# AN EVOLUTIONARY STUDY OF TWO MICROSPORIDIA

A thesis submitted for the degree of
DOCTOR OF PHILOSPHY
of the
Australian National University
Canberra

Department of Botany and Zoology
School of Life Sciences
and
CSIRO Division of Entomology
Black Mountain
Canberra

Robert Norman Rice
March 1999

# STATEMENT

This thesis contains no material that has previously been submitted for an academic record at this or any other university and is the original work of the author, except where acknowledged.

Robert Norman Rice

# ACKNOWLEDGEMENTS

Throughout life we are presented with opportunities for self fulfilment; the completion of this thesis represents one such opportunity. The very nature of this undertaking requires a great personal commitment. However, the final goal cannot be achieved without the commitment and assistance of others. It is those persons to whom I am indebted and wish to acknowledge.

I would like to thank my supervisor Dr Denis Anderson for providing me with the opportunity and the facilities to undertake this study. Doctor Paul Cooper as chairperson of my advisory committee, I thank you for attending to all those administrative requirements of the university. Doctor Paul Keese, I owe a special thank you for the hours spent in discussion and the technical guidance given to complete this study. Professor Adrian Gibbs, for his assistance and guidance relating to all matters of molecular systematics. Collectively, I thank you all for your efforts in reviewing the pre-submission drafts of my thesis.

I would also like to thank the staff, both technical and administrative, of the CSIRO Division of Entomology, Black Mountain, Canberra, for allowing me access to the facilities of the Division to undertake the research. Also, I thank you for the assistance and the friendship provided while in residence at Black Mountain.

I would also like to thank the Australian Beekeeping Industry, who, through its Honeybee Research and Development Council, provided a stipend and research funding (HBDRC Project Number CSE-3H) during the first two years of this study.

Last but not least, I thank my wife Janine and children Timothy, Melanie and Jessica for putting up with the long hours and frustration met with along the way.

# TABLE OF CONTENTS

STATEMENT ........ ii

ACKNOWLEDGEMENTS ........ iii

TABLE OF CONTENTS ........ iv

ABSTRACT ........ vii

CHAPTER 1 ........ 1

GENERAL INTRODUCTION ........ 1

1.1 BACKGROUND ........ 1

1.2 PHYLUM MICROSPORA SPRAGUE ........ 2

1.2.1 Cellular characteristics ........ 3

1.2.2 Molecular characteristics ........ 3

1.2.3 Life cycle ........ 4

1.2.4 Taxonomic classification ........ 5

1.2.5 Phylogenetic analysis ........ 5

*a. Analyses of the SSU rRNA gene* ........ 5

*b. Analyses based on protein sequences* ........ 8

1.3 EUKARYOTIC MITOCHONDRIAL ENDOSYMBIOSIS ........ 11

1.4 CONCLUSIONS AND AIMS ........ 12

CHAPTER 2 ........ 14

GENERAL MATERIALS AND METHODS ........ 14

2.1 SOURCE OF *NOSEMA* SPORES ........ 14

2.1.1 *Nosema apis* isolates ........ 14

2.1.2 *Nosema vespula* isolate ........ 14

2.2 PURIFICATION OF *NOSEMA* SPORES FROM HOST TISSUE ........ 15

2.3 ISOLATION OF GENOMIC DNA FROM PURIFIED *NOSEMA* SPORES ........ 16

2.3.1 *Nosema apis* ........ 16

*a. Isolation by spore germination* ........ 16

*b. Isolation by mechanical disruption of spores* ........ 16

2.3.2 *Nosema vespula* ........ 17

2.4 SELECTION OF PRIMERS FOR PCR ........ 17

2.5 Gene amplification using PCR ..... 19
2.6 Visulation of PCR amplification products ..... 19
2.7 Procedures for molecular cloning and sequencing ..... 19
2.8 Sequence analyses ..... 20
2.9 Phylogenetic analyses ..... 21

CHAPTER 3 ..... 23

LONG PCR OF RIBOSOMAL DNA REPEAT UNITS ..... 23
3.1 Introduction ..... 23
3.2 Materials and methods ..... 24
3.2.1 Isolation of genomic DNA template ..... 24
3.2.2 Selection of primers for LPCR ..... 24
3.2.3 Operon amplification using LPCR ..... 25
*a. Before the availability of commercial LPCR kits* ..... 25
*b. Amplification using the rTth XL™ kit (Perkin Elmer)* ..... 26
3.2.4 Visualisation of the LPCR amplification products ..... 26
3.2.5 Restriction analyses of the LPCR amplification products ..... 27
3.2.6 Procedures for molecular cloning and sequencing ..... 27
3.2.7 Sequence analyses ..... 27
3.3 Results and discussion ..... 27
3.3.1 LPCR amplification ..... 27
3.3.2 Restriction analysis ..... 29
*a. N. vespula LPCR-amplified fragment* ..... 29
*b. Yeast LPCR-amplified fragment* ..... 29
3.3.3 *N. vespula* sequence analysis ..... 30
3.3.4 Yeast sequence analysis ..... 30
3.4 Summary ..... 31

CHAPTER 4 ..... 37

*Nosema apis* - INTRASPECIFIC COMPARISON ..... 37
4.1 Introduction ..... 37
4.2 Materials and methods ..... 38
4.2.1 Isolation of genomic template ..... 38
4.2.2 Selection of primers for PCR amplification ..... 38
4.2.3 Amplification of target sequence ..... 39

4.2.4 Visualisation of the LPCR amplification products ........ 39
4.2.5 Procedures for molecular cloning ........ 39
4.2.6 Procedures for sequencing ........ 40
4.2.7 Sequence alignments and phylogenetic analysis ........ 40
4.3 Results and discussion ........ 40
4.3.1 Sequence length and variation ........ 40
4.3.2 Nucleotide variation including transition and transversion analysis ........ 41
4.3.3 Phylogenetic analysis ........ 42
4.4 Summary ........ 44
CHAPTER 5 ........ 60
*Nosema apis* - LARGE RIBOSOMAL SUBUNIT: SEQUENCE AND SECONDARY STRUCTURE ........ 60
5.1 Introduction ........ 60
5.2 Materials and methods ........ 62
5.3 Results and discussion ........ 62
5.3.1 LSU rRNA gene sequence ........ 62
5.3.2 Internal transcribed spacer sequence ........ 63
5.3.3 Comparison of the *N. apis* LSU rRNA ........ 64
5.3.4 A comparison of the *N. apis* and generic eukaryotic LSU rRNA secondary structure 64
*a Non-canonical base-pairs and other eukaryotic signature events* ........ 64
*b Secondary structure comparisons* ........ 65
*c Tertiary interactions comparisons* ........ 70
5.4 Summary ........ 70
CHAPTER 6 ........ 85
*Nosema vespula* rDNA REPEAT UNIT: ........ 85
SEQUENCE, GENE ARRANGEMENT AND SUBUNIT SECONDARY STRUCTURES ........ 85
6.1 Introduction ........ 85
6.2 Materials and methods ........ 86
6.3 Results and discussion ........ 87
6.3.1 rDNA repeat unit sequence ........ 87
6.3.2 Gene arrangement ........ 87

6.3.3 Sub-repeats of the intergenic spacer ........ 90
6.3.4 Secondary Structure Models ........ 90
*a. Small subunit ribosomal RNA* ........ 90
*b. Large subunit ribosomal RNA* ........ 93
*c. 5S sub-unit ribosomal RNA* ........ 97
6.4 . SUMMARY ........ 98

CHAPTER 7 ........ 125

*Nosema vespula*: ANALYSIS OF DIVERGENCE AND PHYLOGENETIC RECONSTRUCTION ........ 125
7.1 INTRODUCTION ........ 125
7.2 MATERIALS AND METHODS ........ 127
7.2.1 Sequence alignment ........ 127
7.2.2 Analysis of sequence divergence ........ 127
7.2.3 Phylogenetic Analysis ........ 128
7.3 RESULTS AND DISCUSSION ........ 129
7.3.1 Sequence alignments ........ 129
7.3.2 Phylogenetic inheritance and sequence divergence ........ 130
7.3.3 Phylogenetic reconstruction ........ 141
7.4 SUMMARY ........ 143
CHAPTER 8 ........ 160
GENERAL CONCLUSION ........ 160
8.1 CONCLUSION ........ 160

APPENDECES ........ 162
APPENDIX 1 SSU rRNA MODEL ........ 162
APPENDIX 2A, B LSU rRNA MODEL ........ 162

REFERENCES ........ 166

# ABSTRACT

This thesis describes evolutionary studies of two protozoan parasites of insects from the phylum Microspora. The first, *Nosema apis* Zander, is an economically important pathogen of adult stages of the cosmopolitan honey bee *Apis mellifera* Linnaeus. The second, *Nosema vespula*, while not formally described, was first discovered as a pathogen of the European Wasp *Vespula germanica* Fabricius, in Australia. Unlike *N. apis,* this microsporidian kills at the larval stage and has an extensive host range.

Typically, the eukaryotic rDNA repeat unit occurs as part of a tandem array. These arrays occur at one or several chromosomal locations. The total number of repeat units within each array is variable, ranging from only a few in lower eukaryotes to hundreds in higher eukaryotes. Within each repeat unit is a gene operon that encodes a leader sequence, the small subunit (SSU) ribosomal RNA (rRNA), the first internal transcribed spacer (ITS1), the 5.8S rRNA, the second internal transcribed spacer (ITS2) and the LSU rRNA. Also within the repeat unit is a non-transcribed spacer. The microsporidians differ from most other eukaryotes as they lack ITS2 such that the 5.8S and LSU rRNA genes are covalently linked. The Microspora has been placed as a basal group within the eukaryotic lineage based on the published 16S-like rRNA sequence of the microsporidian, *Viarimorpha necatrix* Pilley. This organism has eukaryotic ultrastructure, but prokaryotic-like rRNA features including 70S ribosomes and the covalently linked 5.8S-like and LSU rRNA genes.

