–	–	–	–	–	+	–	–	+	–	–	–	
Trimeresurus	+	–	–	+	–	+	–	–	–	–	–	–	+	–	–	+	–	–	–	
Artemia	+	–	–	–	–	–	–	–	+	–	–	–	+	–	–	+	–	–	–	
Drosophila	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–	
Onchocerca	–	+	–	–	–	–	+	–	–	–	+	–	–	–	+	–	–	–	–	
Caenorhabditis	–	–	–	–	–	–	–	–	–	–	+	–	–	–	+	–	–	–	–	
Candida	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	Fungi
Saccharomyces	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Aspergillus	+	–	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–	–	+	
Schizosacch.	–	–	+	+	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	
Pneumocystis	+	–	–	+	–	–	–	–	–	–	–	–	–	+	–	–	–	+	–	
Nosema	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Zea	+	–	–	+	+	–	–	+	–	+	–	–	–	+	–	–	+	–	–	Plants
Arabidopsis	+	–	–	+	+	–	–	+	–	+	–	–	–	+	–	–	+	–	–	
Acanthamoeba	+	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–	–	–	–	Protists
Tetrahymena	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Entamoeba	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Plasmodium	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Archaebacteria	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
	↑			↑										↑						

Figure 3-4 Phylogenetic distribution of introns among currently known TBP genes

Intron positions present in the alignable portion of the TBP gene are numbered consecutively with major eukaryotic groups indicated to the right. Presence of an intron is indicated by "+" and absence by "-". Potentially "old" introns are indicated by arrows: intron positions shared by animals, fungi and plants (black), and the position shared by fungi, plants and *Acanthamoeba* (gray).

Intron position analysis of TBP

Of the forty taxa in the TBP dataset, only approximately half of them are represented by genomic sequences, the rest come from cDNA sequencing. Examining the genomic sequences for spliceosomal intron positions reveals that there are 19 different intron positions across TBP genes from diverse taxa. (Collating the intron data for the TBP gene was the collaborative contribution of J.M. Logsdon.) A close examination of the distribution of introns at each position shows that a small number of the introns are present at identical positions in diverse plants, animals and fungi (see Figure 3-4). Introns conserved at this phylogenetic depth are rare and likely predate the animal/fungal/plant divergence, and therefore such introns have been designated as "old introns." In TBP, old introns are found at positions 1 and 4, along with another likely old intron (shared between plants, fungi and *Acanthamoeba*) found at position 14.

The completed genomic sequence of the *N. locustae* TBP coding sequence revealed that there were no introns in the gene. By accepting the fungal ancestry of the microsporidia, as suggested above, this observation indicates that the lack of introns in the *N. locustae* TBP gene is a result of intron loss.

DISCUSSION

The origin of the microsporidia has become a topic of interest of late due to conflicting phylogenetic evidence and the discovery that microsporidia possess genes of mitochondrial origin—making it highly unlikely that the microsporidia diverged prior to the acquisition of the mitochondrion, as the Archezoa hypothesis originally proposed (Cavalier-Smith 1983). Being secondarily amitochondriate certainly questions the microsporidia's position as early-diverging eukaryotes, but since other putative early-branching eukaryotes (for example, G. *lamblia* (Roger et al. 1998)) have now been found to also harbour such genes, it may be that there are no known extant eukaryotes that actually predate the mitochondrial acquisition.

Phylogenetic analyses with microsporidian genes can be problematic, since their sequences tend to be extremely divergent, producing long branches on trees that are difficult to resolve with accuracy. Long-branched taxa often cluster at the base of a tree (see Philippe and Adoutte 1998; Philippe and Laurent 1998), which raises questions about the validity of the primitive, deep-branching position of microsporidia in trees such as small subunit rRNA and EF-1alpha. Such cases are eroding our faith in the small subunit rRNA tree, which is the best support for the microsporidia being early-branching eukaryotes.

Two independent studies of tubulin phylogenies offer a distinct alternative origin for the microsporidia, both retrieving a microsporidia-fungi relationship with strong bootstrap support (Keeling and Doolittle 1996; Edlind et al. 1996). Trees constructed with HSP 70 sequences weakly support a similar relationship (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998). Here, I have presented phylogenetic analyses of the *Nosema locustae* TBP sequence. Again, the microsporidian sequence does not branch at the base of the eukaryotic tree, but instead branches within the "crown" and shows a weak (yet consistent) affinity for branching with the fungal sequences.

Although the bootstrap support for the microsporidian sequence branching with the fungi is weak (42 and 47%; Figure 3-3), its position within the crown is well supported (98 and 100% bootstrap support for the clade that includes the microsporidian and the rest of the eukaryotes to the exclusion of *Plasmodium, Entamoeba, Tetrahymena* and the archaebacteria). Therefore, the TBP phylogeny adds to a growing body of evidence supporting a late divergence of the microsporidia, and favours a fungal origin.

In addition to phylogenetic evidence supporting a microsporidia-fungi sisterhood, there are several shared features that also unite the two groups. In the EF-1α gene of only microsporidia, fungi and animals there is an insertion conserved in position and length (Kamaishi et al. 1996b). Both also have split thymidylate synthase and dihydrofolate reductase genes that are otherwise fused in plants and *Plasmodium* (Vivarès et al. 1996). Similarities in the meiotic cycle have also been proposed (Flegel and Pasharawipas 1995).

Accepting the fungal ancestry of microsporidia, as suggested here and by others, indicates that the lack of introns in *N. locustae* TBP gene is a result of intron loss. Indeed, all three "old" introns in TBP have been lost at least once in other lineages. As well, the presence of these three introns in some fungi indicates that microsporidia have lost them separately from other fungal lineages, unless microsporidia are specifically related to ascomycete yeasts (like *Saccharomyces cerevisiae*). Although it is possible that high titers of spliceosomal introns were never present in microsporidian ancestors, these analyses suggest that at least some intron loss has occurred. Such a loss of non-coding elements may be one way to reduce genome size, concomitant with the adoption of a parasitic lifestyle (Biderre et al. 1995; Keeling and McFadden 1998). Genome size reduction is apparent in the tiny sizes of microsporidian genomes. *N. locustae* possesses a genome of 5.4 Mb which is very small by typical eukaryotic standards (Streett 1994), but the reduction in genomic size in microsporidia can be even more extreme. For example, the *Encepalitozoon cuniculi* genome, at 2.9 Mb, is smaller than many bacterial genomes (Biderre et al. 1995).

During the course of this study, a putative spliceosomal intron has been documented in a ribosomal protein gene of *E. cuniculi* (Biderre, Méténier and

Vivarès 1998). The putative intron resides adjacent to the ATG start codon and is very short: only 28 nucleotides. Although no RNA work is associated with this discovery, it is likely a true intron and its short size may be evidence of genomic reduction. Counting this one intron, and noting the dearth of introns in every other microsporidian gene so far studied, the intron density in microsporidia is extremely low, and may be on par with *S. cerevisiae* (J. Logsdon pers. comm.; see also Logsdon 1998).

The sequence of the *N. locustae* TBP gene provides evidence that microsporidia are related to fungi, in sharp contrast to the contention that microsporidia are among the earliest of eukaryotes. Fungal ancestry of the microsporidia significantly affects the interpretation of many of the unusual features they possess. While characteristics such as the apparent lack of mitochondria, peroxisomes, stacked golgi (and formerly spliceosomal introns), as well as the possession of tiny genomes were once thought to embody the ancestral situation, these features should instead be taken to indicate the highly derived nature of the microsporidia.

CHAPTER IV:

The TPI Gene of *Nosema locustae*

INTRODUCTION

As obligate intracellular parasites, microsporidia exhibit a unique mode for infecting cells—piercing host cells with an everted polar tube that is otherwise tightly coiled within the chitinous spore of the protist, and then passing the spore contents into the host via the tube. Phylogenetic information regarding the origin of the microsporidia is conflicting, with analysis of small subunit rRNA and elongation factor sequences indicating an early origin (Vossbrinck et al. 1987; Kamaishi et al. 1996a and 1996b), while analysis of tubulins, the largest subunit of RNA polymerase II and chaperonin sequences suggest a much later divergence (Keeling and Doolittle 1996; Edlind et al. 1996; Hirt et al. 1999; Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998). Resolution of the question of microsporidian origin will require the development of additional phylogenetic markers.

Triosephosphate isomerase (TPI) catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehyde 3-phosphate and is a component of both the glycolytic and Calvin cycle pathways. The gene sequence is generally well-conserved and has been widely sampled from a diverse sampling of eukaryotes and prokaryotes.

TPI has also been well documented in terms of intron evolution, as it was long ago postulated to provide a good example of introns separating units of protein structure, hence providing support for the early evolution of genes through exon shuffling ("introns-early") (Gilbert, Marchionni, and McKnight

1986; Straus and Gilbert 1985). TPI stood as a model for this theory for only a short time however, since as more TPI gene sequences were generated new intron positions were discovered, and protein structural units became further split, making it highly unlikely that the TPI gene (or for that matter, any gene) could have been assembled in this way (Logsdon et al. 1995; Stoltzfus et al. 1994; Logsdon 1998). In addition, the finding that TPI is likely of alpha-proteobacterial origin, possibly derived from the mitochondrial progenitor, makes it even more unlikely that present day introns are ancient relics that were involved in separating protein structural units involved in the construction of genes (Keeling and Doolittle 1997). There remain, however, five introns at identical positions in the TPI amino acid sequence from diverse animals, fungi and plants and likely predate the divergence of these groups (Gilbert, Marchionni, and McKnight 1986). As described in the previous chapter for the case of TBP, introns conserved at this phylogenetic depth are rare, and are termed "old" (as defined in Logsdon et al. 1995). The presence of such conserved intron positions in the TPI gene makes it a good candidate to possess a microsporidian intron.

In an attempt to examine both the phylogenetic placement of and intron evolution within the microsporidia, the full-length TPI gene from *Nosema locustae* was sequenced. The gene lacks spliceosomal introns and the sequence is extremely divergent. Consequently, the long branch possessed by the *N. locustae* TPI sequence in phylogenetic trees makes the position of the microsporidia in this analysis suspect. Furthermore, the overall ability of TPI to resolve well-known eukaryotic relationships is uncertain, reducing the value of TPI as a phylogenetic marker.

Figure 4-1 Alignment of TPI homologs
The conserved, central region of the *N. locustae* TPI sequence is shown aligned with a selection of eukaryotic and proteobacterial homologs. Numbering is based on the *N. locustae* conceptual translation product and gaps are indicated with a dash (-). Regions of ambiguity in the alignment were not included in the phylogenetic analysis. Taxon abbreviations: *N.lo, Nosema locustae; S.ce, Saccharomyces cerevisiae; S.po, Schizosaccharomyces pombe; A.ni, Aspergillus nidulans; C.ci, Coprinus cinereus; H.sa, Homo sapiens; M.mu, Mus musculus; D.me, Drosophila melanogaster; C.el, Caenorhabditis elegans; S.ma, Schistosoma mansoni; Z.ma, Zea mays; A.th, Arabidopsis thaliana; G.ve, Gracilaria verrucosa; P.in, Phytopthora infestans; P.fa, Plasmodium falciparum; T.cr, Trypanosoma cruzi; G.la, Giardia lamblia; E.hi, Entamoeba histolytica; R.et, Rhizobium etli; X.fl, Xanothobacter flavus; E.co, Escherichia coli.*

```
      81                                                               146
N.lo  MTLCAQDCSQFAQ-GAHTGEVGAYMLQEMGVKYVILGHSERRHI--LKEPDSVLHSKLRCCLEAN-LNVV
S.ce  VTVGAQNAYLKAS-GAFTGENSVDQIKDVGAKWVILGHSERRSY--FHEDDKFIADKTKFALGQG-VGVI
S.po  IGVGAQNVFDKKN-GAYTGENSAQSLIDAAITYTLTGHSERRTI--FKESDEFVADKTKFALEQG-LTVV
A.ni  IGVAAQNVFDKPN-GAFTGEISVQQLREANIDWTILGHSERRVI--LKETDEFIARKTKAAIEGG-LQVI
C.ci  VKVAAQNAYFKES-GAFTGEISPKQISDAGIPYVILGHSERRTL--FHETSEVVALKTRAALDNG-LKVI
H.sa  IAVAAQNCYKVTN-GAFTGEISPGMIKDCGATWVVLGHSERRHV--FGESDELIGQKVAHALAEG-LGVI
M.mu  IAVAAQNCYKVTN-GPFTGEISPGMIKDLGATWVVLGHSERRHV--FGESDELIGQKVSHALAEG-LGVI
D.me  LGLAGQNAYKVAK-GAFTGEISPAMLKDIGADWVILGHSERRAI--FGESDALIAEKAEHALAEG-LKVI
C.el  VLVAAQNCYKVPK-GAFTGEISPAMIKDLGLEWVILGHSERRHV--FGESDALIAEKTVHALEAG-IKVV
S.ma  IHVAAQNCYKVSK-GAFTGEISPAMIRDIGCDWVILGHSERRNI--FGESDELIAEKVQHALAEG-LSVI
Z.ma  FHVAAQNCWVKKG-GAFTGEVSAEMLVNLGVPWVILGHSERRAL--LGESNEFVGDKVAYALSQG-LKVI
A.th  FFVAAQNCWVKKG-GAFTGEVSAEMLVNLDIPWVILGHSERRAI--LNESSEFVGDKVAYALAQG-LKVI
G.ve  FDTSAQNAWISKG-GAFTGELDAAMVKDVGAEWVILGHSERRHIAQLKESDHTIAMKAAYALQHASLGVI
P.in  VRVSGQDVWKQGN-GAFTGETSAEMLKDLGAEYTLVGHSERR-E--KGETNEVVAKKAAYALEKG-LGVI
P.fa  FSTGIQNVSKFGN-GSYTGEVSAEIAKDLNIEYVIIGHFERRKY--FHETDEDVREKLQASLKNN-LKAV
T.cr  FQIAAQNAITRS--GAFTGEVSLQILKDYGISWVVLGHSERRLY--YGETNEIVAEKVAQACAAG-FHVI
G.la  LKIAAQNVYLEGN-GAWTGETSVEMLLDMGLSHVIIGHSERRRI--MGETNEQSAKKAKRALDKG-MTVI
E.hi  ILVSAENAWTKS--GAYTGEVHVGMLVDCQVPYVILGHSERRQI--FHESNEQVAEKVKVAIDAG-LKVI
R.et  VFTVAQDVSRFGNMGAYTGEVSAELLKDSQIEYVLIGHSERREY--FAESAAILNAKAQNALNAG-LKVI
X.fl  LMIGGQDCHPAES-GAHTGDISAEMLRDAGAVAVILGHSERRID--HQEGDAIVKAKVKAAWRAG-LLPV
E.co  IMLGAQNVDLNLS-GAFTGETSAAMLKDIGAQYIIIGHSERRTY--HKESDELIAKKFAVLKEQG-LTPV

                                                                       208
N.lo  LCVGETLEHRESGKTISVVKSQLSLLSNFR----ECEK-ISVAYEPVWAIGTGK--HP-EIKDVETVVEC
S.ce  LCIGETLEEKKAGKTLDVVERQLNAVLEEV----KDWTNVVVAYEPVWAIGTGLAATPEDAQDIHASIRK
S.po  ACIGETLADREANETITVVVRQLNAIADKV----QNWSKIVIAYEPVWAIGTGKTGTPEEAQEVHAEIRK
A.ni  FCIGETLEEREANKTIDVVTRQLNAAAKEL--SKEQWAKVVIAYEPVWAIGTGKVATTEQAQEVHSAIRK
C.ci  LCIGETLKEREEGRTAAVCEEQLSAVVKQL--KEEDWSNIVIAYEPVWAIGTGKVATTSQAQETHVDVRK
H.sa  ACIGEKLDEREAGITEKVVFEQTKVIADNV----KDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRG
M.mu  ACIGEKLDEREAGITEKVVFEQTKVIADNV----KDWSKVVLAYEPVWAIGTGKTATPQQAQEVHEKLRG
D.me  ACIGETLEEREAGKTNEVVARQMCAYAQKI----KDWKNVVVAYEPVWAIGTGKTATPDQAQEVHASLRQ
C.el  FCIGEKLEEREAGHTKDVNFRQLQAIVDKG----VSWENIVIAYEPVWAIGTGKTASGEQAQEVHEWIRA
S.ma  ACIGETLSERESNKTEEVCVRQLKAIANKI-KSADEWKRVVVAYEPVWAIGTGKVATPQQAQEVHNFLRK
Z.ma  ACVGETLEQREAGSTMDVVAAQTKAIAEKI----KDWSNVVVAYEPVWAIGTGKVATPAQAQEVHASLRD
A.th  ACVGETLEEREAGSTMDVVAAQTKAIADRV----TNWSNVVIAYEPVWAIGTGKVASPAQAQEVHDELRK
G.ve  YCIGELLEERESGQTIAVCERQLQALSDAI----SDWSDVVIAYEPVWAIGTGKVATPEQAEQVHEAVRA
P.in  ACIGETKDHREANQTVAYITEQLDAYAAEI----NDWTNVVIAYEPIWAIGTGLTASPEQAQEAHASIRA
P.fa  VCFGESLEQREQNKTIEVITKQVKAFVDLI----DNFDNVILAYEPLWAIGTGKTATPEQAQLVHKEIRK
T.cr  VCVGETNEEREAGRTAAVVLTQLAAVAQKL--SKEAWSRVVIAYEPVWAIGTGKVATPQQAQEVHELLRR
G.la  FCTGETLDERKANNTMEVNIAQLEALKKEIGESKKLWENVVIAYEPVWSIGTGVVATPEQAEEVHVGLRK
E.hi  ACIGETEAQRIANQTEEVVAAQLKAINNAI--SKEAWKNIILAYEPVWAIGTGKTATPDQAQEVHQYIRK
R.et  YCVGESLEQRESGQAEVVVLQQICDLASVV--TAEQWPHIVIAYEPIWAIGTGKTASPEDAQTMHAKIRE
X.fl  VCVGETLVERDAGEAAAVVTRQVRQSVPE----GATAVSLVIAYEPIWAIGTGRTPTTDDVAEVHGAIRH
E.co  LCIGETEAENEAGKTEEVCARQIDAVLKTQ--GAAAFEGAVIAYEPVWAIGTGKSATPAQAQAVHKFIRD
```