There were two major aims of the work. The first aim was to determine the sequence for a region of the ribosomal DNA (rDNA) operon from nine *N. apis* isolates and to use this data to investigate genetic variability among *N. apis* isolates. The second aim was to sequence the entire *N. vespula* rDNA repeat unit as well as the *N. apis* large subunit (LSU) rRNA gene and use this data to investigate and clarify the divergence of the Microspora within the eukaryotic lineage.

The region sequenced for the *N. apis* isolates included the three prime (3') end of the SSU rRNA gene, the ITS rDNA and the five prime (5') end of the LSU rRNA gene. The rDNA sequence of the nine isolates were aligned and comparisons made. Amongst the nine isolates, both length and positional character-state variation were detected. The sequences ranged in length from 663 to 666 base pairs (bp). Thirty-one character-state substitution events were identified in the alignment. These substitution events occurred within the ITS rDNA and LSU rRNA gene. The substitution events identified within the LSU rRNA gene

correspond to regions of the LSU rRNA that are known to be variable across the eukaryotic lineage.

The rDNA repeat unit of *N. vespula* is 7292 basepairs (bp) in size and consists of an SSU rRNA gene of 1245 bp, an LSU rRNA gene of 2549 bp and a 5S rRNA gene of 120 bp. The tandem arrangement and size of the repeat units were confirmed by long PCR. The arrangement of the rRNA genes within the rDNA operon of *N. vespula* was found to be prokaryotic-like in two ways: 1) the SSU and LSU rRNA genes are separated by one ITS such that the 5.8S-like and the LSU rRNA genes are covalently linked; and 2) the 5S rRNA gene is located downstream from the LSU rRNA gene within the non-transcribed spacer.

From the sequence data obtained in this study, secondary structure models were developed for the LSU rRNA of *N. apis* and the SSU, LSU, and 5S rRNAs of *N. vespula.* The developed models are based on models proposed for other eukaryotes. The putative 5' and 3' termini of each rRNA gene were identified from the secondary structure models. Comparisons made between the proposed *N. apis* and *N. vespula* models and the generic eukaryotic models, revealed a number of unusual structural features. A number of the conserved structural elements apparent within the eukaryotic models, either differed in structure, size, or are absent in the proposed *N. apis* and *N. vespula* models. Indeed, several prokaryotic-like, structural features were found within the rRNA subunits of *N. vespula.* Furthermore, at some alignment positions within the SSU and LSU rRNAs that are conserved, but differ in character-state between the prokaryotes and eukaryotes, *N. vespula* possesses the prokaryotic character-state.

Additionally, the SSU and LSU rRNA sequences of *N. vespula,* were aligned with ten other eukaryotes, two eubacteria and two archaebacteria and used to construct evolutionary trees. These trees were inferred from unambiguously aligned positions (excluding alignment gaps) using maximum parsimony (MP) and neighbor joining (NJ) (of Galtier and Gouy corrected distance measures) analyses. The SSU and LSU rRNA MP trees were inconsistent with each other. The SSU rRNA MP tree agreed with previously published trees placing the Microsporidia at or near the base of the eukaryotic lineage (bootstrap support 86%). The LSU rRNA MP tree placed the Microsporidia further up the tree, diverging after several other protist taxa (60% support). The SSU and LSU rRNA NJ trees are congruent with each other placing the divergence of the Microsporidia at the base of the eukaryotic lineage (100% support). In this respect these trees are congruent with previously published SSU rRNA trees.

# CHAPTER 1

# GENERAL INTRODUCTION

## 1.1 BACKGROUND

Throughout history, humanity has endeavoured to sort and categorise the biological world. Linnaeus (1758) formalised a hierarchical system of nomenclature to classify and organise plants and animals. Evolutionists, like Lamarck (1809), Darwin (1859), and Haechel (1866) used this system of nomenclature to produce a classification based on phylogenetic relationships (Moritz and Hillis, 1996). However, it was not until the latter half of this century that attempts were made to reconstruct evolutionary history based on the shared attributes of extant and fossil organisms (Walter Zimmermann and Willi Hennig, see Moritz and Hillis, 1996).

Initially, phylogenetic relationships of organisms could only be studied at the macro-phenotypic level. The advent of the microscope increased the resolution of macro studies, but also revealed the world of micro-organisms; a new world of biological diversity to be sorted, described and categorised had been discovered. Among these micro-organisms are the members of the Kingdom Protoctista (John Hogg, 1861, see Hausmann and Hülsmann 1996; Margulis 1974 a, b; 1980; 1989). Recent phenotypic analyses of these single-celled free-living eukaryotes have led to the assignment of many Protoctistans to various phyla and other taxonomic sub-groups (Hausmann and Hülsman, 1996), but have failed to refine the taxonomic relationships, particularly at higher taxonomic levels (Hausmann and Hülsman, 1996). The taxon of parasitic Protoctista have proved very challenging to resolve. Reductions in physical, ultrastructural, and biochemical properties—a consequence of their parasitic lifestyles (Cavalier-Smith, 1987)—has apparently limited the phenotypic character-set by which the Protoctista can be unequivocally assigned. Hence, alternate character-sets such as DNA sequences are being used to resolve taxonomic relationships.

The DNA sequences that have been most widely used to attempt to resolve taxonomic relationships among the Protoctista are those of the SSU and LSU rRNA genes (Olsen and Woese, 1993). These genes are characterised by the presence of highly conserved and divergent domains and thus can be used to resolve taxonomic relationships at all levels (Hillis, Moritz, and Mable, 1996, and references therein). The following general descriptions of the Microspora include specific taxonomic information, and published phylogenetic conclusions.

## 1.2 PHYLUM MICROSPORA SPRAGUE

Members of the Phylum Microspora are generally referred to as the Microsporidia (Weber *et al.* 1994). All are nonflagellate microscopic spore-forming, obligate, intracellular parasites, that live either freely or in parasitophore vacuoles within the cytoplasm of the host cell. First described by Näegeli in 1857 (Weber *et al.,* 1994) as a newly identified parasite of silkworms (namely, *Nosema bombycis* Näegeli), there have now been more than 100 genera and nearly 1,000 species described (Canning, 1993; Sprague *et al.*, 1992; Weber *et al.,* 1994; Hausmann and Hülsmann, 1996).

Most commonly found in arthropods and bony fish, Microsporidia also parasitise Apicomplexans, Myxozoans, Ciliophorans and the metazoans: coelenterates, platyhelminths, nematodes, bryozoans, annelids, molluscs, and other invertebrates (Hausmann and Hülsmann, 1996). Two species, *N. bombycis* and *N. apis*, respectively, are economically important pathogens of silkworms (Ishihara and Hayashi, 1968) and honeybees (Bailey and Ball, 1991), while *N. acridophagus* Henry, *N. cuneatum* Henry, and *N. locustae* Canning have been trialed as micro biological control agents for the grasshopper *Melanoplus sanguinipes* Fabricius (Erlandson *et al.,* 1985; Lockwood and Debrey, 1990). Another species *Vairimorpha necatrix,* has been used to control caterpillars (Hausmann and Hülsmann, 1996).

The host range of the Microsporidia is not limited to the aforementioned taxonomic groups, but also includes mammals. Parasitism of mammals by microsporidians was first reported by Wright and Craighead (1922). *Encephalitozoon cuniculi* Levaditi *et al.* has been isolated from rodents (Wright and Craighead, 1922; Pakes *et al.,* 1975), canines, and primates (Canning and Hollister, 1987). The first substantiated report of a microsporidial infection in humans was reported by Matsubayashi *et al.* (1959), but reports of infections in humans continue to grow (Weber *et. al.*, 1994). However, most of these infections are opportunistic associated with the human immunodeficiency lentivirus (HIV) (Cali and Owen, 1988; Vossbrinck *et al.,* 1993; Zhu *et al.,* 1993, Weber *et al.,* 1994). Human infections have been reported from the genera *Encephalitozoon* spp., *Enterocytozoon* spp., *Septata* spp., *Pleistophora* sp., *Nosema* spp., and from several unclassified microsporidians (Weber *et al.,* 1994 for a review on human microsporidial infections).

### 1.2.1 Cellular characteristics

Microsporidians, by definition, are true eukaryotes. They possess a nucleus, an endomembrane system, and the separation of chromosomes on a mitotic spindle (Canning, 1989). However, ultrastructural and molecular studies of extant species demonstrate the presence of possible plesiomorphic character-states (Vossbrinck *et al.*, 1987). Several prokaryotic features of the Microsporidia have been described. The Microsporidia lack mitochondria (Marquardt and Demeree, 1985; Cavalier-Smith, 1987), permanent well-developed Golgi dictyosomes, and peroxisomes (Cavalier-Smith, 1993). Their ribosomes have sedimentation coefficients of 70S for the monosome and 30S and 50S for the subunits (Ishihara and Hayashi, 1968; Curgy *et al.*, 1980). Other prokaryotic features have also been identified at the molecular level.

### 1.2.2 Molecular characteristics

First discovered in *V. necatrix,* the rDNA operon of the Microsporidia lacks a separate 5.8S rRNA gene, but possess a sequence complementary to the 5.8S rRNA gene in the 5' region LSU rRNA gene (Vossbrinck and Woese, 1986). Excluding the microsporidians, the 5.8S rRNA gene is believed to be a universal characteristic of eukaryotes (Erdmann *et al.*, 1985). Cavalier-Smith (1993) however, has suggested that the absence of a separate 5.8S rRNA gene in the Microsporidia may be the consequence of secondary shortening. He proposed that a single deletion may have removed the RNA processing site from the pre-rRNA that is recognised by the enzyme responsible for cleaving the 5.8S rRNA from the LSU rRNA. There are also other examples of secondary loss in the microsporidial RNA genes that lend support to Cavalier-Smith's notion. From sequence data and secondary structure models, universally conserved regions of the SSU rRNA have also been reported as absent in *V. necatrix* (Neefs *et al.*, 1991) and *Vairimorpha lymentriae* Weiser (Vossbrinck *et al.*, 1993). Unusual molecular features also appear when considering the entire genome. At 2.9 mega bases (Mb), the genome of *E. cuniculi* is the smallest eukaryotic genome yet reported (Biderre *et al.*, 1995). This genome is even smaller than that of the bacterium *Escherichia coli* (Migula) Castellani and Chalmers (4.7 Mb) (Smith *et al.*, 1987).

Despite the small size of the *E. cuniculi* genome, other microsporidians may have much larger genomes, for example, the genome of *Glugea atherinae* Berrebi is estimated at 19.5 Mb (Biderre *et al.,* 1995). Genome size does not however necessarily reflect chromosome number. Karyotypic studies of several microsporidian species have revealed the presence of a haploid genome whose chromosome number varies from at least 8 chromosomes each in *N. costelytrae* and *Vairimorpha sp.* (Malone and McIvor, 1993) to 11 in *E. cuniculi* (Biderre *et al.*, 1995).

### 1.2.3 Life cycle

The following description of the life-cycle of Microsporidia has been reported by Tanada and Kaya (1993). Their life cycle has two phases: merogony and sporogony. Spores, that are the end-result of sporogony, and can survive outside their hosts, have an imperforate spore coat and are usually 3-6μm x 2-4μm, but may be up to 20μm in diameter. The interior of the spore, the sporoplasm, consists of a plasmalemma, cytoplasm, one or two nuclei, endoplasmic reticulum and ribosomes all enclosed in a coiled polar tube (diameter 0.1μm; length 100-400μm). In response to ingestion, the spore germinates and the polar tube everts penetrating a host gut cell. Subsequently, the sporoplasm is injected through the polar tube into the host cell. Within the host cell, the sporoplasm develops into a rounded meront with one (unikaryon) or two (diplokaryon) nuclei. During merogony these meronts divide mitotically and their progeny infect other host cells. Each meront subsequently develops into sporonts that have different cell membrane structures from meronts. Sporogony varies. In microsporans, such as in *Encephalitozoon* spp. and *Nosema spp.*, the first nuclear division is followed by cytokinesis and this produces two potential sporoblasts. In *V. necatrix*, cytokinesis does not follow nuclear division, and a pansporoblast or sporophorous vesicle is produced which later buds into unicellular sporoblasts (Larsson, 1986). Polymorphic forms however, containing both free spores and spores in sporophorous vesicles, have been found in both adult and larval forms of the same host (Canning, 1989). Sporoblasts develop into spores by the differentiation of the polar tube, the cell organelles and the spore coat. Spores are released when the host dies (Tanada and Kaya, 1993).

The above description of the life cycle of the Microsporidia has been determined from a few well-studied species. However, the life cycles of most are incompletely known (Flegel and Pasharawipas, 1995). Many species differ in morphology and ultrastructure at different life cycle stages of both the parasite and the host. For example, species from the genus *Amblyospora* exhibit three spore types: two in the mosquito host *Culex salinarius* Coquillet (Andreadis and Hall, 1979) and one in an alternate copepod host (Sweeney *et al.*, 1985; Andreadis, 1985). For this and the aforementioned reasons, the system of taxonomic identification for the Phylum Microspora has been based on the number of hosts they infect and differences in the chromosomal cycle.

### 1.2.4 Taxonomic classification

Generally, all species can be categorised into three groups (Canning, 1989): 1) specific to a single host and showing a mitotically reproducing uninucleate condition at all developmental stages (eg. *Encephalitozoon spp.*); 2) specific to a single host and showing only the diplokaryotic condition with synchronous mitotic division (eg. *Nosema spp.* and *Vairimorpha spp.*); and 3) infects alternate species and exhibits both uninucleate and diplokaryotic cells (eg. *Amblyospora spp.*).

The genus *Nosema* has been formally described by Sprague *et al.* (1992). Belonging to the phylum Microspora and of the class Dihaplophasea, each species has the nucleus paired as a diplokaryon during part of their life cycle. The Dihaplophasea belong to the order Dissociodihaplophasea in which haplosis occurs by nuclear dissociation resulting in unpaired nuclei. The Dissociodihaplophasea belong to the super family Nosematoidea Labbe, 1899, the members of which have bi-nucleate homospores that dissociate after the sporoplasm invades a new host. The cycle of gamete production ends with plasmogamy and nuclear dissociation. Finally, the genus *Nosema* belongs within the family Nosematidae Labbe, 1899 in which all reproduction stages occur in host cell hyaloplasm.

### 1.2.5 Phylogenetic analysis

*a. Analyses of the SSU rRNA gene*

The prokaryotic features of the microsporidia led Cavalier-Smith (1987) to propose that they are one of the earliest offshoots of the eukaryotic lineage. Molecular systematic studies of SSU rRNA gene sequence (Vossbrinck *et al.*, 1987) supported Cavalier-Smith's proposition. Further studies suggest that the Diplomonadida (Sogin *et al.*, 1989) and Trichomonadida (Olsen and Woese, 1993) also emerge early in eukaryotic evolution. The debate about the branching order of the Microsporidia, Diplomonadida and Trichomonadida however remained unresolved. These three lineages share a common feature, they all lack mitochondria. The most parsimonious hypothesis to explain this commonality is that the three lineages diverged from a common ancestor before mitochondrial symbiosis; an alternate hypothesis is that during divergence the three lineages suffered several independent losses of the organelle (Germot *et al.*, 1997). The remainder of this chapter reviews the debate about the divergence of the early eukaryotes, and includes recent evidence about the secondary loss of mitochondria.

A search of the GenBank database (National Centre for Biotechnology Information) in December 1997 using the search criteria 'Organism', returned 110 sequences for the term 'Microsporidia'. Sequence types and numbers are listed as follows: 1) the complete and partial SSU rDNA (54); 2) partial LSU rDNA (21); 3) SSU/ITS/LSU rDNA (12); 4) 5S rDNA (1); 5) pseudo rRNA (1); 6) alpha tubulin (4); 7) beta tubulin (5); 8) elongation factor 1 alpha (2); 9) elongation factor 2 (1); 10) isoleucyl-tRNA synthetase (1); 11) glutamyl-tRNA synthetase (2); 12) U2 homologue (1); 13) HSP70 (2); and 14) genomic fragments (3).

Most of these sequences were determined for the express purpose of intraspecific phylogenetic studies, and to determine the likely branch point of the Microsporidia on the eukaryotic tree. Vossbrinck *et al.* (1987) were the first to include a microsporidian sequence in a phylogenetic reconstruction. This report includes a distance-based phylogenetic tree reconstructed from the SSU RNA sequences of eight eukaryotes (*V. necatrix, Euglena gracilis* Klebs, *Trypanosoma brucei* Plimmer and Bradford, *Dictyostelium discoideum* Raper, *Paramecium tetraurelia* Sonneborn, *Saccharomyces cerevisiae* (Meyen) Hansen, *Zea mays* Linnaeus, *Xenopus laevis* Daudin), using one archaebacteria (*Sulfolobus solfataricus* Zillig *et al.*), and one eubacteria (*E. coli*) as the outgroup to the tree. The results suggested that microsporidians are a sister group to all other eukaryotes.

These results were challenged in a similar phylogenetic reconstruction that also included the SSU rRNA sequence from the Diplomonadida *Giardia lamblia* Lamblia (Sogin *et al.*, 1989). Analysis of these sequences using parsimony (PAUP: Swofford, 1985) and distance methods (Fitch and Margoliash, 1967) produced trees similar to Vossbrinck *et al.* except for the branching order of *V. necatrix* and *G. lamblia*. The topology of the distance method placed *G. lamblia* as the earlier offshoot below *V. necatrix*, and was favoured as the true topology. The tree generated by parsimony analysis was considered unlikely because the branch length for *V. necatrix* in the distance analysis was abnormally long. This abnormal length was considered to indicate abnormal rates of change in one or more lineages over time; a violation of the hypotheses of homogeneity (Galtier and Gouy, 1995). Under parsimonious phylogenetic reconstructions, such violations can cause anomalous and deep or alternate branching patterns in comparison to analyses by distance methods (Felsenstein, 1978; Lake, 1987). Another important indicator in this report was that a debate about the branching order of the Diplomonadida and the Microsporidia would follow, and that this question was not going to be easily resolved.

Three reports on phylogenetic reconstruction of the eukaryotic tree were published in 1991.

Each reconstruction analysed at least 40 SSU rRNA gene sequences (Eschbach *et al.,* 1991; Hendricks *et al.*, 1991; Wolters, 1991). Some differences in branching order existed among these reports. All for the most part however were congruent, placing *Giardia* at the base of the eukaryote tree followed by *Vairimorpha.*

One of the largest SSU rRNA phylogenetic reconstructions (153 sequences) attempting to resolve the divergence of the eukaryotic lineage included two diplomonads (*G. lamblia* and *Giardia intestinals* (Lambl) Alexeieff and *V. necatrix*. This reconstruction used a modified distance matrix using the archaebacteria *Halobacterium cutirubrum* (Lochhead) Elazari-Volcani as the outgroup (Van de Peer *et al.,* 1993). Predicably, the Diplomonadida branched earliest followed by *V. nectarix.* Leipe *et al.* (1993) however provide contradictory evidence as to the branching order of the diplomonads and microsporidians.

In an attempt to address the problem associated with violations of homogeneity in the data, Leipe *et al.* undertook a phylogenetic reconstruction of SSU rRNA sequence data that included the free-living diplomonad *Hexamita inflata* Dujardin (51% G + C), the trichomonad *Tritrichomonas foetus* (Riedmuller) Wenrich and Emerson (48% G + C), *G. lamblia* (75% G + C) and *V. necatrix* (35% G + C). Using a corrected (Jukes and Cantor, 1969) distance method (Olsen, 1988) and tree reconstruction (Saitou and Nei, 1987), they were able to demonstrate the effect of outlying prokaryotes with differing G + C content on tree topology. The inclusion of *S. solfataricus* (63% G + C) placed *G. lamblia* at the base of the eukaryotic tree but split the diplomonads into a paraphyletic group. This is an unlikely outcome based on ultrastructure data that strongly links *Hexamita* and *Giardia* (Leipe *et al.,* 1993). However, the use of prokaryotes with a normal G + C content as outlying groups caused the diplomonads to appear as a monophyletic group, but were preceded by *V. necatrix* and *T. foetus*. Leipe *et al.* concluded that while the branching order of *T. foetus* and *V. necatrix* was uncertain, they preceded the diplomonads. More specifically, that *V. necatrix* diverged before the diplomonads and that this was in contradiction to their original report (Sogin *et al.,* 1989).