Figure 4-1 Alignment of TPI homologs

```
                                                             266
N.lo  --AEKALKEYGLRPRILYGGSVNKANCTSLARIKGLDGFLVGNASLT-TELFDIADV
S.ce  FLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLK-PEFVDIINS
S.po  WATNKLGASVAEGLRVIYGGSVTGGNCKEFLKFHDIDGFLVGGASLK-PEFPTNIVN
A.ni  WLKDAISAEAAENTRIIYGGSVSEKNCKDLAKEADIDGFLVGGASLK-PAFVDIVNA
C.ci  YLATAVSPKVASETRVIYGGSVNAANSKDLASQQDIDGFLVGGASLK-PEFVDIINA
H.sa  WLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLK-PEFVDIINA
M.mu  WLKSNVNDGVAQSTRIIYGGSVTGATCKELATPADVDGFLVGGASLK-PEFVDIINA
D.me  WLSDNISKEVSASLRIQYGGSVTAANAKELAKKPDIDGFLVGGASLK-PEFLDIINA
C.el  FLKEKVSPAVADATRIIYGGSVTADNAADVGKKPDIDGFLVGGASLK-PDFVKIINA
S.ma  WFKTNAPNGVDEKIRIIYGGSVTAANCKELAQQHDVDGFLVGGASLK-PEFTEICKA
Z.ma  WLKTNASPEVAESTRIIYGGSVTAANCKELAAQPDVDGFLVGGASLK-PEFIDIINA
A.th  WLAKNVSADVAATTRIIYGGSVNGGNCKELGGQADVDGFLVGGASLK-PEFIDIIKA
G.ve  WLANNVSPQVAASTRILYGGSVSPANCESLAKQPNIDGFLVGGASMK-PTFLEIVDS
P.in  WLKEKVSPEAAEQTRVIYGGSVGAKNAPELSQKEDIDGFLVGGASLK-PDFLQIINA
P.fa  IVKDTCGEKQANQIRILYGGSVNTENCSSLIQQEDIDGFLVGNASLK-ESFVDIIKS
T.cr  WVRSKLGTDIAAQLRILYGGSVTAKNARTLYQMRDINGFLVGGASLK-PEFVEIIEA
G.la  WFAEKVCAEGAQHIRIIYGGSANGSNCEKLGQCPNIDGFLVGGASLK-PEFTTMIDI
E.hi  WMTENISKEVAEATRIQYGGSVNPANCNELAKKADIDGFLVGGASLDAAKFKTIINS
R.et  GLTQITSH--GANMAILYGGSVKAENAVELAACPDINGAL
X.fl  ILA-ERFGAEANGIRILYGGSVKPDNAAALLATANVDGALVGGASLKAADFLAIARA
E.co  HIA-KVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKADAFAVIVKA
```

Figure 4-1 Alignment of TPI homologs

RESULTS

PCR amplification of TPI coding sequence from *Nosema locustae*

Four degenerate PCR primers were designed to highly conserved regions of the TPI sequence, and were used in all possible combinations in amplification reactions with *N. locustae* genomic DNA as template. Only one of the combinations (TPIF2-TPIR1) produced an amplification product of the expected size (~200 bp) which was subsequently cloned and then sequenced. Sequencing revealed that many of the clones contained identical insert sequence, which when compared to the database proved to be highly similar to TPI sequences.

Full length sequence of the *N. locustae* TPI gene

As the *N. locustae* PCR product only accounts for approximately one-quarter of the TPI coding sequence, it was necessary to retrieve the full-length sequence to determine intron presence or absence, and also to increase the number of informative sites for phylogenetic analysis with other TPI sequences. To this end, one of the positive PCR product clones (TPIF2R1-7) was arbitrarily chosen as a probe to screen the *N. locustae* genomic library.

Eleven primary PFUs were carried through to secondary screens and *in vivo* excision. Of these, restriction enzyme digestion with *Kpn*I and *Sac*I followed by Southern blotting with the TPIF2R1-7 probe indicated that there were three positive, independent genomic clones. As the TPI gene was expected to be relatively short, the entire TPI open reading frame was sequenced directly from the genomic clones, and subcloning was not undertaken. In fact, the entire TPI coding region was sequenced from a representative of each of the three independent genomic clones and ABI walking reactions completed double-

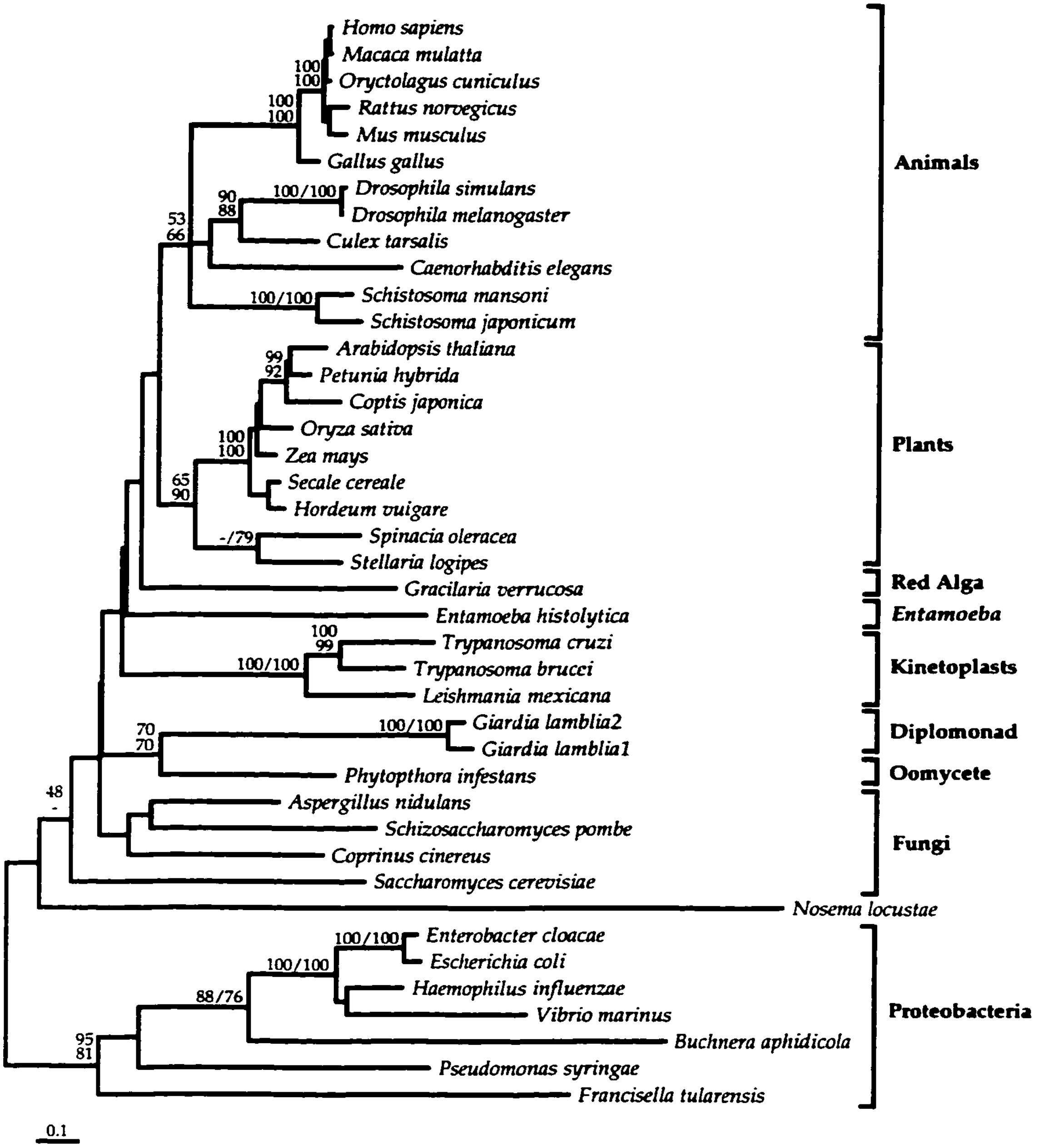


Figure 4-2 Fitch tree of TPI sequences
Bootstrap values based on 100 resampling replicates are indicated for major groups with >50% support. (Fitch results above those for neighbor-joining.) The position of the microsporidian is not supported in this tree or in the BioNJ tree (not shown).

stranded sequences of each. All clones contained an identical open reading frame of 798 bp, with an inferred protein length of 265 amino acids. The *N. locustae* TPI open reading frame is not interrupted by any introns, and is shown aligned with other TPI homologs in Figure 4-1. Based on the sequence similarity evident in Figure 4-1, the *N. locustae* sequence is a TPI homolog.

Sequencing results indicated that the *N. locustae* TPI gene is likely single copy as all the PCR products sequenced were identical, and all independent genomic clones contained the same TPI sequence along with identical flanking sequence. Single copy number and *N. locustae* origin of the TPI gene are consistent with results from a genomic Southern blot (data not shown).

Phylogenetic analysis of TPI sequences

An amino acid alignment of all available eukaryotic TPI sequences was constructed, excluding extremely partial sequences and almost identical sequences (for example, those of several insects). Proteobacterial sequences were also included in the alignment to provide an outgroup for phylogenetic analysis. The choice of proteobacteria as the outgroup stems from a previous phylogenetic study which indicated that eukaryotic TPI homologs are more closely related to those from proteobacteria than those from archaebacteria (Keeling and Doolittle 1997). In sum, the alignment included 274 characters and 44 taxa, nine of which were proteobacterial.

Maximum likelihood distances were calculated accounting for invariant sites with eight gamma categories to correct for site-to-site rate variation. Trees were constructed with BioNJ and Fitch. The *N. locustae* sequence is extremely divergent compared to other TPI homologs, and possesses the longest branch length in tree reconstructions. In fact, when the complete dataset was used both Fitch and BioNJ algorithms placed the microsporidian sequence with the next longest branches on the tree - with the only alpha-proteobacterial sequences available, *Rhizobium etli* and *Xanthobacter flavus*. Removing the *R. etli* and *X. flavus* sequences from the dataset resulted in two different placements of the microsporidian sequence. In the BioNJ tree, *N. locustae* branched as a sister to the only red algal sequence available, *Gracilaria verrucosa*, and in the Fitch tree the microsporidian sequence branched as the most basal eukaryote. Trees were constructed again, this time removing the next longest branch on the tree, that of *Plasmodium falciparum*. The Fitch tree produced is shown in Figure 4-2. All major relationships seen in this tree are the same as those seen with the full dataset, and also the dataset lacking only the alpha-proteobacterial sequences. As indicated in the figure, the microsporidian is again the most basal eukaryote. The BioNJ tree is identical except that *G. verrucosa* branches as a sister to the animals, along with the *N. locustae* sequence. This pattern mimics the position in the BioNJ tree constructed from the dataset lacking only the alpha-proteobacterial sequences.

Bootstrap support values for both NJ and Fitch were calculated based on the PAM250 matrix (PROTDIST) with 100 resampling replicates, and are shown in Figure 4-2. In general, the overall support for the topology of the tree is low; support is seen only for very closely related taxa and relationships among major groups are not resolved (<50%). The position of the microsporidian sequence is not supported for either of the two positions in which it falls. In the Fitch tree (Figure 4-2) its basal position is supported only by a 48% Fitch bootstrap value. This position is not consistent with the BioNJ tree, where the *N. locustae* sequence branches with *G. verracusa* with 43% support. It is likely that the long branch of the *N. locustae* sequence accounts for its attraction to the base of the tree or to the red algal sequence, which also possesses one of the longer branches in the tree.