Galtier and Gouy (1995) proposed an algorithm to account for estimating pairwise evolutionary distances without assuming homogeneity or stationarity (constancy of base composition within each lineage) of the evolutionary process. Application of this corrected distance method on a data set containing 12 archaebacteria and 14 eukaryotic SSU rRNA sequences, including *G. lamblia* and *V. necatrix,* placed *V. necatrix* as the earliest diverging eukaryote. Other corrected distance methods compared in this analysis reversed this order, placing *G. lamblia* as the earliest diverging eukaryote. It was suggested that the G + C-rich *G. lamblia* was artificially attracted to the G + C-rich prokaryotes. Also, that other distance method, which did not compensate for G + C, failed to produce the correct branching of order. The above analyses also include comparisons of phylogenetic reconstruction using maximum parsimony (Fitch, 1971) and maximum likelihood (Felsenstein, 1981; Olsen *et al.,* 1994). Maximum parsimony was found to be positively misleading in cases of composition bias. Maximum likelihood was however robust for computer simulations with constant rates of change amongst sites (stationarity), but performed poorly for the empirical data.

In addition to the effects of composition bias on the branching order of the early eukaryotes, it has been suggested that the deep branching of the amitochondrial Metamonada (Sogin *et al.,* 1989), Microsporidia (Vossbrinck *et al.,* 1987) and Parabasala (Leipe, *et al.,* 1993; Cavalier-Smith and Chao, 1996) may be an artefact caused by exceptionally high rates of molecular evolution; a consequence of a parasitic lifestyle (Wolters, 1991; Siddall *et al.,* 1992). Cavalier-Smith and Chao (1996) attempted to resolve this question and the question of divergence for these three phyla from the eukaryotic tree by applying maximum likelihood analysis. Maximum likelihood analysis is less vulnerable to violations of a constant substitution rate among lineages (Nei, 1991; Kuhner and Felsenstein, 1994). A computer computation of two weeks reconstructed a phylogenetic tree of 79 eukaryote and 5 archaebacterial SSU rRNA sequences. Among the eukaryotic sequences were those of both free-living and parasitic diplozoa and ten microsporidians. All forms of diplozoa were grouped together at the base of the eukaryotic tree. The grouping of the diplozoa was taken to discount the view held by Wolters (1991) and Siddall *et al.* (1992) that exceptionally high rates of molecular evolution in parasitic diplozoa was a consequence of lifestyle. It was concluded that this tree strengthens support for the *Giardia* as the earliest eukaryotic branch but concluded that the question still remained unanswered.

### *b. Analyses based on protein sequences*

It has been suggested that phylogenetic trees reconstructed from rRNA sequence data may be misleading because of composition bias in the organisms studied (Loomis and Smith, 1990; Hasegawa and Hashimoto, 1993). Protein phylogenies however, are believed to give a more robust estimate of early eukaryotic divergence as they are free from drastic bias of genome G + C content (Kamaishi *et al.,* 1995).

Amino acid sequences for translation elongation factor (EF) EF1 alpha and EF2 have been determined for a number of species (including Microsporidia and Diplomonadida) and used to reconstruct phylogenetic trees (Kamaishi *et al.,* 1996a; Kamaishi *et al.,* 1996b, Yamamoto *et al.,* 1997). All these analyses used maximum likelihood estimates of protein phylogeny (Kishino and Hasegawa, 1990) and assumed the JTT model for the amino acid substitution process (Cao *et al.,* 1994a) for EF-1 alpha, and the JTT-F model (Cao *et al.,* 1994b) for EF-2. In all analyses the Microsporidia were represented by *Glugea plecoglossi* Takahashi and Egusa (a parasite of fish) and the Diplomonadida by *G. lamblia.* These analyses concluded: that the sequence EF-1 alpha and EF2 are highly divergent; that *G. plecoglossi* was likely to be the earliest offshoot; that the microsporidians might be extremely ancient eukaryotes; and that they may have diverged before mitochondrial endosymbiosis.

Similarly, the protein sequences of both alpha and beta tubulins have been used for phylogenetic reconstruction of the early diverging eukaryotes. Edlind *et al.* (1996) reconstructed phylogenetic trees from 24 beta- (430 residues), two alpha, and two gamma-tubulin sequences using parsimony and distance methods. Both methods produced very similar trees consisting of two major lineages (fungal-animal and protozoan-plant) preceded by four independently branching sequences. Three of the four independently branching sequences were from the amitochondrial protists *Entamoeba histolytica* Schaudinn, *Trichomonas vaginalis* Donne, and *G. lamblia* respectively, with *Physarum polycephalum* branching between *T. vaginalis* and *G. lamblia.* An unexpected result was the branching of *Encephalitozoon hellem* Dider *et al.* within the fungal clade. These results were subsequently verified by additional parsimony and distance analysis of partial (152 residues) beta-tubulin sequences that included the sequence of the free living amoeba *Acanthamoeba polyphaga* (Puschkarew) Page. In this analysis, the branching position of *P. polycephalum* moved to within the protozoan-plant lineage and *G. lamblia* branched at the base of the animal-fungal lineage. *E. histolytica* and *T. vaginalis* remained as the first and second branches respectively, while *E. hellem* remained within the fungi. These results, regarding the amitochondrial protists presented, are incongruent with the aforementioned SSU rRNA and EF 1 alpha phylogenetic reconstructions. Similarly, Keeling and Doolittle (1996) analysed 58 alpha tubulin sequences including three microsporidial sequences: *E. hellem, N. locustae,* and *Spraguea lophii* (Doflein) Weissenberg. All three sequences grouped together, and as previously, within the fungal sequences. However, this paper does acknowledge that the alpha- and beta-tubulin genes of the Microsporidia and fungi are highly divergent, and that the result obtained may be a consequence of "long branch attraction" (Felsenstein, 1978) (Keeling and Doolittle, 1996).

Recently, Philippe and Adoutte (1998) have published trees for the SSU rRNA sequence and the protein sequences of beta-tubulin and actin. These trees contained at least 75 sequences from a range of species and were congruent with previously published trees of each respective type. They were able to demonstrate by plotting the observed verses inferred number of substitutions that all three genes were mutationally saturated, with saturation more accentuated in the beta-tublin and SSU rRNA sequences but also clearly apparent in the actin sequences. They also demonstrated that in trees containing a number of fast evolving species, that their branch lengths could be substantially underestimated and yet the species still be artefactually attracted to the outgroup by the phenomenon of long branch attraction.

An observation of Mooers *et al.* (1995) suggests that high levels of saturation leads to high levels of homoplasy in the data. Affected trees would appear as "unbalanced". In these trees, the species branch in a paraphyletic array on one side of the basal node rather than appearing as succession of dichotomies having a similar number of taxa on each side of successive nodes. The SSU rRNA, beta-tubulin and actin phylogenetic reconstructions appear as asymmetric trees. In such trees the basal region is unbalanced while the top forms a succession of dichotomies (Philippe and Adoutte, 1998). Philippe and Adoutte conclude that reconciliation of the rRNA and protein trees is pointless. Rather, that phylogenetic inferences be made using the balanced portion from the tops of several trees.

Two other papers of interest examine the aminoacyl-tRNA synthetase gene of *N. locustae* (Brown and Doolittle, 1995) and the U2 RNA homologue from *V. necatrix* (Di Maria *et al.*, 1996). The aminoacyl-tRNA synthetase genes are a multigene family whose divergence likely preceded that of the prokaryotes and the eukaryotes. Duplicated genes such as these have been used to determine the root of the universal tree. Maximum parsimony consensus trees of the amino acid sequences for isoleucyl-, leucyl-, and valyl-tRNA synthetases grouped the taxa analysed into three domains: the eukaryotes, the archaebacteria, and the eubacteria. Additionally, the eukaryotes and the archaebacteria appeared as sister clades. The portion of the reconstruction for isoleucyl-tRNA synthetases included the *N. locustae* gene and placed *N. locustae* as the lowest branch, although the internal nodes were not statistically significantly. This placement is congruent with that of the SSU rRNA phylogenies for the lower eukaryotes. However, neither trichomonads nor diplomonads were represented in this reconstruction. The U2 is one of several small nuclear RNAs (snRNAs) which are found in small ribonucleoprotein complexes (snRNPs) that appear to be ubiquitous in eukaryotes. These snRNAs (except for U6) possess a hypermethylated 2,2,7-trimethylguanosine 5' cap structure, while regions of sequence within the U2 gene have been shown to be invariant. The data from the U2 RNA homologue of *V. necatrix* again demonstrates the degree of divergence undergone by the Microsporidia. Several of the invariant nucleotides within the gene have 'alternate character-states', and the 5' cap structure while present is not a hypermethylated 2,2,7-trimethylguanosine 5' cap.

## 1.3 Eukaryotic mitochondrial endosymbiosis

Early in evolution, the free-living progenitor of the alpha-proteobacteria entered into an endosymbiotic association with a eukaryotic host cell. Today, the degenerate descendants of this eubacterium are the mitochondrion (Keeling and Doolittle, 1997). Phylogenetic reconstructions based on rRNA and EF 1 alpha and EF2, place the three amitochondriate protist lineages, Microsporidia, Metamonada (including Diplomonadida), and Parabasala (including Trichomonadida), as the earliest branches of the eukaryotic tree.

Recently, the 70 kilo Dalton (kDa) heat shock proteins (HSP70) were identified in the genomes of the Microsporidia *N. locustae* (Germot *et al.*, 1997) and *V. necatrix* (Hirt *et al.*, 1997). The amino acid sequence of these proteins contained motifs that are shared by mitochondria and proteobacteria HSP70s. Maximum likelihood analysis placed the microsporidial HSP70 sequences within the mitochondrial clade, a sister group to the alpha proteobacteria clade. Also, in both analyses, the microsporidian HSP70 sequences were placed as a sister sequence to that of the fungal mitochondria.

Other evidence exists for the secondary loss of mitochondria and the morphogenesis of mitochondria into hydrogenosomes in the trichomonad *T. vaginalis* (Müller, 1997). The encoding sequences of heat shock proteins HSP70 and HSP10, and the chaperonin cpn60 have been detected in the nuclear genome of *T. vaginalis*. The products of some of these genes were localised in the hydrogenosomes of *T. vaginalis* and other trichomonads.

Evidence for the secondary loss of mitochondria from *G. lamblia* has also been discovered. Soltys and Gupta (1994) demonstrated the presence of a 60 kDa protein from *G. lamblia* that cross-reacts with mammalian mitochondrial cpn60 antibodies. Henze *et al.* (1995) demonstrated the presence of the glyceraldehyde-3-phosphate dehydrogenase gene and Keeling and Doolittle (1997) demonstrated the presence of the triosephosphate isomerase gene in *G. lamblia*. Phylogenetic reconstruction of these sequences placed the *G. lamblia* sequences within the eukaryotic clade, itself a sister group to the alpha proteobacteria clade. Most recently, Roger *et al.* (1998) demonstrated the presence of the cpn60 gene in *G. lamblia*. Phylogenetic reconstruction placed *G. lamblia, E. histolytica,* and *T. vaginalis* as a sister group to the mitochondrial clade and the *G. lamblia*-mitochondrial clade as a sister group to the alpha-proteobacteria clade.

## 1.4 CONCLUSIONS AND AIMS

The consensus of most trees is that the Diplomonadida, Microsporidia and the Trichomonadida represent the three earliest offshoots of the eukaryotic tree. However, there is no agreement amongst these publications as to the correct branching order of these taxa.

To date, the gene sequences obtained from microsporidians provide a contradictory view of their phylogenetic relationships. Phylogenetic analyses of the SSU rRNA suggests that the diplomonads, including *G. lamblia,* are the earliest branching eukaryotes, followed by microsporidians. The sequences of EF 1 alpha and EF2 suggest however that the divergence of microsporidians from other eukaryotes predated the divergence of *G. lamblia.* The U2 RNA homologue from *V. necatrix* and isoleucine aminoacyl-tRNA synthetase from *N. locustae* are also highly divergent from other eukaryotes but there is insufficient sequence data from other protists to assess their phylogenetic significance. In contrast, the alpha and beta tubulin genes of microsporidians suggest a more recent phylogenetic affinity with fungi. Most recently, it has been demonstrated that the amitochondrial lineages are likely to be secondarily amitochondriate.

To date, the LSU rRNA genes have not been used to investigate the divergence of the amitochondriate lineages. The LSU rRNA gene has been determined for many organisms providing comprehensive data for phylogenetic reconstruction. In addition, the secondary structure of the LSU rRNA is highly conserved amongst all organisms, allowing useful comparisons of widely divergent organisms.

In this thesis I investigate the ribosomal RNA sequences of two parasitic Protoctistan species from the genus *Nosema* in the phylum Microspora. The sequence of the LSU rRNA gene was determined for *N. apis* and a model for the LSU rRNA secondary structure proposed. Additionally, intraspecific variation for nine isolates of this species was considered along with the phylogenetic relationship of this species within the genus *Nosema.* Secondly, the sequence of the entire rDNA repeat unit of a yet formally undescribed species of Microspora was determined. The location of each ribosomal gene within the operon of the rDNA repeat unit was identified and a secondary structure model for each rRNA subunit developed. The sequence data obtained and the secondary structure models developed were used to attempt to clarify the likely evolutionary relationships of the Microsporidia within the eukaryotes. I did this by investigating the following four aspects:

1. The arrangement of the rRNA genes within the rDNA operons of the Microsporidia, eukaryotes and prokaryotes;

2. A comparison of SSU and LSU rRNA secondary structure elements of the Microsporidia with those from both prokaryotes and eukaryotes;

3. The degree of divergence of the microsporidial SSU and LSU rRNAs; and

4. Phylogenetic inference using the SSU and LSU rRNA sequences.

# CHAPTER 2

# GENERAL MATERIALS AND METHODS

## 2.1 SOURCE OF *NOSEMA* SPORES

### 2.1.1 *Nosema apis* isolates

*N. apis* infects and replicates in the epithelial cells of the honeybee midgut (Fries, 1993). Nine isolates of *N. apis* were used in this study (Table 2.1). Except for the Java isolate, all were obtained from the European honeybee, *Apis mellifera* (Apoidea: Apidae). The isolate from Java was obtained from the Asian honeybee *A. cerana* Fabricius. The isolates from Canada, New Zealand, and Sweden were provided as purified spores, while the remainder of the isolates were obtained as infections in whole bees from which the spores were recovered.

Samples of 35 honeybees, potentially infected with *N. apis,* were collected from the entrances of beehives. Five of the 35 bees were chosen at random. Their alimentary tracts were removed by crushing the thorax between the fingers, grasping the sting and terminal sclerites with tweezers, and gently pulling the alimentary tract away from the abdomen. A small piece of the midgut was removed from each honeybee, crushed between a microscope slide and a cover slip and microscopically examined at 400x magnification for the presence of *N. apis* spores. On confirmation of a *N. apis* infection within the beehive, the alimentary tracts of the remaining 30 bees were removed and stored at $4^{0}$C awaiting spore recovery and purification.

### 2.1.2 *Nosema vespula* isolate

The *N. vespula* isolate was a gift from Dr Denis Anderson, CSIRO Division of Entomology, Canberra, Australia. This as yet undescribed species of microsporidian was originally isolated by Dr Anderson from the infected larvae of the European wasp *Vespula germanica* Fabricius (Vespoidea: Vespidae). From preliminary morphological and ultrastructure studies as well as studies on the reproductive cycle, this species was tentatively name "*Nosema vespula*".

Experimental research has shown that this isolate infects an extensive host range including hymenopterans, dipterans, and lepidopterans. Since discovery, this species has been maintained *in vivo* using *Helicoverpa armigera* Hübner caterpillars as hosts. The organism initially infects the host's epithelial gut cells and then moves to the host's fat bodies where parasite replication occurs (Dr Denis Anderson, personal communication). A stock of the organism had been reared in caterpillars that had in turn been stored at -20$^0$C. Spore numbers, sufficient for use in molecular biological techniques, were obtained from a mass infection of 250 caterpillars. The caterpillars were a gift from Dr Peter Christian, CSIRO Division of Entomology, Canberra. The inoculum, used to infect these caterpillars, was recovered (Section 2.2) from frozen stocks. Host caterpillars were maintained on a soybean flour and wheat germ diet (Teakle and Jensen, 1985). Third- and fourth-instar caterpillars were starved for 3 hours and then fed 0.25 gm blocks of diet inoculated with 10 µl of a $10^5$ spores ml$^{-1}$ dilution in distilled water. The caterpillars were maintained at 28$^0$C and fed on demand for nine days. The infected caterpillars were then stored at -20$^0$C awaiting spore recovery and purification.

## 2.2 PURIFICATION OF *NOSEMA* SPORES FROM HOST TISSUE

Spores grown in bees or caterpillars were liberated from tissue by macerating either five entire caterpillars or 30 alimentary tracts in 30 ml of distilled water using a mortar and pestle. To remove large particulate matter, the spore suspension was filtered through four layers of Kimwipe (Kimberly-Clark, Australia). The resulting spore suspension was centrifuged at 1000 g for 20 minutes, the supernatant removed, the spores resuspended in 1ml of distilled water and centrifuged into a discontinuous gradient of neutralised Percoll (Sigma) according to the method of Sato and Watanabe (1980). The Percoll gradient was constructed by the sequential layering of 7 ml 100%, 8 ml 75%, 8 ml 50% and 8 ml 25% Percoll in a 50 ml ultracentrifuge tube (25 x 89 mm). A 1 ml aliquot of spore suspension was immediately overlayed onto the gradient and centrifuged at 3,000 g for 90 minutes using a Beckman 28S rotor in a Beckman L8-70M ultracentrifuge. Spores became visible as a white band at the 75-100% interface, while some spores passed through the gradient and were visible as a pellet among the debris in the bottom of the ultracentrifuge tube. The band at the interface was removed with a Pasteur pipette, placed in a 50 ml ultracentrifuge tube and diluted with 35 ml of distilled water. The suspended spores were pelleted at 3,000 g for 30 minutes and the supernatant removed. The spores were resuspended in 1 ml of distilled water, transferred to an eppendorf tube and washed three times with distilled water to remove Percoll. Washing of the spores was by centrifugation at 1,000 g for 5 minutes, removal of the supernatant, and resuspension in distilled water. At the third wash the spore concentration was determined using a Neubauer Counting Chamber (Cantwell, 1970). The spores were then pelleted once more, the supernatant removed, and the spores suspended in distilled water to a concentrations of aapproximately $10^8$ spores $ml^{-1}$. Each spore suspension prepared this way was stored at $4^0C$ in 00.5 ml aliquots pending the isolation of genomic DNA. The spore and debris sediment were also washed by the above method. However, following the final wash, spores were suspendedd at concentrations of approximately $10^5$ spores $ml^{-1}$ for later use as inocula for infecting bees oor caterpillars.

## 2.3 ISOLATION OF GENOMIC DNA FROM PURIFIED *NOSEMA* SPORES

### 2.3.1 *Nosema apis*

#### a. *Isolation by spore germination*

A 0.5 ml aliquot of purified spores was pelleted at 1,000 g for 5 minutes and the supernatant removed. The pellet of spores was resuspended in 200 µl of freshly prepared germination buffer (0.5 M sodium chloride, 0.5 M sodium hydrogern carbonate, pH to 6.0 with 0.1 M orthophosphoric acid; De Graaf *et al.,* 1993) and incubatted at $37^0C$ for 15 minutes allowing spores to germinate. One ml of DNA isolation buffer (0.55% SDS, 124 mM EDTA pH 8.0, 250 mM Tris-HCl pH 9.2, 0.2% 2-mercaptoethanol) wass added to the germinating spore suspension, and the mixture shaken for 2 minutes. The mmixture was incubated at $55^0C$ for 1 hour and then chilled on ice. The SDS and proteins were precipitated by adding of 300 µl of 5 M potassium acetate pH 7.2 and incubating on ice for 15 mminutes. The SDS-protein precipitate was pelleted by centrifugation at 15,000 g for 20 minutes aand the supernatant removed to a new Eppendorf tube. The DNA was precipitated from the supernatant by the addition of 950 µl of absolute ethanol and incubation on ice for 15 minutes. TThe DNA precipitate was pelleted by centrifugation at 15,000 g for 20 minutes, the supernatant removed, and the DNA pellet washed with 200 µl of ice cold 70% ethanol. The pellet was then air dried at room temperature, suspended in 20 µl of TE buffer and stored at $-20^0C$.