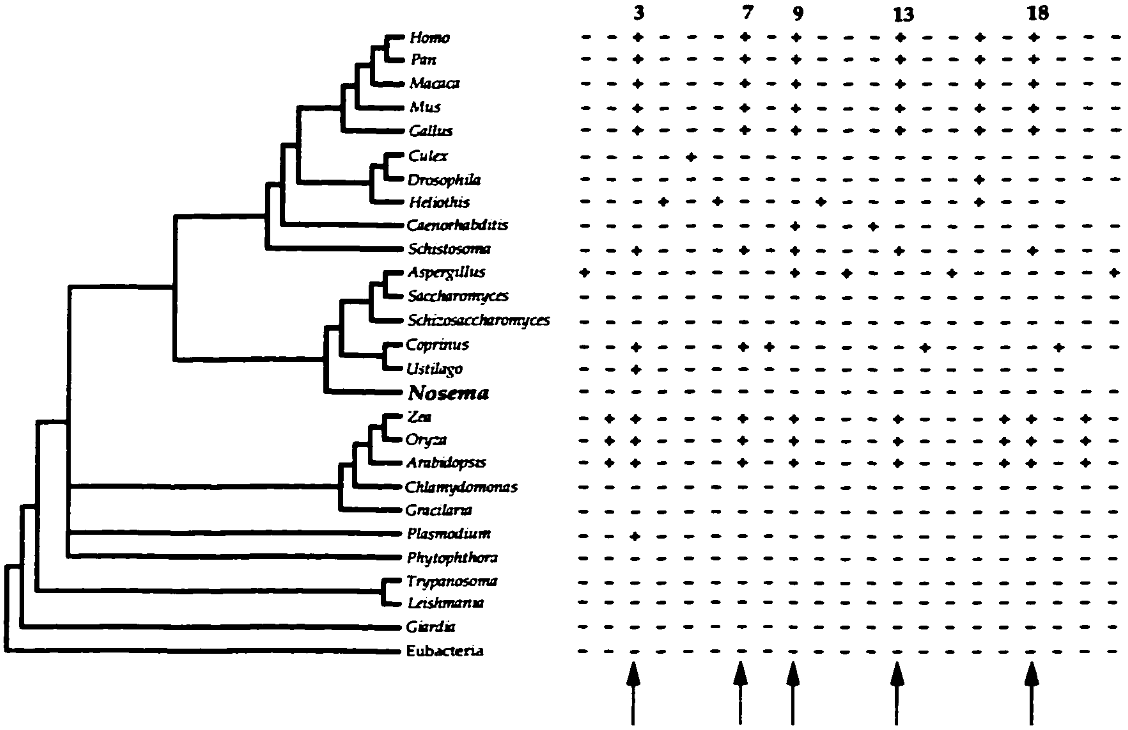


Figure 4-3 Phylogenetic distribution of introns among currently-known TPI genes

Presented as in Figure 3-4. Black arrows indicate potentially "old" introns. This figure is a revised version of that appearing in Logsdon et al. 1995.

Spliceosomal intron loss in the microsporidian TPI gene

As outlined in Figure 4-3, and described in Logsdon et al. 1995, the TPI gene contains at least twenty-one different intron positions, five of which can be classified as "old." Despite the presence of such old introns in known TPI genes, the *N. locustae* TPI gene is intron-lacking. In Figure 4-3, the microsporidia are shown as relatives of fungi, since current evidence (from tubulin, RPB1, TBP and HSP70 phylogenies) indicate that this is the likeliest scenario. Assuming a fungal relationship of the microsporidia, the lack of introns in TPI is a result of intron loss.

DISCUSSION

Determining the origin of the microsporidia will require the development of more phylogenetic markers; however, it seems unlikely that TPI has the resolving power to be particularly useful in addressing this question. The overall topology of the tree is inconsistent in different analyses and is only weakly supported; even the nodes uniting the undisputed groups of animals, plants, and fungi are not extremely well supported. In addition, the unrelated diplomonad *Giardia* and the heterokont *Phytopthora* branch together, while other well documented relationships like the sisterhood of animals and fungi (Baldauf and Palmer 1993) are not retrieved in this analysis. In fact, the fungal sequences branch at the base of the tree after the microsporidian sequence, but not with any support. The focus of this study, the *N. locustae* sequence, does not branch in the same place in the BioNJ and Fitch analyses, and furthermore the placement of the microsporidian sequence in both trees is unsupported. The extremely divergent nature of the *N. locustae* sequence (it possesses the longest branch on the tree!) would cause any phylogenetic placement it possessed to be highly suspect. In sum, the highly divergent microsporidian TPI sequence, coupled with the overall weakness of the TPI tree topology, would make drawing any conclusions about the origin of the microsporidia from this analysis unwarranted.

The *N. locustae* TPI sequence is completely intron-free even though there are five introns in TPI that likely predate the divergence of animals, plants and fungi (Gilbert, Marchionni and McKnight 1986; Logsdon et al. 1995). Originally, such a lack of introns in a microsporidian sequence would be ascribed to the ancient origin of these protists, once believed to primitively lack spliceosomal introns, along with several other "eukaryotic" features like mitochondria and peroxisomes. However, recent evidence has indicated that microsporidia not only possess genes of mitochondrial origin (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998), but a single putative spliceosomal intron has also been found in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi* (Biderre, Méténier and Vivarès 1998). Additional phylogenetic information (from tubulins, RPB1, HSP70, TBP) also suggests that microsporidia are not primitive, but instead are related to fungi. As a fungal relationship seems most likely, the intron-void TPI in *N. locustae* seems less likely to represent a primitive condition, but rather more likely represents spliceosomal intron loss (Figure 4-3). As shown in Figure 4-3, the microsporidian TPI is not the only TPI gene to completely lack introns, so there is a precedent for such a loss to occur. Intron loss in the microsporidia may exemplify genome size reduction; microsporidian genomes are quite small, at the extreme the *E. cuniculi* genome is only 2.9 Mb (Biderre et al. 1995). The loss of non-coding elements like introns may be one way to reduce genome content, and likely occurred recently in the course of microsporidian evolution.

CHAPTER V:

Phylogenetic Analysis of the Second Largest Subunit of *Nosema locustae* RNA Polymerases

INTRODUCTION

Phylogenetic analysis of alpha- and beta-tubulin genes suggested in 1996 that microsporidia are not ancient eukaryotes, but are instead close relatives of fungi (Keeling and Doolittle 1996; Edlind et al. 1996). This was the first molecular evidence supporting such a relationship, and in the last three years molecular phylogenies of HSP 70 (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Peyretaillade et al. 1998), the largest subunit of RNA polymerase II (RPB1) (Hirt et al. 1999), and TATA box binding protein (see Chapter 3) have provided additional evidence for a relationship between microsporidia and fungi. Support for the microsporidia - fungi relationship varies among molecules and, bearing in mind that comparison is hindered somewhat by differences in phylogenetic method, the greatest bootstrap support is seen with alpha-tubulin, followed by RPB1 and beta-tubulin, with TBP and HSP70 providing weak support. Support values aside, the fact that multiple datasets indicate the same alternative is significant. In addition, several molecular datasets that have indicated an ancient origin for the microsporidia (EF-2, EF-1alpha, and large subunit rRNA) have recently been re-analyzed, and the strength seen for the deep position of the microsporidia has been attributed to phylogenetic artifacts arising from compositional biases and accelerated rates of evolution for the microsporidian sequences (Hirt et al. 1999; Peyretaillade et al. 1998). In the case of the elongation factors, saturation of the microsporidian sequences (particularly in EF-1alpha) make their phylogenetic position unreliable (Hirt et al. 1999; Roger et al. 1999). Furthermore, when more rigourous attempts were made to account for base compositional biases and rate variation, a relationship between microsporidia and fungi in elongation factor phylogenies can not be excluded (Hirt et al. 1999). Similarly, a new analysis of LSU sequences employing a method that purportedly uses a more realistic model of molecular evolution that accounts for site-to-site rate variation does not support a basal position for the microsporidia, and favours a divergence within crown eukaryotes (though in this case, not with the fungi) (Peyretaillade et al. 1998).

Although there appears to be a relationship between microsporidia and fungi, it is unclear whether microsporidia arose from a fungus or from a protozoan fungal ancestor. In tubulin phylogenies (which include a selection of both ascomycete and basidiomycete fungi), microsporidia branch within the fungi (Keeling and Doolittle 1996; Edlind et al. 1996). A different situation is seen with analyses of RPB1, HSP70 and TBP where the microsporidia branch as a sister-group to the fungi (Germot, Philippe and Le Guyader 1997; Hirt et al. 1998; Chapter 3). However, the fungal representation in these latter cases is limited to two taxa for RPB1 and HSP70 and five for TBP, and in all three cases all are ascomycete fungi. In a collaborative effort to try to see if the microsporidia are related to a specific fungus, the second-largest subunit of RNA polymerase II (RPB2) was sequenced from two microsporidia and representatives of each of the four fungal divisions.

The second-largest subunit of eukaryotic RNA polymerases is involved in binding the substrate nucleotide during transcription and is most similar to the β subunit of the prokaryotic RNA polymerase, with nine regions highly conserved between the two (Archambault and Friesen 1993; Sweetser, Nonet and Young 1987). Degenerate primers designed to these domains are capable of amplifying the second-largest subunit of all three polymerases at one time. Such is the case described here for the microsporidian *Nosema locustae*. Partial coding sequences for the second-largest subunits of RNA polymerase I, II and III (RPA2, RPB2 and RPC2) were amplified and were confirmed as such by phylogeny. The RPB2 subunit was then focused upon since of the three, it possesses the greatest sampling. The nearly full-length sequence of the *N. locustae* RPB2 gene was determined and included in a phylogenetic analysis that also included sequences from the microsporidian *Vairimorpha necatrix*, and representatives of all four fungal divisions: ascomycetes, basidiomycetes, zygomycetes and chytrids.

Phylogenetic analysis of the RPB2 sequences resolved the expected relationships among the fungal divisions, but indicated only a weak microsporidia-fungi relationship, placing the microsporidia as a sister to the fungi. A basal position for the microsporidia was more strongly supported by bootstrapping and, although Kishino-Hasegawa tests indicated that the best tree included a relationship between the microsporidia and the fungi, a basal placement of the microsporidia was not significantly worse. Phylogenetic analyses were also carried out including only the microsporidian and fungal sequences, and the microsporidia showed no particular affinity for any fungal division. Altogether, although RPB2 phylogeny does indicate a relationship between the microsporidia and the fungi, the weakness of the relationship precludes drawing any reliable conclusions about whether the microsporidia arose from within the fungi, or from a protist ancestor of the fungi.

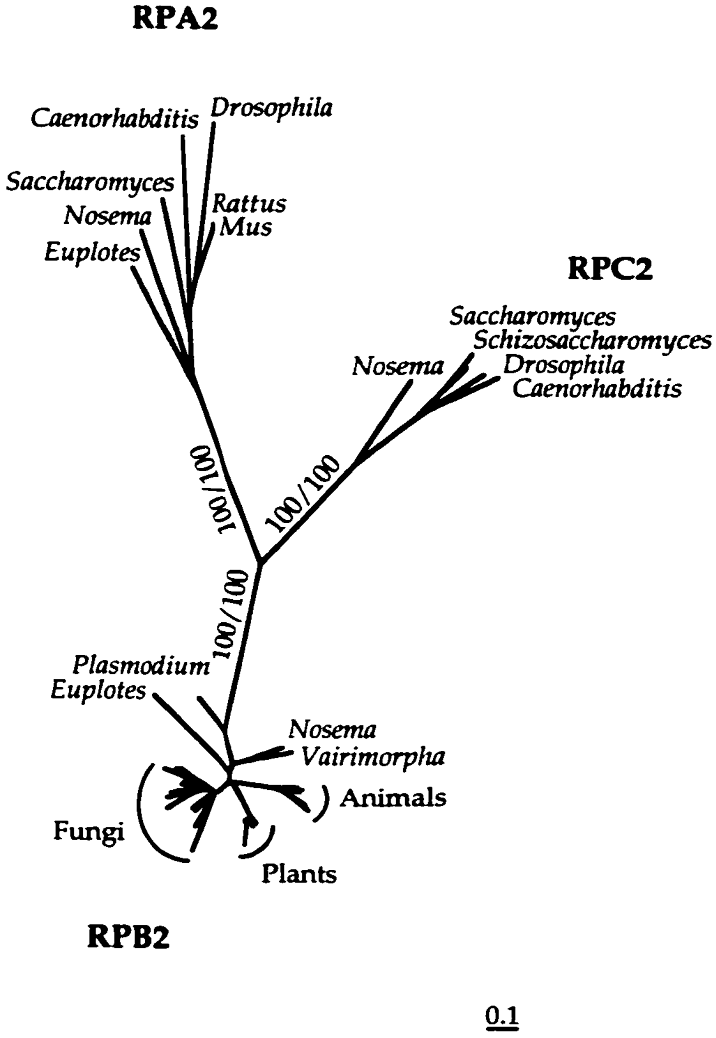


Figure 5-1 Fitch tree of RPA2, RPB2 and RPC2 partial sequences
Each microsporidian partial sequence is the second largest subunit of a different RNA polymerase, as both neighbor-joining and Fitch bootstraps support the monophyly of each polymerase (including the *N. locustae* sequences) with 100% support.

RESULTS

Partial coding sequences of the second-largest subunits of RNA Polymerases I, II and III

Primers designed to amplify the second-largest subunit of RNA Polymerase II were generously provided by R.P. Hirt and T.M. Embley and were used in a standard PCR reaction with *N. locustae* genomic DNA as template. A strong amplification product of the expected size (~1100 bp) was cloned and several copies were sequenced, revealing that there were three major amplification products of a nearly identical length. Comparing these sequences against sequences deposited in Genbank showed that the three products bore strongest similarities to the second-largest subunit of RNA polymerases I, II and III (RPA2, RPB2, and RPC2 genes, respectively). Multiple clones of each were sequenced on both strands.

To confirm that these partial sequences were indeed three different polymerases, a phylogenetic analysis including sequences representing all three polymerases was performed. The inferred amino acid sequences of the *N. locustae* PCR products were aligned with the homologous region of all available sequences of the second-largest subunit of all three eukaryotic RNA polymerases. Excluding very partial, and almost identical sequences, the alignment contained 43 taxa and 345 characters. Maximum likelihood distances were calculated with the JTT substitution matrix with site-to-site rate variation corrected with eight gamma categories and one invariable category. Trees based on these distances were constructed with BioNJ and Fitch. The Fitch tree is shown in Figure 5-1, with both NJ and Fitch bootstraps indicated (based on one hundred resampling replicates and distances calculated with the PAM250 substitution matrix). The BioNJ tree is identical except that the microsporidian RPB2 sequences branch