#### b. *Isolation by mechanical disruption of spores*

This method was used to extract genomic DNA from the spores of *N. apis* isolates where there was insufficient material available to use the germination protocol. A 200 µl aliquot of purified spores was placed in a 0.5 ml Eppendorf tube and pelletted by centrifugation at 1,000 g for 5 minutes. The supernatant was removed and the pellet sussspended in 150 µl of STE buffer (100 mM sodium chloride, 1 mM EDTA, and 10 mM Tris-HCl, pH 8.0; Baker *et al.,* 1995). One hundred and fifty mg of 0.45 mm glass beads (Sigma) was added and the tube vortexed at maximum speed for 20 seconds to disrupt the spore coat. Immediately, the tube was placed in a 95$^0$C heating block for 5 minutes to denature proteins. The tube was then centrifuged at 15,000 g for 3 minutes and the supernatant removed to a new tube and stored at -20$^0$C until needed.

### 2.3.2 *Nosema vespula*

The method of Crozier (1991) was devised for the isolation of genomic DNA from honeybees. The CTAB buffer used in Crozier's method was found to induce germination of *N. vespula* spores when the spores were incubated in the buffer at 37$^0$C. A pellet of purified spores was suspended in 500 µl of CTAB buffer (100 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0, 1.4 M sodium chloride, 2% Hexadecyltrimethylammonium bromide w/v) containing 0.2% 2-mercaptoethanol and incubated at 37$^0$C for 30 minutes (to allow spores to germinate) and then at 65$^0$C for 90 minutes. The mixture was extracted with phenol/chloroform twice (Sambrook *et al.,* 1989) and the DNA precipitated by adding of 750 µl of isopropanol and incubating overnight at -20$^0$C. The DNA precipitate was centrifuged at 15,000 g for 20 minutes to pellet the DNA. The supernatant was removed and the pellet washed twice with 200 µl of cold 70% ethanol. The pellet was air dried at room temperature and then suspended in 20 µl of TE buffer and stored at -20$^0$C.

## 2.4 SELECTION OF PRIMERS FOR PCR

Primers for PCR amplification were synthesised with an Applied Biosystems model 381A DNA synthesiser. Primer concentration was calculated from the absorbancy measured at OD260 by the method of Sambrook *et al.* (1989). A total of 19 primers were used to amplify and sequence the entire rDNA repeat unit of *N. vespula,* and the 3' region of the SSU rRNA gene, the entire ITS and the 5' region of the LSU rRNA gene of *N. apis.* Eight of these primers were used for amplification of target template in both species, while three were specific for *N. apis* and eight specific for *N. vespula.* Two of the primers (NV1703F and NV3493R) were specific to nucleotide sequences within the LSU rRNA gene that are conserved among eukaryotes (Guttel, 1994b). A third primer (NV457R) was selected from the SSU rRNA gene of *V. necatrix* (Vossbrinck *et al.,* 1987). The remaining primers were selected from sequence data obtained for *N. apis* and *N. vespula* (Chapters 5 and 6).

The primers with the largest G + C content (ie. > 50%) but minimal likelihood of primer dimer formation were chosen. Also, an *Eco*RI restriction site was included in some primers to facilitate cloning of the amplified fragment. The primers were assigned an identification code that indicated: 1) the 5' nucleotide position of the primer template sequence that corresponded to the complete *N. vespula* operon sequence; 2) whether the primer was complementary to the non-coding strand (F) or coding strand (R); and 3) if the primer was used for direct sequencing (D). For example, NV765RD = *N. vespula*, sequence position 765, complementary to the coding strand, and used for direct sequencing from PCR product. The template was either *N. vespula* (Nv) DNA, *N. apis* (Na) DNA, or both (Nv/Na).

The primer, template, and primer sequences were as follows:

| | | |
|---|---|---|
| NV411FD | Nv | ACT TGT TCC AAG AGT GTG TAT G |
| NV457R | Nv | GAG GAA TTC*GGA CTT TTT AAC TGC ATC AAT C |
| NV740FD | Nv | ACG GAA GAA TAC CAC AAG GAG |
| NV765RD | Nv | ATC CAC TCC TTG TGG TAT TCT TCC GTC |
| NV984RD | Nv | GTC TGA GGG CAT AAC GGA CCT G |
| NV1161F | Nv/Na | GTT GTG GAA TTC GTG CAA GCT ACT TGA ACA ATA TG |
| NV1222RD | Nv | CAA CAG CAA CCA TGT TAC GAC |
| NV1373FD | Nv/Na | GAT AAC CCT TTG AAC TTA AG |
| NV1584RD | Nv/Na | ACT ACC AAG CAG CCC TAC TCA |
| NV1629FD | Nv/Na | ATG GTA TAC CGA TAG CAA AT |
| NV1690RD | Nv/Na | GGC TAA CAC CCA CAC ATT TTC AC |
| NV1703F | Nv/Na | GCC CGT CTT GAA ACA CGG A |
| NV1851R | Nv/Na | GAA TGA GAA TTC GGA TCC AAT AAA CTG TTG CTT ATC |
| NV2394FD | Nv | AGT AAC AAT ATA TTT GTA TAG ATA TAG |
| NV3408FD | Na | GTC GTC TCT TCT GAT CAT CGT AG |
| NV3442F | Nv | GCA GAA TTC GAA GTG TTG GAT TGT TCA C |
| NV3493R | Nv/Na | GTC TAA ACC CAG CTC ACG TTC C |
| NV3753RD | Na | TAA AAT TAA CCT ACT CCA ACA CAT |
| NV4196RD | Na | CCA AGC GGT CTC CCA TCT TAG |

*The *Eco*RI restriction site (underlined) was included to facilitate cloning.

Diagrammatic presentation of the attachment sites for these primers are shown in the chapters that follow.

## 2.5 Gene amplification using PCR

Target sequences were amplified in a GeneAmp PCR System 2400 thermal sequencer (Perkin Elmer). Thermal cycle amplifications were performed in 100 µl final volume using rTth-XL DNA polymerase and 3.3x XL buffer (Perkin Elmer) containing dNTPs at 200 µM, primers at approximately 0.5 µM, sample DNA (100 ng), and 2.5 units of rTth-XL DNA polymerase (Perkin Elmer). The "hot start" protocol for long PCR using paraffin wax beads was used according to the vendor's instructions (Perkin Elmer). The extension time for each PCR reaction was optimised for each expected product length; a longer product requiring a longer period of extension. Following an initial denaturation of 90 sec at $94^{0}$C, the reaction cycles consisted of denaturation for 10 sec at $94^{0}$C, primer annealing for 20 sec at $55^{0}$C and extension for 3 (product < 1kb) or 10 (product > 1 kb) minutes at $68^{0}$C. Cycles were repeated 30 times. The final cycle also included an extension of 10 min at $72^{0}$C. Negative controls, all the above reagents except template DNA, were included to screen for possible foreign DNA contamination. Positive controls were included to screen for host DNA contamination of the template.

## 2.6 Visulation of PCR amplification products

PCR-amplified products (5 µl aliquots) were separated and analysed by electrophoresis through a horizontal 1% agarose (Promega) gel containing 500 ng/ml ethidium bromide. Molecular markers, either lambda/*Hind* III or lambda/*Spp* I/*Eco* RI (New England Biolabs), were included to assess the size of PCR-amplified fragments. Electrophoresed PCR-amplified fragments and molecular markers were visualised using a ultraviolet transilluminator (UVP Inc.) and the gels photographed using a Polaroid MP4 Landcamera and Polaroid 667 film (ISO $3000/36^{0}$).

## 2.7 Procedures for molecular cloning and sequencing

The vector pBluescript® $SK^{+}$ plasmid (Stratagene) was prepared by linearisation with *Eco*RI or *Sma*I (Promega) and then dephosphorylated with calf intestinal phosphatase (Boehringer

Mannheim) according to the manufactures instruction. The PCR-amplified fragments were purified using the Wizard® PCR Preps system (Promega) and ligated into the prepared *Eco*RI or *Sma*I cloning site of the Bluescript® SK$^+$. Plasmids were transformed into *E. coli* strain TG1 by the heat shock method (Sambrook *et al.*, 1989). Clones with inserts were selected by their resistance to ampicillin and blue-white screening on X-gal/IPTG plates. Plasmid DNA was prepared by the alkaline lysis method (Sambrook *et al.*, 1989). Subclones for dye primer sequencing (Applied Biosystems) were generated using the Erase-a-Base® system (Promega). Sequence gaps were determined by ABI Prism® Dye Terminator Cycle Sequencing (Applied Biosystems). Sequencing was performed using an ABI 373A model sequencer (Applied Biosystems).

## 2.8 Sequence analyses

The sequence data obtained from the ABI sequencer was checked visually for sequence overlaps. Overlaps were marked on the chromatograms and the chromatograms ordered sequentially 5' to 3'. The completed sequences were entered manually into the sequence alignment program "DCSE" (Dedicated Comparative Sequence Editor version 2.54, De Rijk and Wachter, 1993). Sequence data for comparative analyses (sequence similarity match) was submitted by electronic mail (e-mail) to GenBank (National Centre for Biotechnology Information) using the proprietary software "BLAST". Sequence data was deposited into the GenBank database by e-mail using the proprietary software "BankIt".

Pre-aligned SSU and LSU rRNA sequences (Van de Peer *et al.*, 1998; De Rijk *et al.*, 1998), in DCSE format, were obtained for *Escherichia coli, Thermotoga maritima* Hubner *et al., Halobacterium marismortui* Ginzburg, *Sulfolobus acidcaldarius, Giardia muris* Grassi, *G. intestinalis, G. ardeae, Entamoeba histolytica, Physarum polycephalum, Trypanosoma brucei, Euglena gracillis, Tetrahymena thermophila* Nanney and McCoy, *Prorocentrum micans* Ehrenberg, *Plasmodium falciparum* Welch, *Cryptococcus neoformans* (Sanfelice) Vuillemin, and *Arabidopsis thaliana* Linnaeus. Information about secondary structure interactions is also contained in alignments from this SSU and LSU rRNA sequence database.

Secondary structure models were obtained, where possible, for the SSU and LSU rRNAs of the aforementioned species (Gutell, 1994a; Gutell *et al.*, 1993). Also, generic models for the SSU and LSU rRNAs of the eubacteria and eukaryotes were obtained. These generic models indicate conserved nucleotide-positions within the core of the rRNA subunits. Additionally, where the character-state of a conserved nucleotide-position is also conserved, the character-state of this position is given (Appendix 1 and 2; Gutell, 1994; Gutell *et al.*, 1993). The presence and character-state of conserved positions in the archaebacterial SSU and LSU rRNAs was determined from 83 SSU and 42 LSU pre-aligned rRNA sequences (Van de Peer *et al.*, 1998; De Rijk *et al.*, 1998). Using the above data, the character-state of highly conserved nucleotides for the eubacteria, archaebacteria and eukaryotes were added to the SSU and LSU rRNA sequence alignments of the aforementioned species.

Using the conserved nucleotide and secondary structure data as 'anchor points', the sequences of the *N. apis* LSU rRNA and the *N. vespula* SSU and LSU rRNA were aligned visually to the aforementioned species. As there are no introns in either the *N. apis* or *N. vespula* rRNA genes, their sequence is synonymous with the sequence of their rRNA subunits only requiring a thymine (T) to uracil (U) substitutions in alignments. From these alignments, and by comparison to existing secondary structure models, models were developed for the *N. apis* LSU rRNA and the *N. vespula* SSU and LSU rRNAs. Furthermore, a secondary structure model was developed for the *N. vespula* 5S rRNA by comparison of sequence data to the published generic 5S rRNA models (Wolters and Erdmann, 1986; Barciszewska *et al.*, 1996). The developed models were drawn manually using the program CorelDRAW™ (Corel Corporation Limited).

## 2.9 Phylogenetic analyses

The aligned sequence data stored in DCSE format was converted manually to "MASE+" format (Faulkner and Jurka, 1988) for analyses using the programs "SEAVIEW" and "PHYLO_WIN" (Galtier *et al.*, 1996). SEAVIEW is a sequence alignment program that also allows for the storage of additional information about the sequence alignment within the MASE+ data-file. In this regard, SEAVIEW is the precursor program for phylogenetic analysis using PHYLO_WIN. Unambiguous sequence positions for phylogenetic analysis were determined from secondary models of the SSU and LSU rRNAs (Gutell, 1994a; Gutell *et al.*, 1993) and recorded in the MASE+ data file using SEAVIEW. Phylogenetic analyses of the unambiguous rRNA alignment positions was conducted using the maximum parsimony, maximum likelihood and neighbor joining (with various distance corrections options) options of the PHYLO_WIN program.

| Isolate | Origin | Supplier |
|---|---|---|
| CN | Canberra, ACT, Australia | Dr Denis Anderson<br>CSIRO, Division of Entomology, Black Mountain, Canberra, ACT, Australia. |
| BB | Batemans Bay, NSW, Australia | Mr Noel Bingley<br>Sutton, NSW, Australia. |
| WA | Perth, WA, Australia | Mr Jeff Beard<br>Western Australia Department of Agriculture, Perth, WA, Australia. |
| KI | Kangaroo Island, WA, Australia | Mr Bruce White<br>Western Australia Department of Agriculture, Perth, WA, Australia. |
| NZ | Auckland, New Zealand | Dr Louise Malone<br>Hort+Research, Mount Albert, Research Centre, Auckland, New Zealand. |
| SW | Uppsala, Sweden | Dr Ingemar Fries<br>Swedish University of Agricultural Science, Bee Division, Uppsala, Sweden. |
| CA | Dawson Creek, BC, Canada | Mr John Gates<br>Agriculture Canada, Dawson Creek, BC, Canada. |
| AC | Bogor, Java, Indonesia | Didik<br>Bumi Ciluar Indah AI/1, Ciluar, Bogor, Java, Indonesia. |
| JV | Semarang, Java, Indonesia | Ministry of Forestry, Semarang, Java, Indonesia. |

Table 2.1

The identification, origin and supplier of each isolate of *N. apis* used in this thesis.

# CHAPTER 3

# LONG PCR OF RIBOSOMAL DNA REPEAT UNITS

## 3.1 INTRODUCTION

Saiki *et al.* (1985) first proposed the *in vitro,* enzymatic-amplification of specific DNA sequences using DNA polymerase, primers and thermocycling (PCR). The subsequent development of PCR (Mullis and Faloona, 1987; Saiki *et al.,* 1988) has provided a powerful, cost-effective tool for the amplification of specific DNA sequences. Today, PCR is widely used in many applications of science including molecular biology, evolutionary studies, forensic biology, and pathology (Cheng *et al.,* 1994a). However, the amplification of target regions > 5 kilobases (kb) has been found to be difficult and has limited the use of PCR for some tests (Cheng *et al.,* 1994b). Recent developments in PCR have been aimed at increasing the length and quantity of the amplified fragments. A number of factors, known or suspected, impose limits on the length of amplifiable product. These include: complete denaturation of target sequences; extension times sufficiently long to allow strand synthesis in each PCR cycle; protection of template DNA against damage (such as depurination) during thermocycling; and reduced efficiency in amplifying long sequence due to misincorporation of nucleotides (Cheng *et al.,* 1994a). Misincorporation of nucleotides may prematurely terminate strand synthesis (Huang *et al.,* 1992) and will affect sequences longer than 10 kb (Cheng *et al.,* 1994a). The use or addition of DNA polymerase with 3' to 5' proofreading activity will result in increased yields of longer products (Cheng *et al.,* 1994a). Cheng *et al.* (1994a) and Barnes (1994) demonstrated the practicality of long PCR (LPCR) on lambda clones that contained very large inserts. They were able to successfully amplify target sequences of 35 kb but LPCR has not been demonstrated for more complex eukaryotic genomes.

The rRNA genes of most cytoplasmic organisms are arranged in operons separated by a non-transcribed spacer, together forming the rDNA repeat unit. The repeat units occur in tandem arrays (Long and Dawid, 1980; Appels and Honeycutt, 1987; Reeder, 1990). However, the number of repeat units and the length of these units is highly variable. For example, the repeat unit length of *Bacillus thuringiensis* Berliner is approximately 5 kb whereas for *Mus musculus* Linnaeus it is approximately 44 kb (Appels and Honeycutt, 1987).

This chapter describes the use of LPCR in determining the size of the rDNA repeat unit from *N. vespula*, *N. apis,* and an as yet undescribed yeast. Furthermore, it describes the use of LPCR in determining the arrangement of the rDNA repeat units within these species.

## 3.2 Materials and methods

My initial attempts at LPCR amplification of rDNA repeat units were before commercial manufactured LPCR kits such as rTth XL™ (Perkin Elmer), Expand™ (Boehringer Mannheim), or Vent$_R$® (New England Biolabs) were available. Instead, a method was devised based the earlier work of Cheng *et al.* and Barnes. This method included, and was optimised for, the use of rTth DNA polymerase (Perkin Elmer) and Vent DNA polymerase (New England Biolabs) as the amplification enzymes. The rTth XL™ kit also uses these enzymes. For this reason, and because of the earlier work using rTth DNA polymerase and Vent DNA polymerase, preference was given to the rTth XL™ kit.

### 3.2.1 Isolation of genomic DNA template

Genomic DNA for the *Nosema* species was isolated as previously described (Sections 2.2, 2.3.1a, and 2.3.2). Genomic DNA from the undescribed yeast was provide as a gift from Dr Richard Jefferson, Centre For The Application Of Molecular Biology To International Agriculture (CAMBIA).

### 3.2.2 Selection of primers for LPCR

Primers for PCR amplification were synthesised and prepared as previously described (Section 2.4). The following five features were considered when choosing suitable priming sites: 1) a large, highly conserved region within the eukaryotic rDNA operon was required; 2) if possible, this region should be within the LSU rRNA gene in keeping with the focus of the thesis; 3) the primer annealing region should be sufficiently long as to allow the simultaneous annealing of two adjacent primers placed 'back to back' thus allowing for the amplification of a product spaning two adjacent rDNA repeat units; 4) the G + C content of the primers should be maximised; and 5) the primers should not form primer dimers.

An analysis of aligned LSU rRNA sequences (De Rijk *et al.*, 1998) and the eukaryotic LSU rRNA secondary structure model (Gutell, *et al.*, 1993) identified a conserved region encompassing helices 31a to 35a (Figure 6.4a) as the most suitable for priming sites. Previously obtained sequence data (Figure 6.2) indicated that a region from helix 31a to helix 34 was most suitable for *N vespula* and *N. apis*, as this region was well conserved between the two species. The available aligned LSU rRNA sequence data from 16 fungi (including yeast) (De Rijk *et al.*, 1998) indicated that the 5' regions of helices 31a and 35 were well conserved.

Primers for *N. vespula* and *N. apis*, NV1817F and NV1813R, were designed using the *N. vespula* sequence (Figure 6.2). These primers will amplify all except two nucleotides of the rDNA repeat unit. Oligonucleotide primer NV1817F (helix 35) is downstream from primer NV1813R (helix 31a). By choosing these primer annealing regions, the amplified fragment would span two contiguous rDNA repeat units but contain the sequence equivalent to one repeat unit. The primers also contain the *Bam*HI restriction site to aid cloning of single representative rDNA repeat unit fragments.