Figure 5-2 Alignment of partial RPA2 sequences
The partial coding sequence of the RPA2 gene of *N. locustae* is shown aligned with the homologous region of all available eukaryotic RPA2 sequences. Regions of ambiguity in the alignment were not included in the phylogenetic analysis. Taxon abbreviations: *N.lo, Nosema locustae; S.ce, Saccharomyces cerevisiae; C.el, Caenorhabditis elegans; D.me, Drosophila melanogaster; M.mu, Mus musculus; R.no, Rattus norvegicus; E.oc, Euplotes octocarinatus.*

```
N.lo  PFPDHNQSPRNMHQCQMAKQSIGVPFLNTKYRVDNKSYHLVYTQDPVVKTHFYDMYNLSSYPIGI
S.ce  PFSDFNQSPRNMYQCQMGKQTMGTPGVALCHRSDNKLYRLQTGQTPIVKANLYDDYGMDNFPNGF
C.el  PFPDHNQSPRNVYQCQMGKQTMGTAVHAWHSRADNKMYRLQFPQQPMLKLEAYEKYEMDEYPLGT
D.me  PMPDYNQSPRNMYQCQMGKQTMGTPCLNWPKQAANKLYRLQTPGTPLFRPVHYDIIQLDDFAMGT
M.mu  PFSDHNQSPRNMYQCQMGKQTMGFPLLTYQNRSDNKLYRLQTPQSPLVRPCMYDFYDMDNYPIGT
R.no  PFSDHNQSPRNMYQCQMGKQTMGFPLLTYQDRSDNKLYRLQTPQSPLVRPCMYDHYDMDNYPIGT
E.oc  VFAEYNQSPRNMYQCQMAKQTMGTPYHNHQFRTDNKIYRLLFPHRPIVKTRTQVDFDIEEYPSGT


N.lo  NAVVAVLSYTAYDMEDAMVINRSSIDRGFFKGEIYKTETVNLEKDCHVMDMPDIGDIID-----T
S.ce  NAVVAVISYTGYDMDDAMIINKSADERGFGYGTMYKTEKVDLALN---RNRGDPITQHFGFGNDE
C.el  NACVAVISYTGYDMEDAMTINKASYQRGFAHGTVIKVERINLVTE---RER----KTIFYRNPRE
D.me  NAIVAVISYTGYDMEDAMIINKAAYERGFAYGSIYKTKFLTLDKK---S-S------YFARHPHM
M.mu  NAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEK---FKQGE-DNLVFGVKPGD
R.no  NAIVAVISYTGYDMEDAMIVNKASWERGFAHGSVYKSEFIDLSEK---FKQGD-DSLVFGVKPGD
E.oc  NAVVAVISYTGYDLEDAMIINKSSYERGFGHGVVYKSYTHDLNESNSQSTRGIKSSVRYKFLNNV


N.lo  -----DDVILRYKDSVGDMHAIRYKGLEAGCVD----------------------SVRLFSNQS
S.ce  WPK----EWLEKLDEDGLPYIGTYVEEGDPICAYFDDTLNKTKIKTYHSSEPAYIEEVNLIGDES
C.el  E--------IKTVGPDGLPIPGRRYFLDEVYYVTFNMETGDFRTHKFHYAEPAYCGLVRIVEQGE
D.me  PEL------IKHLDTDGLPHPGSKLSYGSPLYCYFDGEVATYKVVKMDEKEDCIVESIRQLGSFD
M.mu  PRV------MQKLDDDGLPSIGAKLEYGDPYYSYLNLNTGEGFVVYYKSKENCVVDNIKVCSNDM
R.no  PRV------MQKLDNDGLPFIGAKLEFGDPYYGYLNLNTGEGFVVYYKSKENCVVDNIKVCSNDT
E.oc  SQKDKSKIKLENIDPDGLPKIGSQLTKGKPELCIFDTLKRGAKLSKFKDSEKARIETVRVCGNDD


N.lo  VSP-LINTAVVRPRIKRDPTIGDKFCSRHGQKGVCSMRWPNIDMPFSESGLVPDIIINPHAFPSR
S.ce  NKFQELQTVSIKYRIRRTPQIGDKFSSRHGQKGVCSRKWPTIDMPFSETGIQPDIIINPHAFPSR
C.el  GDS-GAKHALIQWRIERNPIIGDKFASRHGQKGINSFLWPVESLPFSETGMVPDIIFNPHGFPSR
D.me  LS--PTKMVAITLRVPRPATIGDKFASRAGQKGICSQKYPAEDLPFTESGLIPDIVFNPHGFPSR
M.mu  GSG-KFKCICITVRIPRNPTIGDKFASRHGQKGILSRLWPAEDMPFTESGMMPDILFNPHGFPSR
R.no  GSG-KFKCVCVTVRVPRNPTIGDKFASRHGQKGILSRLWPAEDMPFTESGMMPDILFNPHGFPSR
E.oc  KNP-DNLSIGYTIRYSRIPVIGDKFSSRHGQKGVLSVLWPQVDMPFTENGITPDLIINPHAFPSR


N.lo  MTIGMLLESIAGKSGCLLGKKQDSTPFQQCCSDRQMHQSKEDMRKAFCEELQKHGFNYYGNEPMY
S.ce  MTIGMFVESLAGKAGALHGIAQDSTPWI--------FNEDDTPADYFGEQLAKAGYNYHGNEPMY
C.el  MTIGMMIESMAGKAAATHGENYDASPFV--------FNEDNTAINHFGELLTKAGYNYYGNETFY
D.me  MTIAMMIETMAGKGAAIHGNVYDATPFR--------FSEENTAIDYFGKMLEAGGYNYYGTERLY
M.mu  MTIGMLIESMAGKSAALHGLCHDATPFI--------FSEENSALEYFGEMLKAAGYNFYGTERLY
R.no  MTIGMLIESMAGKSAALHGLCHDATPFI--------FSEENSALEYFGEMLKAAGYNFYGTERLY
E.oc  MTMGMLIQSMAAKSGSLRGEFKTVETFQ--------RYDDNDIVGHFGKELLDKGFNYHGNELMY


N.lo  SGITGQELRTDIYLGVVYYLRLRHMVNDKFQVRSTGPVQPQTRQPIKGRSKQGGVRLGEMESDC
S.ce  SGATGEELRADIYVGVVYYQRLRHMVNDKFQVRSTGPVNSLTMQPVKGRKRHGGIRVGEMERDA
C.el  SGVDGRQMEMQIFFGIVYYQRLRHMIADKFQVRATGPIDPITHQPVKGRKKGGGIRFGEMERDA
D.me  SGVDGREMTADIFFGVVHYQRLRHMVFDKWQVRSTGAVEARTHQPIKGRKRGGGVRFGEMERDA
M.mu  SGISGMELEADIFIGVVYYQRLRHMVSDKFQVRTTGARDKVTNQPLGGRNVQGGIRFGEMERDA
R.no  SGISGMELEADIFIGVVYYQRLRHMVSDKFQVRTTGARDKVTNQPIGGRNVQGGIRFGEMERDA
E.oc  SGIFGTPLKADIFIGVVYYQRLRHMVSDKSQARGTGPIDILTHQPVKGRKKGGGIRFGEMERDS
```

Figure 5-2 Alignment of partial RPA2 sequences

Figure 5-3 Alignment of partial RPC2 sequences
The partial coding sequence of the RPC2 gene of *N. locustae* is shown aligned with the homologous region of all available eukaryotic RPC2 sequences. Regions of ambiguity in the alignment were not included in the phylogenetic analysis. Taxon abbreviations: *N.lo, Nosema locustae; S.ce, Saccharomyces cerevisiae; S.po, Schizosaccharomyces pombe; C.el, Caenorhabditis elegans; D.me, Drosophila melanogaster.*

```
N.lo  PFPDHNQSPRNTYQCAMGKQAMGFVALNQYRRFDTPLNLLVYPQKPLVSTKITDIVNFEKLPAGQ
S.ce  PYPHHNQSPRNTYQCAMGKQAIGAIAYNQFKRIDTLLYLMTYPQQPMVKTKTIELIDYDKLPAGQ
S.po  PYPHHNQSPRNTYQCAMGKQAIGAIAYNQLQRIDTLLYLMVYPQQPMVKTKTIELIGYDKLPAGQ
C.el  PYPHHNQSPRNTYQCAMGKQAMGTIAYNQQKRIDSIMYLLCYPQRPLVKSKTIELTNFEKLPAGA
D.me  PYPHHNQSPRNTYQCAMGKQAMGMIGYNHNNRIDSLMYNLVYPHAPMVKSKTIELTNFDKLPAGQ

N.lo  NGMVAVMAYAGYDIEDALIINKGALDRGFGRCEVYRSYKHSLKRYSNGMCDRISGSKHS------
S.ce  NATVAVMSYSGYDIEDALVLNKSSIDRGFGRCETRRKTTTVLKRYANHTQDIIGGMRVD-ENGDP
S.po  NATVAIMSYSGYDIEDALVLNKSSIDRGFGRCQVFHKHSVIVRKYPNGTHDRIGDPQRDPETGEV
C.el  NGIIAVMSYSGYDIEDALVLNKASLDRGYGRCLVYKHVKGTAKKYPNQTFDRLLGPALDPNTRKP
D.me  NATVAVMSYSGYDIEDALILNKASIDRGYGRCLVYKNSKCTVKRYANQTFDRIMGPMKDALTNKV

N.lo  ---------DGVRGPGERVFDGDFFVYKESPTEDADFKF----------SGELYKNL--------
S.ce  IWQHQSLGPDGLGEVGMKVQSGQIYINKSVPTNSADAPN--------PNNVNVQTQYREAPVIYR
S.po  VWKHGVVEDDGLAGVGCRVQPGQIYVNKQTPTNALDNSIT------LGHTQTVESGYKATPMTYK
C.el  IFKHKNLDQEGIVFAGARIMPKQTIINKHMPVVSGESGPGASASANTIGIAGRQVDYKDVSITYK
D.me  IFKHDVLDTDGIVAPGEQVQNKQIMINKEMPAVT-------SMNPLQGQS--AQVPYTAVPISYK

N.lo  --SPSVVDKVIVARSGEDQFMVKTCLRQVRQPEVGDKFSSGHGQKGVIGLVVPETDLPFSEDGQV
S.ce  GPEPSHIDQVMMSVSDNDQALIKVLLRQNRRPELGDKFSSRHGQKGVCGIIVKQEDMPFNDQGIV
S.po  APEPGYIDKVMLTTTDSDQTLIKVLMRQTRRPELGDKFSSRHGQKGVCGVIVQQEDMPFNDQGIC
C.el  TPTPSYAERVLLTYNEDEAHLFKVLLRQTRRPELGDKFSSRHGQKGVCGLIAQQEDMPFNDLGMV
D.me  GPEPSYIERVMVSANAEEDFLIKILLRQTRIP-RGDKFSSRHGQKGVTGLIVEQEDMPFNDFGIC

N.lo  PDIIMNPHGFPSRMTVGKIIELISGKAAVFTGRQSDATVFREDITEDLSHILIKHGYSYCGKDTF
S.ce  PDIIMNPHGFPSRMTVGKMIELISGKAGVLNGTLEYGTCFGGSKLEDMSKILVDQGFNYSGKDML
S.po  PDIIMNPHGFPSRMTVGKMIELLSGKVGVLRGTLEYGTCFGGTKVEDASRILVEHGYNYSGKDML
C.el  PDMIMNPHGYPSRMTVGKLMELLSGKAGVVNGTYHYGTAFGGDQVKDVCEELAACGYNYMGKDML
D.me  PDMIMNPHGFPSRMTVGKTLELLGGKAGLLEGKFHYGTAFGGSKVEDIQAELERHGFNYVGKDFF

N.lo  TNGITGDVYEGYVFFGPVYYQRLKHMVADKMHARSRGPRAMLTRQPTEGKSREGGLRLGEMERDC
S.ce  YSGITGECLQAYIFFGPIYYQKLKHMVLDKMHARARGPRAVLTRQPTEGRSRDGGLRLGEMERDC
S.po  TSGITGETLEAYIFMGPIYYQKLKHMVMDKMHARARGPRAVLTRQPTEGRSRDGGLRLGEMERDC
C.el  TSGITGQPLSAYIYFGPIYYQKLKHMVLDKMHARARGPRAALTRQPTEGRSREGGLRLGEMERDC
D.me  YSGITGTPLEAYIYSGPVYYQKLKHMVQDKMHARARGPKAVLTRQPTQGRSREGGLRLGEMERDC
```

Figure 5-3 Alignment of partial RPC2 sequences

after the *Euplotes* sequence in the BioNJ tree. Clearly, the three *N. locustae* partial sequences correspond to the second-largest subunit of the three RNA polymerases, as the monophyly of each polymerase (including the *N. locustae* sequence) is supported by 100% bootstrap support. The similarity among subunits of the same polymerase is also evident from aligning their sequences, and the *N. locustae* RPA2 and RPC2 partial coding sequences are shown aligned with available homologs in Figures 5-2 and 5-3, respectively.

Since there is the greatest representation of sequences available for RPB2, it was chosen for further analysis. This diversity stems from the inclusion of unpublished fungal sequences provided collaboratively by B. Hall, and the unpublished microsporidian *Vairimorpha necatrix* sequence provided collaboratively by R.P. Hirt and T.M. Embley.

Microsporidian RPB2 sequences

As the *N. locustae* PCR product included only approximately one-quarter of the full-length sequence, the *N. locustae* genomic library was screened using the PCR product as a probe and an almost full-length sequence was retrieved. (Library screening and genomic clone sequencing were the collaborative contributions of J.M. Logsdon.) Aligning the *N. locustae* sequence with that of *V. necatrix* and all other available eukaryotic RPB2 sequences revealed that the microsporidian sequences possess the motifs characteristic of RPB2. These include the conserved regions shared by eukaryotic RPB2 and prokaryotic RNA polymerase subunit β, which were identified early-on and designated as "homology domains" (Sweetser, Nonet and Young 1987). There are nine such domains (A through I) and the available microsporidian sequences allow for comparison of C through I.

Figure 5-4 *N. locustae* possesses the conserved "homology domains" of RPB2

Alignment of the conserved domains C through I comparing the sequences of *N. locustae* and *V. necatrix* with those from a selection of eukaryotes and with subunit β of *E. coli* RNA polymerase. Numbering reflects the position of the domain in the full-length *Saccharomyces cerevisiae* sequence. Taxon abbreviations: *N.lo, Nosema locustae; V.ne, Vairimorpha necatrix; S.ce, Saccharomyces cerevisiae; A.bi, Agaricus bisporus; M.hi, Mucor hiemalis; A.ma, Allomyces macrogynus; H.sa, Homo sapiens; C.el, Caenorhabditis elegans; P.fa, Plasmodium falciparum; E.oc, Euplotes octocarinatus; E.co, Escherichia coli.*

Domain C (391-406)

N.lo	NRKGEDDRDHYGKKRM
V.ne	GRKKEDDRDHYGKKRM
S.ce	DRKDQDDRDHFGKKRL
A.bi	ERRELDDRDHFGKKRL
M.hi	ERRELDDRDHYGKKRM
A.ma	GRRQVDDRDHFGKKRL
H.sa	GRRELDDRDHYGNKRL
C.el	GRRELDDRDHIGNKRL
P.fa	GRIKEDDRDHFGKKRL
E.oc	GRIKEDDRDHYGKKRL
E.co	GKGEVDDIDHLGNRRI

Domain D (512-541)

N.lo	RQLHNTHWGMICPAETPEGQSCGLVKNLSL
V.ne	RQLHNTHWGMICPAETPEGQACGLVKNLSL
S.ce	RQLHNTHWGLVCPAETPEGQACGLVKNLSL
A.bi	RQLHNTHWGMVCPAETPEGQACGLVKNLAL
M.hi	RQLHNTHWGLVCPAETPEGQACGLVKNLAL
A.ma	RQLHNTHWGMVCPAETPEGQACGLVKNLAL
H.sa	RQLHNTLWGMVCPAETPEGHAVGLVKNLAL
C.el	RQLHNTQWGMVCPAETPEGQAVGLVKNLAL
P.fa	RQLHNTHWGMICPFETPEGQSVGLVKNLSL
E.oc	RQLHNTHWGMVCPAETPEGQACGLVKNLSL
E.co	RDVHPTHYGRVCPIETPEGPNIGLINSLSV

Domain E (748-766)

N.lo	ILGICASMIPFPDHNQSPR
V.ne	ILGICASVFPFPDHNQSPR
S.ce	ILGVAASIIPFPDHNQSPR
A.bi	ILGICASIIPFPDHNQSPR
M.hi	ILGICASIIPFPDHNQSPR
A.ma	ILGICASIIPFPDHNQSPR
H.sa	ILGVCASIIPFPDHNQSPR
C.el	ILGVCASIIPFPDHNQSPR
P.fa	ILGVCASIIPFSDHNQSPR
E.oc	ILGVCASIIPFPDHNQSPR
E.co	VVSVGASLIPFLEHDDANR

Domain F (816-851)

N.lo	ELPSGQNAIVTIACYTGYNQEDSIIMNQSAIDRGLF
V.ne	ELPPGQNAIVAIASYTGYNQEDSIIMNQSAIDRGLF
S.ce	ELPAGQNAIVAIACYSGYNQEDSMIMNQSSIDRGLF
A.bi	ELPAGQNAIVAILCYSGYNQEDSVIMNQSSIDRGLF
M.hi	ELPAGQNAIVAILCYSGYNQEDSVIMNQSSIDRGLF
A.ma	ELPAGQNAVVAIAVYSGYNQEDSIIMNQSSIDRGLF
H.sa	ELPAGINSIVAIASYTGYNQEDSVIMNRSAVDRGFF
C.el	ELPAGINAIVAILSYSGYNQEDSVIMNNSAIDRGLF
P.fa	ELPAGINAIVAIMCYTGYNQEDSLIMNQSSIDRGLF
E.oc	ELPPGCDSIVAITCYTGYNQVDSVIMSQAPIDRGCF
E.co	ELALGQNMRVAFMPWNGYNFEDSILVSERVVQEDRF

Domain G (887-917)

N.lo	NLNYEKLDSDGLINSQVRVTGVDILMGKAVP
V.ne	NLNYSKLDEDGIIPIGVRVTGDDVLIGKVSP
S.ce	HGTYDKLDDDGLIAPGVRVSGEDVIIGKTTP
A.bi	HGTYDKLEDDGLIAPGTGVRGEDIIIGKTAP
M.hi	HGTYEKLEDDGLIAPGTRVSGDDIIIGKTAP
A.ma	HGSYDKIEEDGLVAPGTRVTGDDIIIGKTAP
H.sa	HAIYDKLDDDGLIAPGVRVSGDDVIIGKTVT
C.el	HSLYDKLDEDGIISPGMRVSGDDVIIGKTVA
P.fa	RGDYTKLDDDGLIAPGIRVLGDDIIIGKVSP
E.oc	NGDYTKLDIDGLIFPGKNVLGDDIIIGKTAL
E.co	EAALSKLDESGIVYIGAEVTGGDILVGKVTP

Figure 5-4 *N. locustae* possesses the conserved "homology domains" of RPB2

Domain H (961-1032)

```
                                                           .
N.lo  YKFAKVRVRSNRIPQTGDKFASRHGQKGTIGITLRQEDMPFTKEGIVPDIIINPHAIPNRMTIGHLIEALLG
V.ne  YKFAKVKVRTSRIPQMGDKFASRHAQKGTMGISLRQEDMPFTSEGIVPDIIINPHAIPSRMTIGHLIECLLG
S.ce  LKFVKVRVRTTKIPQIGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLLS
A.bi  QKFVKVRVRATRIPQIGDKFASCHGQKGTVGITYRQEDMPFTAEGIVPDLIINPHAIPSRMTIGHLVECLLS
M.hi  LKFVKVRVRSTRVPQMGDKFASRHGQKGTIGMTYRQEDFPFSAEGITPDLIINPHAIPSRMTIGHMIECLLG
A.ma  YKFVKVRVRSTRVPQMGDKFASRHGQKGTVGMTYRQEDMPFSADGVTPDLIVNPHAIPSRMTIGHLVECLLS
H.sa  YKFCKIRVRSVRIPQIGDKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQG
C.el  NKFVKIRMRSVRLPQIGDKFASRHGQKGTMGIMYRQEDMPFTAEGLTPDIIINPHAVPSRMTIGHLIECLQG
P.fa  NKFAKVKVRSVRIPQIGDKFASRHGQKGTIGITYRTEDMPFSSLGIFPDIIMNPHAVPSRMTIGHLVECLTG
E.oc  DSFTKVKMRAIRIPQIGDKFASRHGQKGTVGMTYRQEDIPFTQEGIIPDIIVNPHAIPSRMTIGHLIECLAS
E.co  LKIVKVYLAVKRRIQPGDKMAGRHGNKGVISKINPIEDMPYDENGTPVDIVLNPLGVPSRMNIGQILETHLG
```

Domain I (1058-1156)

```
N.lo  KLESLGYQKRGLEVMYSGFTGRKLEAQVFVGPTYYQRLKHMVEDKIHARAR
V.ne  KLKEFNYQQRGFEVMYNGMTGHKLRAQIFIGPTYYQRLKHMVEDKIHARAK
S.ce  LLREHGYQSRGFEVMYNGHTGKKLMAQIFFGPTYYQRLRHMVDDKIHARAR
A.bi  FLRQKGYQSRGLEVMYHGHTGRKLQAQIYLGPTYYQRLKHMVDDKIHSRAR
M.hi  ALQAQGYQSRGFEVMYNGFTGRKLNVQVFLGPTYYQRLKHMVDDKIH
A.ma  RLRSCGYQQRGFEIMYNGHTGKKLAAQVFFGPTYYQRLKHMVDDKIHSRAR
H.sa  LLSDYGYHLRGNEVLYNGFTGRKITSQIFIGPTYYQRLKHMVDDKIHSRAR
C.el  LLCEYGYHLRGNEVMYNGHTGKKLTTQIFFGPTYYQRLKHMVDDKIHSRAR
P.fa  KLHNLGYEKYGNEMLYNGHNGRMLKSKIFIGPTYYQRLKHMVEDKIHARSR
E.oc  DLHRLGYQKRGNEVMYDGWTGRKMDTMIFLGPTYYQRLKHMVDDKIHSRSR
E.co  LLKLGDLPTSGQIRLYDGRTGEQFERPVTVGYMYMLKLNHLVDDKMHARST

N.lo  GPVQIMTRQPVEGRSREGGLRFGEMERDCIISHGASSFLKERLFDVSD
V.ne  GPLQILTRQPVEGRSRDGGLRFGEMERDCM
S.ce  GPMQVLTRQPVEGRSRDGGLRFGEMERDCMIAHGAASFLKERLMEASD
A.bi  GPVQILTRQPVEGRSRDGGLRFGEMERD
A.ma  GPLQILTRQPVEGRS
H.sa  GPIQILNRQPMEGRSRDGGLRFGEMERDCQIAHGAAQFLRERLFEASD
C.el  GPIQMMNRQPMEGRARDGGLRFGEMERDCQISHGATQFLRERLFEVSD
P.fa  GPLTMITRQPTEGRSRDGGLRFGEMERDCMISHGSAKMLKERLFEESD
E.oc  GPLQILTRQPTEGRSRHGGLRFGEMERDCMVSHGAARFLKERLFDVSD
E.co  GSYSLVTQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSD
```

Figure 5-4 *N. locustae* possesses the conserved "homology domains" of RPB2

```
N. locustae       DVSDPYSLSVCELCGLFAVSTPE---IAECRGCKNRT
S. cerevisiae     EASDAFRVHICGICGLMTVIAKLNHNQFECKGCDNKI
S. pombe          DCSDAYRVIVCDICGLIAIASYK-KDSYECRSCQNRT
H. sapiens        EASDPYQVHVCNLCGIMAIANTR-THTYECRGCRNKT
C. elegans        EVSDPYHVYVCNNCGLIVVANLR-TNSFECKACRNKT
A. thaliana       DQSDAYRVHVCEVCGLIAIANLK-KNSFECRGCKNKT
P. falciparum     EESDAYRVHVCDNCGLCCIADIN-KNAYECTVCNSKT
E. octocarinatus  DVSDCYTVHVCRICGLICEANLR-QQKYLCRGCQNST
```

Figure 5-5 *N. locustae* possesses the Zinc finger motif characteristic of RPB2

The characteristic Zn finger is found at the expected position in *N. locustae* RPB2. This motif falls at amino acid position 1153-1195 in the *S. cerevisiae* sequence.

An alignment of these regions of the microsporidian sequences with a selection of eukaryotic RPB2 sequences, in addition to sequence from the β subunit of *E. coli* RNA polymerase, is shown in Figure 5-4. In addition to possessing these domains, RPB2 sequences characteristically also contain a Zinc finger motif of type $CX_2CGX_{7\text{-}24}CX_2C$ (Archambault and Friesen 1993). The *N. locustae* RPB2 sequence contains such a motif at the expected position, as shown in Figure 5-5. (Unfortunately, the motif cannot be identified in *V. necatrix* as the sequence is truncated prior to the expected position of the Zinc finger.)

Phylogenetic analysis of RPB2 sequences

To test the position of the microsporidia, a phylogenetic analysis was carried out with eukaryotic RPB2 sequences, using the eukaryotic RPA2 sequences as an outgroup. All available taxa were included, except those sequences which are almost identical to others, or are extremely partial. Maximum likelihood distances were calculated for this 32 taxon, 790 position dataset, based on the JTT substitution matrix with eight gamma categories and one invariable category. BioNJ and Fitch algorithms were employed for tree reconstruction, and the resulting trees are shown in Figure 5-6a and 5-6b, respectively. In both trees, animal, plant and fungal clades are well-supported, and the relationships retrieved among the fungal groups are in agreement with the current understanding of fungal phylogeny (Bruns et al. 1992) with chytrids branching first, followed by zygomycetes and then the ascomycete-basidiomycete group (although the two chytrid sequences do not branch together in these analyses). In contrast, other accepted relationships, like the sisterhood of animals and fungi (Baldauf and Palmer 1993), are not obtained. Moreover, the bootstrap support uniting the animals and plants is extremely strong (98% Fitch; 95% NJ). The BioNJ and Fitch trees differ only in the position of the

Figure 5-6 Phylogeny of RPB2

(a) BioNJ tree of RPB2

Analysis of RPB2 sequences, employing RPA2 sequences as an outgroup. In the BioNJ tree the microsporidia branch in a basal position. Fitch and neighbor-joining bootstraps are shown (for 100 resampling replicates) with Fitch above neighbor-joining.

(b) Fitch tree of RPB2

In contrast with Figure 5-6a, in the Fitch tree the microsporidia branch as a sister-group to the fungi, but the relationship is not well supported.

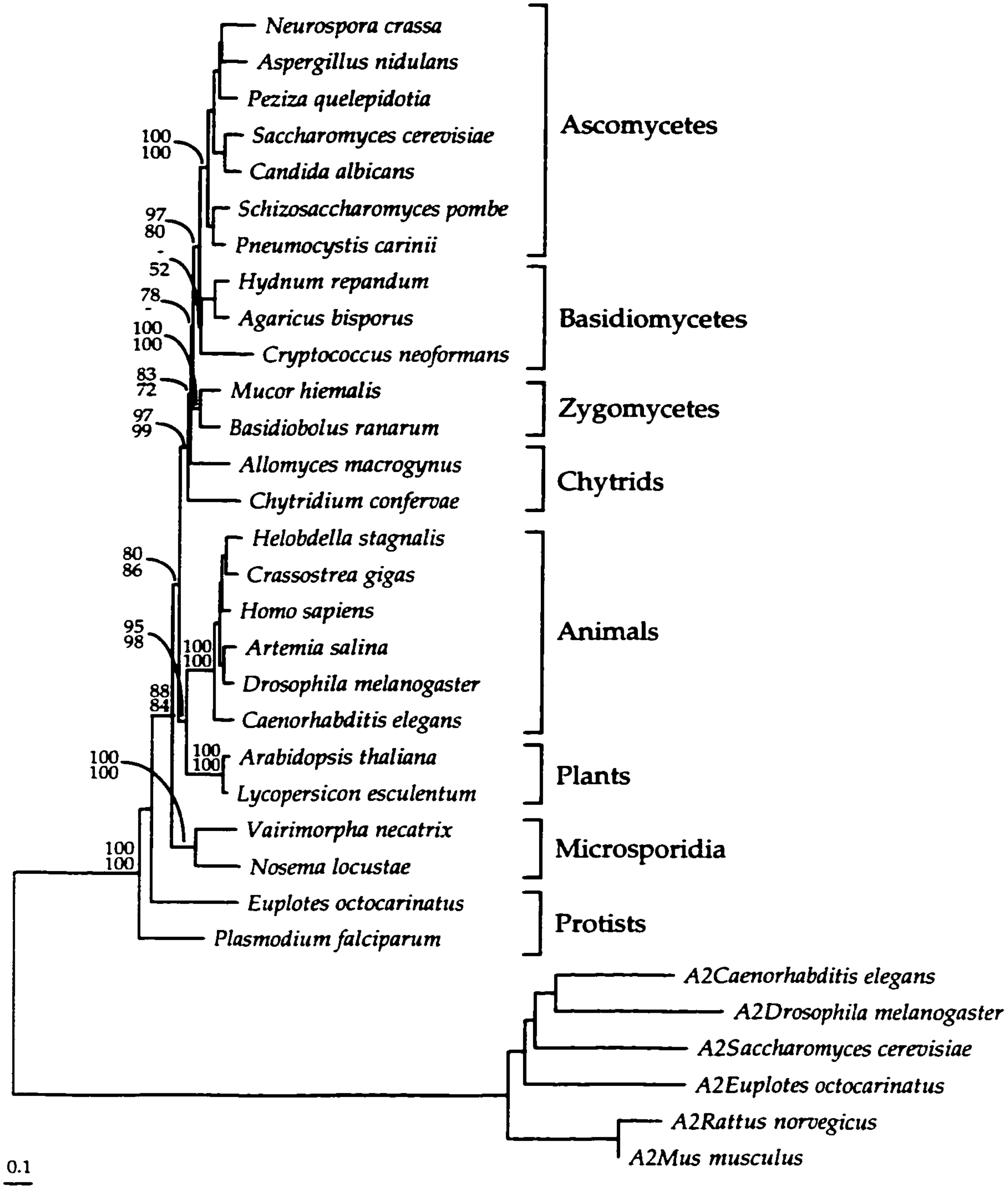


Figure 5-6a BioNJ tree of RPB2

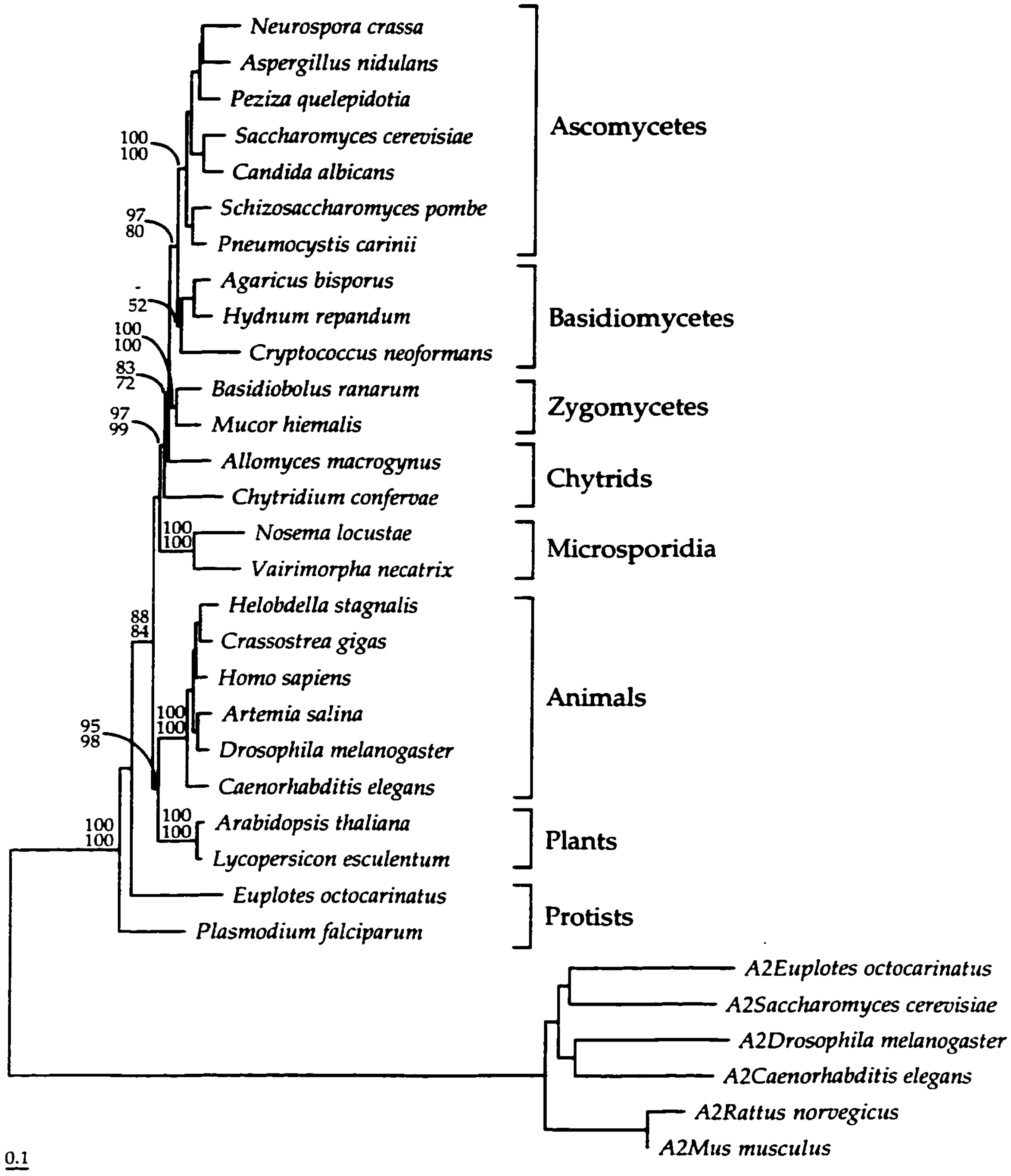


Figure 5-6b Fitch tree of RPB2

microsporidia, where the BioNJ tree places the microsporidia branching prior to the divergence of fungi, animals and plants, and the Fitch tree favours the microsporidia branching as a sister group to the fungi. Bootstrap values for the position of the microsporidia do not support a sisterhood of the microsporidia and fungi; in fact the more basal position of the microsporidia is quite strongly supported (80 and 86% Fitch and NJ bootstrap support for microsporidia branching outside fungi, animals and plants). Note that the positions of the microsporidia in this analysis differ from the Fitch tree that includes all partial RPA2, RPB2 and RPC2 sequences (Figure 5-1), however the BioNJ tree including all partial sequences (described previously) gives the RPB2 topology seen in Figure 5-6a.

An exhaustive ProtML search (Adachi and Hasegawa 1996) was undertaken with a constrained tree, where the strongly supported groups evident in the distance analysis were constrained (i.e. animals, plants, fungi, microsporidia, and the RPA2 outgroup). Of the one hundred most likely trees, the best tree shared the basic topology of the BioNJ tree shown in Figure 5-6a, with the microsporidia branching outside the fungi+animal/plant clade. Consensus trees from the ProtML analysis were constructed in two different ways with conflicting results. Using mol2con (from Dr. A. Stoltzfus) the consensus tree is identical to the best tree and the RELL support for the microsporidia branching outside of the fungi+animal/plant clade is 57%. Constructing a consensus tree with TreeCons (Jermiin et al. 1997) with a Class V (exponential) weighting scheme at 0.05-0.001 significance levels resulted in a tree identical to the Fitch tree shown in Figure 5-6b, except that the two alveolate sequences branch together in the TreeCons consensus with weak support (RLS <50%). The fact that TreeCons constructs a consensus tree that includes the microsporidia+fungi clade indicates that there are many trees not significantly less likely than the best tree that indicate a relationship between the microsporidia and the fungi. However, the RLS value supporting the sisterhood of microsporidia and fungi ranges from 44-45% with α=0.05-0.001. Keeping in mind that RLS values are inflated compared to bootstrap values, the relationship between microsporidia and fungi in this analysis is very weak indeed.

To try to reconcile the conflicting results regarding the position of the microsporidia in RPB2 phylogeny, Kishino-Hasegawa tests (accounting for site-to-site rate variation with six gamma categories) were performed to test the strength of differing topologies. Sixteen different topologies were tested including those shown in Figure 5-6a and 5-6b, as well as topologies where the microsporidian sequences were grafted onto each external node, and internal nodes with animals, plants, each of the alveolates, and with each fungal division. The best tree was that shown in Figure 5-6b, with microsporidia branching as a sister group to fungi. Five other topologies were not significantly worse (at a level of 0.05) and these include placing the microsporidia with each of the chytrids, between the chytrids, outside of the zygomycetes, and branching outside the fungi+animal/plant clade (Figure 5-6a topology). This test therefore did not favour either of the conflicting topologies over the other; however, it is interesting that several positions of the microsporidia within the fungi were excluded by the K-H tests.

In an attempt to test whether the exclusion of the microsporidia from certain positions within the fungal radiation was truly significant, phylogenetic reconstruction was carried out with only the fungal and microsporidian sequences. Maximum likelihood distances were calculated as above, and Fitch and BioNJ trees constructed. Expected fungal relationships were resolved in both trees, and the BioNJ tree placed the microsporidian sequences branching outside the zygomycetes, while the Fitch tree placed the microsporidia branching as a sister to the ascomycete+basidomycete clade, the latter relationship being a topology excluded from the previous K-H test. It should also be noted that the only relationships supported by bootstrap values of over 50% are the monophyly of ascomycetes, the ascomycete+basidimycete clade, the monophyly of zygomycetes, and the monophyly of microsporidia. Furthermore, K-H tests moving the microsporidia to seven different external and internal nodal positions within the fungi do not find any tree to be significantly worse than any other.

DISCUSSION

Although it is only recently that molecular evidence has come to suggest that microsporidia could be related to fungi, similarities between microsporidia and fungi have been noted before. However, these characters were often not specific to fungi (for example, the presence of chitin, separate thymidilate synthase and dihydrofolate reductase genes, and mitotic similarities) or did not unite the microsporidia with any particular fungal type (for example, apparent similarities in meiotic cycles). Very recently Cavalier-Smith has suggested that microsporidia could have evolved from the parasitic trichomycetes (members of the zygomycete division) (Cavalier-Smith 1998). This idea is based on apparent similarities between the microsporidian polar tubule and an organelle possessed by some harpallalean trichomycetes that could be extrusive, and is predicted to be involved in attaching the fungus to its host. It is likely best to be wary when considering this latest classification of the microsporidia based on their infective organelle, since other anomalous classifications of the microsporidia (as myxosporidia, for example) have had a similar basis (Lom and Vávra 1962).

Testing whether the microsporidia are affiliated with a specific type of fungus, or alternatively if they arose from a protozoan relative of fungi, has been hindered by the lack of a molecular dataset with appropriate taxon sampling. As noted in previous chapters, there are relatively few microsporidian protein-coding genes in the database, and of these none has been sampled in fungi to the extent that includes sequences representing all four fungal divisions. In an attempt to develop such a molecular dataset, the second largest subunit of RNA polymerase II has been sequenced from two microsporidia and a selection of ascomycete, basidiomycete, zygomycete and chytrid fungi.

Phylogenetic analysis with the RPB2 sequences was unfortunately less informative regarding the fungal relationship of the microsporidia than was hoped. However, that microsporidia are related to fungi is weakly supported by RPB2 phylogeny. Microsporidia do branch as a sister-group to the fungi in Fitch trees with an RPA2 outgroup and a consensus tree based on a constrained ProtML analysis includes the microsporidia-fungi clade. Support in both these analyses is weak, and although Kishino-Hasegawa topology testing indicated that the best tree has the microsporidia branching as a sister-group to the fungi, a basal origin for the microsporidia was not excluded at a 95% confidence level.

K-H testing did exclude a number of positions for the microsporidia within the fungal radiation, specifically excluding microsporidia branching with zygomycetes as well as at positions both outside and within the basidiomycete-ascomycete clade. To focus more closely on the relationship between the microsporidia and the fungi, analyses were also carried out including only those sequences. Resulting trees were poorly resolved; only the monophyly of ascomycetes was strongly supported, with moderate support for the basidiomycete-ascomycete clade. Topology tests by the Kishino-Hasegawa method were indicative of the lack of resolution. The best tree had microsporidia branching with the zygomycetes, but all other positions within the fungal tree were not significantly worse at a 95% confidence level. In sum, the phylogenetic relationship between the microsporidia and fungi in RPB2 trees appears to be too weak to address specific questions about the fungal origin of the microsporidia.

Comparison by eye of the conserved regions of RPB2 (partial alignments shown in Figure 5-4) revealed a synapomorphy in Region C shared by microsporidia and zygomycetes to the exclusion of all other eukaryotes (methionine in microsporidia and zygomycetes, leucine in all others). In the same region, microsporidia and zygomycetes share a tryptophan residue where all other fungi possess phenylalanine; however, in this case other non-fungal eukaryotes also have tryptophan at this position. No other such sequence similarities could be noted in the other domains and quite clearly it would be unwarranted to base any strong conclusions about a zygomycete relationship for the microsporidia on such a scant comparison that is not mirrored in the phylogenetic results. Recall, though, that it is with the zygomycetes that Cavalier-Smith has most recently allied the microsporidia (Cavalier-Smith 1998). Additional evidence that microsporidia could have evolved at least after the divergence of chytrid fungi includes the presence of stacked golgi in chytrid fungi, but unstacked cisternae in all higher fungi and in microsporidia. Furthermore, microsporidia, ascomycetes, basidiomycetes, and zygomycetes lack 9+2 microtubule structures, whereas chytrid fungi in general are known to be flagellated during some life cycle stages.

Although the RPB2 phylogenetic analysis described here does not significantly bear on the issue of whether microsporidia are derived from a protozoan fungal ancestor or from within the fungi, there is a possibility that if more microsporidia and/or zygomycete and chytrid fungi were added to dataset, results would be more fruitful. In any event, just as several molecular phylogenetic results have had to converge on a single alternative origin of the microsporidia for the microsporidia-fungi relationship to be taken seriously, it will take several more well-developed analyses to convincingly determine the specifics of this relationship.

APPENDIX A

MEDIA AND SOLUTIONS

LB (Luria-Bertani) Medium

1L:	10 g	Bacto-tryptone
	5 g	yeast extract
	10 g	NaCl

For solid media, 15 g of agar was added.

NZY

1L:	5 g	NaCl
	2 g	$MgSO_4\text{-}7H_2O$
	5 g	yeast extract
	10 g	NZ amine (casein hydrolysate)

For solid media, 15 g of agar was added.

For top agar, 0.7% w/v agarose was added.

50 x TAE

1L:	242 g	Tris base
	57.1 g	Glacial acetic acid
	100ml	0.5 M EDTA (pH8)

TE

10 mM Tris

0.1 mM EDTA

10 x TBE

1L:	108 g	Tris base
	55 g	Boric acid
	40 ml	0.5M EDTA

SM Buffer

1L:	5.8 g	NaCl
	2 g	$MgSO_4$-$7H_2O$
	50 ml	1M Tris-Hcl (pH 7.5)
	5 ml	2% w/v gelatin

20 x SSC

1L:	175.3 g	NaCl
	88.2 g	sodium citrate

Church's Buffer

1L:	64.08 g	Na_2HPO_4
	19.32 g	NaH_2PO_4
	0.37 g	EDTA
	70 g	SDS

TEN

40 mM Tris-HCl

1 mM EDTA
150 mM NaCl

APPENDIX B

WALKING PRIMERS

(Orientation is 5' to 3')

N. locustae TBP gene:

TBP.seq1	AGAATGGACGCGCCGGATCT
TBP.seq2	ATAATGCGCATAAGAGACCC
TBP.seq3	AGGCGAATGCTGAACTGTGT
TBP.seq4	CCATGAAGTAGCTCTGATAC
TBP3'seq	AGCTGCGACACACAGTTCAG
TBP5'seq	TCTCGAGAACTTCTGCGCTG

N. locustae TPI gene:

TPI.seq	AGAGAATTCTTGGTCTGAG
TPI.seq1	ACTTATGCGCAAACAGCTATC
TPI.seq2	GTGTTACACAGTAAGTTGAG
TPI.seq3	AGCATCCGGAGATAAAAGAC
TPI.seq4	ATATGACGTCTCTCAGAATG

T. vaginalis PRP8 gene:

816.seq1	TCATTATGAAGACACCTCCG
816.seq2	CTCGTCATCGATATCTATGC
816.seq3	AGATCTCTGAGTCTGATTCG
816.seq4	TGACAAATAACTTTGGACGG
816.seq5	GTTACAGTCTTAAGAGCTGC
816.seq6	TCCTAATCACGACTTCTTGG
816.seq7	CACAGTTGTCAAAGGATCAG
816.seq8	ATAAAACAACAATAAAATAC
816.seq9	TGCATCACGAGGTTCAAGGG
816.seq10	ACTTGCATGACGTCCATGAC
816.seq11	AGATACACATTGTGGTGGTC
816.seq12	CCAAATTGTGGTCGACGACC
816.seq13	GCCATCCTTTGTCGTAAGCC
816.seq14	CATGTTGAGATTGGTCGTCG
816.seq15	AACACGATGTAAATCTCGG
816.seq16	ATGCCGAACAGATTCCCTCC
816.seq17	ATAAGTTATCGATTTCTCAG
816.seq18	GCGTAAATTTCGTAGTTCCGG
816.seq19	CCGGAACTACGAAATTTACGC
816.seq20	AGAGTTGATTCTTTCGCACC
816.seq21	GAAGTAATTCCAGGTGGTCC

816.seq22	ATTTGTCGAATATGGCCTGG
816.seq23	ATGGAGACTGGAGACACTGG
816.seq24	AGGATACCTTCAACATTCTC
816.seq25	AACCATTCGTAAACAGCCGC
816.seq26	GTTAAAGGGTGATGGAGCCC
816.seq27	CGCAAGAGATGGAACAGTCG
816.seq28	CGACTGTTCCATCTCTTGCG
816.seq29	TGAACAGATCTCATCACAGG
816.seq30	TGATGGAGAGATGGCTTGGC
816.seq31	TGTACTGTATGTGGCAATCG
816.seq32	ATACTCTGACTTTGACAGGC
816.seq33	TGGAGCAATTGAGATCTTGC
gexR-a	CGAATTCCAGTTCACTTCAC
gexR-b	AACATGAGGAAAGAAGCCGG

T. vaginalis alpha-tubulin gene:

alpha-1	ATTCAGATCCAATTTCATTC
alpha-2	TAGAAATCAAAGTACATGAG
Tub.seq2	ATAAGGAAGCCCTGAAGACC
Tub.seq3	GCCATACAACTCCATTCTCG

N. locustae RNA polymerase I second-largest subunit partial coding sequence:

RPA2.seq1	TGGCCTAACATCGATATGCC
RPA2.seq2	CTCCAGCAGCATTCCTATGG
RPC2.seq1	CTGATTCGGTCGCACATGCC

REFERENCES

Achsel, T., K. Ahrens, H. Brahms, S. Teigelkamp, and R. Luhrmann. 1998. The human U5-220kD protein (hPrp8) forms a stable RNA-free complex with several U5-specific proteins, including an RNA unwindase, a homologue of ribosomal elongation.factor EF-2, and a novel WD-40 protein. *Mol Cell Biol.* **18**:6756-6766.

Adachi, J., and M. Hasegawa. 1996. MOLPHY: Programs for molecular phylogenetics. Inst. of Stat. Math., Tokyo.

Anderson, G. J., M. Bach, R. Luhrmann, and J. D. Beggs. 1989. Conservation between yeast and man of a protein associated with U5 small nuclear ribonucleoprotein. *Nature.* **342**:819-821.

Archambault, J., and J. D. Friesen. 1993. Genetics of eukaryotic RNA polymerases I, II, and III. *Microbiol Rev.* **57**:703-724.

Ares, M., Jr. 1986. U2 RNA from yeast is unexpectedly large and contains homology to vertebrate U4, U5, and U6 small nuclear RNAs. *Cell.* **47**:49-59.

Baldauf, S. L., and J. D. Palmer. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A.* **90**:11558-11562.

Beggs, J. D., S. Teigelkamp, and A. J. Newman. 1995. The role of PRP8 protein in nuclear pre-mRNA splicing in yeast. *J Cell Sci Suppl.* **19**:101-105.

Berget, S. M., C. Moore, and P. A. Sharp. 1977. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A.* **74**:3171-3175.

Biderre, C., G. Méténier, and C. P. Vivarès. 1998. A small spliceosomal-type intron occurs in a ribosomal protein gene of the microsporidian *Encephalitozoon cuniculi. Mol Biochem Parasitol.* **94**:283-286.

Biderre, C., M. Pages, G. Méténier, E. U. Canning, and C. P. Vivarès. 1995. Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium *Encephalitozoon cuniculi. Mol Biochem Parasitol.* **74**:229-231.

Bonen, L. 1993. Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J.* **7**:40-46.

Boulanger, S. C., S. M. Belcher, U. Schmidt, S. D. Dib-Hajj, T. Schmidt, and P. S. Perlman. 1995. Studies of point mutants define three essential paired nucleotides in the domain 5 substructure of a group II intron. *Mol Cell Biol.* **15**:4479-4488.

Breckenridge, D. G., Y. Watanabe, S. J. Greenwood, M. W. Gray, and M. N. Schnare. 1999. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. *Proc Natl Acad Sci U S A.* **96**:852-856.

Brody, E., and J. Abelson. 1985. The "spliceosome": yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science.* **228**:963-967.

Brow, D. A., and C. Guthrie. 1988. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature.* **334**:213-218.

Brow, D. A., and C. Guthrie. 1990. Transcription of a yeast U6 snRNA gene requires a polymerase III promoter element in a novel position. *Genes Dev.* **4**:1345-1356.

Brown, J. D., and J. D. Beggs. 1992. Roles of PRP8 protein in the assembly of splicing complexes. *EMBO J.* **11**:3721-3729.

Bruns, T. D.R. Vilgalys, S. M. Barns, et al. 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. *Mol Phylogenet Evol.* **1**:231-241.

Bruzik, J. P., and T. Maniatis. 1992. Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature.* **360**:692-695.

Bruzik, J. P., and J. A. Steitz. 1990. Spliced leader RNA sequences can substitute for the essential 5' end of U1 RNA during splicing in a mammalian in vitro system. *Cell.* **62**:889-899.

Bui, E. T., P. J. Bradley, and P. J. Johnson. 1996. A common evolutionary origin for mitochondria and hydrogenosomes. *Proc Natl Acad Sci U S A.* **93**:9651-9656.

Burley, S. K., and R. G. Roeder. 1996. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem.* **65**:769-799.

Canning, E. U. 1990. Phylum Microspora. in *Handbook of Protoctista*. Boston, Jones and Bartlett. pp. 53-72.

Cavalier-Smith, T. 1983. A 6-kingdom classification and a unified phylogeny. in *Endocytobiology II: Intracellular Space as Oligogenetic*. Berlin, Walter de Gruyter. pp. 1027-1034.

Cavalier-Smith, T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet*. **7**:145-148.

Cavalier-Smith, T. 1993. Kingdom protozoa and its 18 phyla. *Microbiol Rev*. **57**:953-994.

Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biol Rev*. **73**:203-266.

Cavalier-Smith, T., and E. E. Chao. 1996. Molecular phylogeny of the free- living archezoan *Trepomonas agilis* and the nature of the first eukaryote. *J Mol Evol*. **43**:551-562.

Chabot, B., S. Bisotto, and M. Vincent. 1995. The nuclear matrix phosphoprotein p255 associates with splicing complexes as part of the [U4/U6.U5] tri-snRNP particle. *Nucleic Acids Res*. **23**:3206-3213.

Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*. **12**:1-8.

Clark, C. G., and A. J. Roger. 1995. Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*. *Proc Natl Acad Sci U S A*. **92**:6518-6521.

Copertino, D. W., and R. B. Hallick. 1993. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem Sci*. **18**:467-471.

Datta, B., and A. M. Weiner. 1991. Genetic evidence for base pairing between U2 and U6 snRNA in mammalian mRNA splicing. *Nature*. **352**:821-824.

Datta, B., and A. M. Weiner. 1993. The phylogenetically invariant ACAGAGA and AGC sequences of U6 small nuclear RNA are more tolerant of mutation in human cells than in *Saccharomyces cerevisiae*. *Mol Cell Biol*. **13**:5377-5382.

Dexter Dyer, B. 1990. Phylum Zoomastigina Class Parabasalia. in *Handbook of Protoctista*. Boston, Jones and Bartlett. pp. 252-258.

DiMaria, P., B. Palic, B. A. Debrunner-Vossbrinck, J. Lapp, and C. R. Vossbrinck. 1996. Characterization of the highly divergent U2 RNA homolog in the microsporidian *Vairimorpha necatrix*. *Nucleic Acids Res.* **24**:515-522.

Edlind, T. D., J. Li, G. S. Visvesvara, M. H. Vodkin, G. L. McLaughlin, and S. K. Katiyar. 1996. Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. *Mol Phylogenet Evol.* **5**:359-367.

Embley, T. M., and R. P. Hirt. 1998. Early branching eukaryotes? *Curr Opin Gen Dev.* **8**:624-629.

Eschenlauer, J. B., M. W. Kaiser, V. L. Gerlach, and D. A. Brow. 1993. Architecture of a yeast U6 RNA gene promoter. *Mol Cell Biol.* **13**:3015-3026.

Fabrizio, P., and J. Abelson. 1990. Two domains of yeast U6 small nuclear RNA required for both steps of nuclear precursor messenger RNA splicing. *Science.* **250**:404-409.

Fabrizio, P., B. Laggerbauer, J. Lauber, W. S. Lane, and R. Luhrmann. 1997. An evolutionarily conserved U5 snRNP-specific protein is a GTP-binding factor closely related to the ribosomal translocase EF-2. *EMBO J.* **16**:4092-4106.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). distributed by author, University of Washington, Seattle.

Field, D. J., and J. D. Friesen. 1996. Functionally redundant interactions between U2 and U6 spliceosomal snRNAs. *Genes Dev.* **10**:489-501.

Flegel, T. W., and T. Pasharawipas. 1995. A proposal for typical eukaryotic meiosis in microsporidians. *Can J Microbiol.* **41**:1-11.

Frank, D., B. Patterson, and C. Guthrie. 1992. Synthetic lethal mutations suggest interactions between U5 small nuclear RNA and four proteins required for the second step of splicing. *Mol Cell Biol.* **12**:5197-5205.

Garcia-Blanco, M. A., G. J. Anderson, J. Beggs, and P. A. Sharp. 1990. A mammalian protein of 220 kDa binds pre-mRNAs in the spliceosome: a potential homologue of the yeast PRP8 protein. *Proc Natl Acad Sci U S A.* **87**:3082-3086.

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* **14**:685-695.

Germot, A., H. Philippe, and H. Le Guyader. 1996. Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proc Natl Acad Sci U S A*. **93**:14614-14617.

Germot, A., H. Philippe, and H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial- type HSP70 in *Nosema locustae*. *Mol Biochem Parasitol*. **87**:159-168.

Gilbert, W., M. Marchionni, and G. McKnight. 1986. On the antiquity of introns. *Cell*. **46**:151-153.

Grabowski, P. J., S. R. Seiler, and P. A. Sharp. 1985. A multicomponent complex is involved in the splicing of messenger RNA precursors. *Cell*. **42**:345-353.

Green, M. R., T. Maniatis, and D. A. Melton. 1983. Human beta-globin pre-mRNA synthesized in vitro is accurately spliced in Xenopus oocyte nuclei. *Cell*. **32**:681-694.

Guialis, A., M. Moraitou, M. Patrinou-Georgoula, and A. Dangli. 1991. A novel 40S multi-snRNP complex isolated from rat liver nuclei. *Nucleic Acids Res*. **19**:287-296.

Guthrie, C., and B. Patterson. 1988. Spliceosomal snRNAs. *Annu Rev Genet*. **22**:387-419.

Hardy, S. F., P. J. Grabowski, R. A. Padgett, and P. A. Sharp. 1984. Cofactor requirements of splicing of purified messenger RNA precursors. *Nature*. **308**:375-377.

Hartwell, L. H., C. S. McLaughlin, and J. R. Warner. 1970. Identification of ten genes that control ribosome formation in yeast. *Mol Gen Genet*. **109**:42-56.

Hashimoto, T., Y. Nakamura, T. Kamaishi, F. Nakamura, J. Adachi, K. Okamoto, and M. Hasegawa. 1995. Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. *Mol Biol Evol*. **12**:782-793.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Mol Biol Evol*. **11**:65-71.

Hashimoto, T., L. B. Sanchez, T. Shirakura, M. Muller, and M. Hasegawa. 1998. Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc Natl Acad Sci U S A.* **95**:6860-6865.

Hernandez, N., and W. Keller. 1983. Splicing of in vitro synthesized messenger RNA precursors in HeLa cell extracts. *Cell.* **35**:89-99.

Hetzer, M., G. Wurzer, R. J. Schweyen, and M. W. Mueller. 1997. Trans-activation of group II intron splicing by nuclear U5 snRNA. *Nature.* **386**:417-420.

Hinkle, G., D. D. Leipe, T. A. Nerad, and M. L. Sogin. 1994. The unusually long small subunit ribosomal RNA of *Phreatamoeba balamuthi. Nucleic Acids Res.* **22**:465-469.

Hinz, M., M. J. Moore, and A. Bindereif. 1996. Domain analysis of human U5 RNA. Cap trimethylation, protein binding, and spliceosome assembly. *J Biol Chem.* **271**:19001-19007.

Hirt, R. P., B. Healy, C. R. Vossbrinck, E. U. Canning, and T. M. Embley. 1997. A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. *Curr Biol.* **7**:995-998.

Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci U S A.* **96**:580-585.

Hodges, P. E., S. P. Jackson, J. D. Brown, and J. D. Beggs. 1995. Extraordinary sequence conservation of the PRP8 splicing factor. *Yeast.* **11**:337-342.

Horner, D. S., R. P. Hirt, S. Kilvington, D. Lloyd, and T. M. Embley. 1996. Molecular data suggest an early acquisition of the mitochondrion endosymbiont. *Proc R Soc Lond B Biol Sci.* **263**:1053-1059.

Imamura, O., K. Saiki, T. Tani, Y. Ohshima, M. Sugawara, and Y. Furuichi. 1998. Cloning and characterization of a human DEAH-box RNA helicase, a functional homolog of fission yeast Cdc28/Prp8. *Nucleic Acids Res.* **26**:2063-2068.

Jackson, S. P., M. Lossky, and J. D. Beggs. 1988. Cloning of the RNA8 gene of *Saccharomyces cerevisiae*, detection of the RNA8 protein, and demonstration that it is essential for nuclear pre- mRNA splicing. *Mol Cell Biol.* **8**:1067-1075.

Jamieson, D. J., and J. D. Beggs. 1991. A suppressor of yeast spp81/ded1 mutations encodes a very similar putative ATP-dependent RNA helicase. *Mol Microbiol.* **5**:805-812.

Jamieson, D. J., B. Rahe, J. Pringle, and J. D. Beggs. 1991. A suppressor of a yeast splicing mutation (prp8-1) encodes a putative ATP-dependent RNA helicase. *Nature.* **349**:715-717.

Jermiin, L. S., G. J. Olsen, K. L. Mengersen, and S. Easteal. 1997. Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. *Mol Biol Evol.* **14**:1296-1302.

Kamaishi, T., T. Hashimoto, Y. Nakamura, Y. Masuda, F. Nakamura, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996. Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, *Glugea plecoglossi*: implications for the deepest branching of eukaryotes. *J Biochem (Tokyo).* **120**:1095-1103.

Kamaishi, T., T. Hashimoto, Y. Nakamura, F. Nakamura, S. Murata, N. Okada, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. *J Mol Evol.* **42**:257-263.

Kandels-Lewis, S., and B. Seraphin. 1993. Involvement of U6 snRNA in 5' splice site selection. *Science.* **262**:2035-2039.

Kassaretis, G. A., E. T. Butler, D. Rouland, and M. J. Chamberlain. 1982. Bacteriophage SP6-specific RNA polymerase: mapping of SP6 DNA and selective in vitro transcription. *J Biol Chem.* **257**:5779-5788.

Keeling, P. J. 1998. A kingdom's progress: Archezoa and the origin of eukaryotes. *BioEssays.* **20**:87-95.

Keeling, P. J., J. A. Deane, and G. I. McFadden. 1998. The phylogenetic position of alpha- and beta-tubulins from the Chlorarachnion host and Cercomonas (Cercozoa). *J Eukaryot Microbiol.* **45**:561-570.

Keeling, P. J., and W. F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol Biol Evol.* **13**:1297-1305.

Keeling, P. J., and W. F. Doolittle. 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha- proteobacterial origin. *Proc Natl Acad Sci U S A*. **94**:1270-1275.

Keeling, P. J., and G. I. McFadden. 1998. Origins of microsporidia. *Trends Microbiol*. **6**:19-23.

Klenk, H.-P., W. Zillig, M. Lanzendorfer, B. Grampp, and P. Palm. 1995. Location of protist lineages in a phylogenetic tree inferred from sequences of DNA-dependent RNA polymerases. *Arch Protistenkd*. **145**:221-230.

Kole, R., and S. M. Weissman. 1982. Accurate in vitro splicing of human beta-globin RNA. *Nucleic Acids Res*. **10**:5429-5445.

Konarska, M. M., P. J. Grabowski, R. A. Padgett, and P. A. Sharp. 1985. Characterization of the branch site in lariat RNAs produced by splicing of mRNA precursors. *Nature*. **313**:552-557.

Krainer, A. R., T. Maniatis, B. Ruskin, and M. R. Green. 1984. Normal and mutant human beta-globin pre-mRNAs are faithfully and efficiently spliced in vitro. *Cell*. **36**:993-1005.

Kuhn, A. N., Z. Li, and D. A. Brow. 1999. Splicing factor Prp8 governs U4/U6 RNA unwinding during activation of the spliceosome. *Mol Cell*. **3**:65-75.

Kulesza, H., G. G. Simpson, R. Waugh, J. D. Beggs, and J. W. Brown. 1993. Detection of a plant protein analogous to the yeast spliceosomal protein, PRP8. *FEBS Lett*. **318**:4-6.

Lerner, M. R., J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz. 1980. Are snRNPs involved in splicing? *Nature*. **283**:220-224.

Lesser, C. F., and C. Guthrie. 1993. Mutations in U6 snRNA that alter splice site specificity: implications for the active site. *Science*. **262**:1982-1988.

Li, J. M., R. P. Haberman, and W. F. Marzluff. 1996. Common factors direct transcription through the proximal sequence elements (PSEs) of the embryonic sea urchin U1, U2, and U6 genes despite minimal similarity among the PSEs. *Mol Cell Biol*. **16**:1275-1281.

Liston, D. R., and P. J. Johnson. 1998. Gene transcription in *Trichomonas vaginalis*. *Parasitol Today*. **14**:261-265.

Logsdon, J. M. 1998. The recent origins of spliceosomal introns revisited. *Curr Opin Gen Dev.* **8**:637-648.

Logsdon, J. M., Jr., M. G. Tyshenko, C. Dixon, J. D-Jafari, V. K. Walker, and J. D. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc Natl Acad Sci U S A.* **92**:8507-8511.

Lom, J., and J. Vávra. 1962. A proposal to the classification within the subphylum Cnidospora. *Sys Zool.* **11**:172-175.

Lossky, M., G. J. Anderson, S. P. Jackson, and J. Beggs. 1987. Identification of a yeast snRNP protein and detection of snRNP-snRNP interactions. *Cell.* **51**:1019-1026.

Lücke, S., T. Klockner, Z. Palfi, M. Boshart, and A. Bindereif. 1997. Trans mRNA splicing in trypanosomes: cloning and analysis of a PRP8-homologous gene from *Trypanosoma brucei* provides evidence for a U5-analogous RNP. *EMBO J.* **16**:4433-4440.

Lundgren, K., S. Allan, S. Urushiyama, T. Tani, Y. Ohshima, D. Frendewey, and D. Beach. 1996. A connection between pre-mRNA splicing and the cell cycle in fission yeast: cdc28$^+$ is allelic with prp8$^+$ and encodes an RNA-dependent ATPase/helicase. *Mol Biol Cell.* **7**:1083-1094.

Machesky, L. M., R. H. Insall, and R. R. Kay. 1998. The helC gene encodes a putative DEAD-box RNA helicase required for development in *Dictyostelium discoideum. Curr Biol.* **8**:607-610.

MacMillan, A. M., C. C. Query, C. R. Allerson, S. Chen, G. L. Verdine, and P. A. Sharp. 1994. Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev.* **8**:3008-3020.

Madhani, H. D., R. Bordonne, and C. Guthrie. 1990. Multiple roles for U6 snRNA in the splicing pathway. *Genes Dev.* **4**:2264-2277.

Madhani, H. D., and C. Guthrie. 1992. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell.* **71**:803-817.

Madhani, H. D., and C. Guthrie. 1994. Dynamic RNA-RNA interactions in the spliceosome. *Annu Rev Genet.* **28**:1-26.

Metzenberg, S., and N. Agabian. 1996. Human and fungal 3' splice sites are used by *Trypanosoma brucei* for trans splicing. *Mol Biochem Parasitol.* **83**:11-23.

Michel, F., and J. L. Ferat. 1995. Structure and activities of group II introns. *Annu Rev Biochem*. **64**:435-461.

Miranda, R., L. M. Salgado, R. Sanchez-Lopez, A. Alagon, and P. M. Lizardi. 1996. Identification and analysis of the u6 small nuclear RNA gene from *Entamoeba histolytica*. *Gene*. **180**:37-42.

Morin, L., and J.-P. Mignot. 1995. Are Archamoebae true Archezoa? The phylogenetic position of *Pelomyxa* sp. as inferred from large subunit ribosomal RNA sequencing. *Eur J Protistol*. **31**:448

Müller, M. 1997. Evolutionary origins of trichomonad hydrogenosomes. *Parasitol Today*. **13**:166-167.

Müller, M. 1997. What are the microsporidia? *Parasitol Today*. **13**:455-456.

Nevins, J. R. 1979. Processing of late adenovirus nuclear RNA to mRNA. Kinetics of formation of intermediates and demonstration that all events are nuclear. *J Mol Biol*. **130**:493-506.

Newman, A. J. 1997. The role of U5 snRNP in pre-mRNA splicing. *EMBO J*. **16**:5797-5800.

Nilsen, T. W. 1993. Trans-splicing of nematode premessenger RNA. *Annu Rev Microbiol*. **47**:413-440.

Nilsen, T. W. 1995. trans-splicing: an update. *Mol Biochem Parasitol*. **73**:1-6.

Nilsen, T. W. 1998. RNA-RNA interactions in nuclear pre-mRNA splicing. in *RNA Structure and Function*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press. pp. 279-307.

Niu, X. H., T. Hartshorne, X. Y. He, and N. Agabian. 1994. Characterization of putative small nuclear RNAs from *Giardia lamblia*. *Mol Biochem Parasitol*. **66**:49-57.

Noble, S. M., and C. Guthrie. 1996. Identification of novel genes required for yeast pre-mRNA splicing by means of cold-sensitive mutations. *Genetics*. **143**:67-80.

Padgett, R. A., M. M. Konarska, P. J. Grabowski, S. F. Hardy, and P. A. Sharp. 1984. Lariat RNAs as intermediates and products in the splicing of messenger RNA precursors. *Science*. **225**:898-903.

Padgett, R. A., S. M. Mount, J. A. Steitz, and P. A. Sharp. 1983. Splicing of messenger RNA precursors is inhibited by antisera to small nuclear ribonucleoprotein. *Cell*. **35**:101-107.

Palmer, J. D., and J. M. Logsdon, Jr. 1991. The recent origins of introns. *Curr Opin Genet Dev*. **1**:470-477.

Paterson, T., J. D. Beggs, D. J. Finnegan, and R. Luhrmann. 1991. Polypeptide components of *Drosophila* small nuclear ribonucleoprotein particles. *Nucl Acids Res*. **19**:5877-5882.

Peebles, C. L., M. Zhang, P. S. Perlman, and J. S. Franzen. 1995. Catalytically critical nucleotide in domain 5 of a group II intron. *Proc Natl Acad Sci U S A*. **92**:4422-4426.

Peyretaillade, E., C. Biderre, P. Peyret, F. Duffieux, G. Méténier, M. Gouy, B. Michot, and C. P. Vivarès. 1998. Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucl Acids Res*. **26**:3513-3520.

Peyretaillade, E., V. Broussolle, P. Peyret, G. Méténier, M. Gouy, and C. P. Vivarès. 1998. Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin. *Mol Biol Evol*. **16**:683-689.

Philippe, H., and A. Adoutte. 1998. The molecular phylogeny of eukaryota: solid facts and uncertainties. in *Evolutionary Relationships Among Protozoa*. London, Chapman and Hall. pp. 25-56.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? *Curr Opin Genet Dev*. **8**:616-623.

Pinto, A. L., and J. A. Steitz. 1989. The mammalian analogue of the yeast PRP8 splicing protein is present in the U4/5/6 small nuclear ribonucleoprotein particle and the spliceosome. *Proc Natl Acad Sci U S A*. **86**:8742-8746.

Reyes, J. L., E. H. Gustafson, H. R. Luo, M. J. Moore, and M. M. Konarska. 1999. The C-terminal region of hPrp8 interacts with the conserved GU dinucleotide at the 5' splice site. *RNA*. **5**:167-179.

Reyes, J. L., P. Kois, B. B. Konforti, and M. M. Konarska. 1996. The canonical GU dinucleotide at the 5' splice site is recognized by p220 of the U5 snRNP within the spliceosome. *RNA*. **2**:213-225.

Roger, A. J., C. G. Clark, and W. F. Doolittle. 1996. A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. *Proc Natl Acad Sci U S A*. **93**:14618-14622.

Roger, A. J., O. Sandblom, W. F. Doolittle, and H. Philippe. 1999. An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. *Mol Biol Evol*. **16**:218-233.

Roger, A. J., S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin, and M. L. Sogin. 1998. A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc Natl Acad Sci U S A*. **95**:229-234.

Rogers, J., and R. Wall. 1980. A mechanism for RNA splicing. *Proc Natl Acad Sci U S A*. **77**:1877-1879.

Rozario, C., L. Morin, A. J. Roger, M. W. Smith, and M. Muller. 1996. Primary structure and phylogenetic relationships of glyceraldehyde-3-phosphate dehydrogenase genes of free-living and parasitic diplomonad flagellates. *J Eukaryot Microbiol*. **43**:330-340.

Ruskin, B., A. R. Krainer, T. Maniatis, and M. R. Green. 1984. Excision of an intact intron as a novel lariat structure during pre- mRNA splicing in vitro. *Cell*. **38**:317-331.

Scherly, D., W. Boelens, N. A. Dathan, W. J. van Venrooij, and I. W. Mattaj. 1990. Major determinants of the specificity of interaction between small nuclear ribonucleoproteins U1A and U2B" and their cognate RNAs. *Nature*. **345**:502-506.

Sharp, P. A. 1991. "Five easy pieces". *Science*. **254**:663.

Shea, J. E., J. H. Toyn, and L. H. Johnston. 1994. The budding yeast U5 snRNP Prp8 is a highly conserved protein which links RNA splicing with cell cycle progression. *Nucleic Acids Res*. **22**:5555-5564.

Shuster, E. O., and C. Guthrie. 1988. Two conserved domains of yeast U2 snRNA are separated by 945 nonessential nucleotides. *Cell*. **55**:41-48.

Sogin, M. L. 1989. Evolution of eukaryotic microorganisms and their small subunit RNAs. *Am Zool*. **29**:487-499.

Sogin, M. L., J. H. Gunderson, H. J. Elwood, R. A. Alonso, and D. A. Peattie. 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science*. **243**:75-77.

Staley, J. P., and C. Guthrie. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell.* **92**:315-326.

Stoltzfus, A. 1999. On the possibility of constructive neutral evolution. *J Mol Evol.* (in press)

Stoltzfus, A., D. F. Spencer, M. Zuker, J. M. Logsdon, Jr., and W. F. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. *Science.* **265**:202-207.

Straus, D., and W. Gilbert. 1985. Genetic engineering in the Precambrian: structure of the chicken triosephosphate isomerase gene. *Mol Cell Biol.* **5**:3497-3506.

Strauss, E. J., and C. Guthrie. 1991. A cold-sensitive mRNA splicing mutant is a member of the RNA helicase gene family. *Genes Dev.* **5**:629-641.

Streett, D. A. 1994. Analysis of *Nosema locustae* (Microsporida: Nosematidae) chromosomal DNA with pulsed-field-gel-electrophoresis. *J Invertebr Pathol.* **63**:301-303.

Strimmer, K., and A. von Haeseler. 1996. Quartet puzzling: a quartet maximum likelihood method method for reconstructing tree topologies. *Mol Biol Evol.* **13**:964-969.

Sweetser, D., M. Nonet, and R. A. Young. 1987. Prokaryotic and eukaryotic RNA polymerases have homologous core subunits. *Proc Natl Acad Sci U S A.* **84**:1192-1196.

Tang, J., N. Abovich, and M. Rosbash. 1996. Identification and characterization of a yeast gene encoding the U2 small nuclear ribonucleoprotein particle B" protein. *Mol Cell Biol.* **16**:2787-2795.

Tarn, W. Y., and J. A. Steitz. 1997. Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem Sci.* **22**:132-137.

Teigelkamp, S., A. J. Newman, and J. D. Beggs. 1995. Extensive interactions of PRP8 protein with the 5' and 3' splice sites during splicing suggest a role in stabilization of exon alignment by U5 snRNA. *EMBO J.* **14**:2602-2612.

Teigelkamp, S., E. Whittaker, and J. D. Beggs. 1995. Interaction of the yeast splicing factor PRP8 with substrate RNA during both steps of splicing. *Nucleic Acids Res.* **23**:320-326.

Thomas, J., K. Lea, E. Zucker-Aprison, and T. Blumenthal. 1990. The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucleic Acids Res.* **18**:2633-2642.

Umen, J. G., and C. Guthrie. 1995. A novel role for a U5 snRNP protein in 3' splice site selection. *Genes Dev.* **9**:855-868.

Umen, J. G., and C. Guthrie. 1995. Prp16p, Slu7p, and Prp8p interact with the 3' splice site in two distinct stages during the second catalytic step of pre-mRNA splicing. *RNA*. **1**:584-597.

Umen, J. G., and C. Guthrie. 1996. Mutagenesis of the yeast gene PRP8 reveals domains governing the specificity and fidelity of 3' splice site selection. *Genetics*. **143**:723-739.

Urushiyama, S., T. Tani, and Y. Ohshima. 1996. Isolation of novel pre-mRNA splicing mutants of *Schizosaccharomyces pombe*. *Mol Gen Genet.* **253**:118-127.

Vickerman, K. 1990. Phylum Zoomastigina Class Diplomonadida. in *Handbook of Protoctista*. Boston, Jones and Bartlett. pp. 200-210.

Vivares, C., C. Biderre, F. Duffieux, E. Peyretaillade, P. Peyret, G. Metenier, and M. Pages. 1996. Chromosomal localization of five genes in *Encephalitozoon cuniculi* (Microsporidia). *J Eukaryot Microbiol.* **43**:97S.

Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*. **326**:411-414.

Whittaker, E., and J. D. Beggs. 1991. The yeast PRP8 protein interacts directly with pre-mRNA. *Nucleic Acids Res.* **19**:5483-5489.

Whittaker, E., M. Lossky, and J. D. Beggs. 1990. Affinity purification of spliceosomes reveals that the precursor RNA processing protein PRP8, a protein in the U5 small nuclear ribonucleoprotein particle, is a component of yeast spliceosomes. *Proc Natl Acad Sci U S A.* **87**:2216-2219.

Will, C. L., and R. Luhrmann. 1997. Protein functions in pre-mRNA splicing. *Curr Opin Cell Biol.* **9**:320-328.

Wolff, T., and A. Bindereif. 1993. Conformational changes of U6 RNA during the spliceosome cycle: an intramolecular helix is essential both for initiating the U4-U6 interaction and for the first step of slicing. *Genes Dev.* **7**:1377-1389.

Wolff, T., R. Menssen, J. Hammel, and A. Bindereif. 1994. Splicing function of mammalian U6 small nuclear RNA: conserved positions in central domain and helix I are essential during the first and second step of pre-mRNA splicing. *Proc Natl Acad Sci U S A*. **91**:903-907.

Wu, J. A., and J. L. Manley. 1991. Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA. *Nature*. **352**:818-821.

Wyatt, J. R., E. J. Sontheimer, and J. A. Steitz. 1992. Site-specific cross-linking of mammalian U5 snRNP to the 5' splice site before the first step of pre-mRNA splicing. *Genes Dev*. **6**:2542-2553.

Xu, D., D. J. Field, S. J. Tang, A. Moris, B. P. Bobechko, and J. D. Friesen. 1998. Synthetic lethality of yeast *slt* mutations with U2 small nuclear RNA mutations suggests functional interactions between U2 and U5 snRNPs that are important for both steps of pre-mRNA splicing. *Mol Cell Biol*. **18**:2055-2066.

Yu, Y. T., P. A. Maroney, E. Darzynkiwicz, and T. W. Nilsen. 1995. U6 snRNA function in nuclear pre-mRNA splicing: a phosphorothioate interference analysis of the U6 phosphate backbone. *RNA*. **1**:46-54.

Zavanelli, M. I., J. S. Britton, A. H. Igel, and M. Ares, Jr. 1994. Mutations in an essential U2 small nuclear RNA structure cause cold-sensitive U2 small nuclear ribonucleoprotein function by favoring competing alternative U2 RNA structures. *Mol Cell Biol*. **14**:1689-1697.