The yeast primers, SCrDNAF and SCrDNAR, were based on the sequence of the LSU rRNA of *S. cervisiae* (accession numbers J10355 and K01048). *S. cervisiae* was chosen as a representative yeast sequence for this highly conserved region. These primers will amplify all except 49 nucleotides of the rDNA repeat unit.

The primer, template, and primer sequences were as follows:

NV1817F. 5-AAT TCG GAT CC*A TGT ACT GGT TGA AGA CAA-3'

NV1813R AAT TCG GAT CCT AGT TCA CTG TGT TTC GGG

SCrDNAF GGT CTG ACG TGC AAA TCG A

SCrDNAR AGT TCA CCA TCT TTC GGG TC

*The *Bam*HI restriction site (underlined) was included to facilitate cloning.

### 3.2.3 Operon amplification using LPCR

*a. Before the availability of commercial LPCR kits*

LPCR amplifications (50 µl), overlaid with 50 µl of mineral oil (Sigma), were performed in GeneAmp® Thin-Walled Reaction tubes using an FTS-1 Thermal Sequencer

(Corbett Research). All components of the reaction buffer except the magnesium acetate were assembled on ice just before use. rTth DNA polymerase lacks the 3' to 5' proofreading exonuclease activity of Vent DNA polymerase which in the presence of Mg$^{2+}$ will degrade single stranded DNA such as primers. Therefore, the addition of magnesium acetate was delayed until the end of the first thermocycling step of cycle 1. The final reaction mix contained 20 mM Tricine pH9.0, 250 µm of each dNTP, 84 mM potassium acetate, 8% glycerol (v/v), 0.2% dimethylsulfoxide, 300 nM of each primer, 100 ng template, 1 unit of rTth DNA polymerase, 0.02 units of Vent DNA polymerase, and 1.76 mM magnesium acetate. The thermocycling parameters were as follows:

| | |
|---|---|
| Cycle 1 | Step 1 - hold at 80$^0$C for 90 seconds and then add magnesium acetate |
| | Step 2 - denature template at 94$^0$C for 1 minute |
| Cycle 2 to 30 | Step 1 - denature template at 94$^0$C for 10 seconds |
| | Step 2 - anneal primers and template at 55$^0$C for 20 seconds |
| | Step 3 - extension of primer at 68$^0$C for 10 minutes |
| Cycle 31 | Step 1 - extend at 72$^0$C for 10 minutes. |

*b. Amplification using the rTth XL™ kit (Perkin Elmer)*

LPCR amplification (100 ul) using the rTth XL PCR kit were performed according to the manufacturer's instructions in GeneAmp® Thin-Walled Reaction tubes. Thermocycling was carried out using a GeneAmp® PCR System 2400 thermal sequencer. Primer, template, and magnesium acetate concentrations were maintained as previously determined (Section 3.2.3a). This kit uses a two-buffer system separated by a paraffin wax interface that melts as the temperature of the reaction mixture approaches the first denaturation step, at which point the buffers mix. The two-buffer system performs the same function as adding the magnesium acetate after the first step of cycle 1 step 1 by keeping the Mg$^{2+}$-activated Vent DNA polymerase away from the primers until the first denaturation step.

### 3.2.4 Visualisation of the LPCR amplification products

Visualisation of LPCR products was performed as described previously (Section 2.6).

### 3.2.5 Restriction analyses of the LPCR amplification products

The DNA of *N. vespula* and the yeast yielded a single large (>7 Kb) product. To confirm the identity of the putative rDNA repeat unit of *N. vespula* 1 µg of the LPCR-amplified product was digested with *Eco* RI, and *Hind* III. The expected sizes of restriction fragments for these enzymes had been previously determined from the *N. vespula* rDNA repeat unit sequence (Figure 6.2). The identity of the yeast isolate was not known, however, initial restriction mapping of the LPCR-amplified product was performed using 1 µg of LPCR product and the enzymes *Dra* I, and *Kpn* I. The digested *N. vespula*, yeast, and undigested *N. apis* LPCR products were examined by electrophoresis (Section 2.6). Fragment sizes were determined by comparison to a lambda/*Hind* III and lambda/*Spp* I/*Eco* RI (New England Biolabs) molecular size markers.

### 3.2.6 Procedures for molecular cloning and sequencing

The LPCR-amplified fragments from *N. vespula* and the yeast genomic DNA were purified using the Wizard™ PCR Preps DNA Purification System (Promega). The termini of the purified LPCR-amplified products of the *N. vespula* and yeast DNAs were then directly sequenced (Section 2.7). They were also ligated into the predigested *Bam*HI or *Sma*I cloning site of the plasmid vector Bluescript® SK$^+$ (Stratagene) and following transformation into *E. coli*, were screened for inserts (Section 2.7).

### 3.2.7 Sequence analyses

DNA sequences obtained from *N. vespula* were visually compared (Section 2.8) to previously determined sequences (Figure 6.2), whereas, the yeast sequence data was submitted to the GenBank database for comparative similarity analysis (Section 2.8).

## 3.3 Results and discussion

### 3.3.1 LPCR amplification

A single fragment of 7.30 kilobases (kb) was amplified from *N. vespula* DNA and of 9.40 kb from yeast DNA using both the commercial and non-commercial methods (Figures 3.1B and E). The amplification of a single fragment from the DNA of these two species was possible over a wide range of magnesium acetate concentrations, although, yields were concentration dependent.

Amplification of the putative *N. apis* rDNA repeat unit proved to be difficult and unreliable. LPCR product was only obtained when using the commercially available rTth XL™ kit. Three fragments were amplified from *N. apis* DNA. As seen in figure 3.1H, the three fragments are estimated to be 11-12, 16-18, and much greater than 23 kb in length.

The primers NV1817F and NV1813R were designed to amplify the rDNA repeat unit of both microsporidians. Three fragments were amplified from the *N. apis* DNA, while only one fragment was amplified from the *N. vespula* DNA. However, it cannot be assumed that just because *N. vespula* and *N. apis* are taxonomically closely related (Baker *et al.*, 1995) that their rDNA repeat units will be of the same length. However, by comparing the *N. vespula* sequence data (Table 6.1) to the *N. apis* data for the SSU rRNA gene (Malone *et al.,* 1994), the ITS rDNA, and the LSU rRNA gene (Table 5.1), any additional repeat length that might exist must reside in the non-transcribed spacer. Other studies have shown that the microsporidians have unusual rDNA gene dispersion patterns along the chromosomes.

For example, recently it has been shown that the rDNA operons of the microsporidians *E. cuniculi*, *E. hellem*, and *E. intestinalis* do not occur within tandem arrays but are dispersed across several chromosomes (Peyretaillade *et al.*, 1998). In contrast, the rDNA of *N. bombycis* is located in one chromosome (Kawakami *et al.,* 1994). Not all microsporidians are as extreme. For example, the rDNA of *G. atherinae* is shown by hybridisation to occur in 8 of 16 chromosomes and for *S. lophii* 6 of 12 chromosomes (Biderre *et al.*, 1994).

The consistent single LPCR fragment obtained from *N. vespula* DNA suggests that only two rDNA repeat units occur on one chromosome similar to *N. bombycis* or that the rDNA repeat units of *N. vespula* are of equally size on all chromosome in which they occur. The three LPCR fragments amplified from *N. apis* DNA represents a pattern of dispersed rDNA repeat units not too dissimilar from those described for *E. cuniculi*, *E. hellem*, and *E. intestinalis*. For example, to amplify three LPCR fragments at least four rDNA repeat units would have to be present on one chromosome or alternatively six rDNA repeat units on three chromosomes. Also, that the non-transcribed spacer portion of each repeat unit differs in length.

The rRNA gene arrangements of the Microsporidia along chromosomes are highly unusual, but they are not unique among the protists. Several intracellular parasitic protozoans from the phylum Apicomplexa lack the typical tandem organisation of rRNA genes; for example *Theileria spp.* has two copies of the rRNA genes (Kibe *et al.,* 1994), while two to four copies occur in *Babesia spp.* (Dalrymple *et al.,* 1992), and four to eight copies occur in *Plasmodium spp.* (McCutchan *et al.,* 1995).

The question of origin for the three *N. apis* LPCR fragments remains unanswered. Each attempt at LPCR amplification of the *N. apis* rDNA repeat unit produced a similar amount of poor quality product. Consequently, neither cloning nor direct sequencing of any of the *N. apis* LPCR product was successful. A recent publication (Gatehouse and Malone 1998) supports the possibility that the middle sized fragment amplified may be the entire rDNA repeat unit of *N. apis.* Gatehouse and Malone determined that the approximate length of the rDNA repeat unit for *N. apis* was 18 kb, approximately equivalent in length to the second fragment amplified from *N. apis* DNA. Perhaps then, the remaining two fragments simply represent the presence of contaminating genomic DNA. Alternatively, these two fragments may represent strain variation in non-transcribed spacer length possible in a mixed population of *N. apis,* such as that collected from a beehive.

### 3.3.2 Restriction analysis

#### *a. N. vespula LPCR-amplified fragment*

Fragment lengths of the *Eco* RI and *Hind* III restriction digests of the *N. vespula* LPCR products (Figure 3.1C and D) agree with the expected sizes of fragments (Table 3.1) determined from sequence data (Figure 6.2). It was concluded that the entire operon of *N. vespula* had been successfully amplified.

#### *b. Yeast LPCR-amplified fragment*

Fungi have previously been shown to have rDNA with repeat units ranging in length from 9 to 42 kb (Appels and Honeycutt, 1987). Therefore, it was expected that the LPCR product from the yeast would be longer than the *N. vespula* product. The LPCR-amplified yeast product (Figure 3E) is approximately 9.1 kb in length. The length of the observed fragments for either the *Dra* I and *Kpn* I restriction enzyme digests (Table 3.2) indicate that the overall length of this product is greater than 7700 bp. However, this does not account for fragments that are smaller than 560 bp. Fragments less than 560 bp were not visible when electrophoresed because either the quantity of DNA in each putative fragment was insufficient to be seen under

UV trans-illumination or the fragments had migrated off the agarose gel.

### 3.3.3 *N. vespula* sequence analysis

The 5' and 3' ends of the *N. vespula* LPCR-amplified product were sequenced (Figure 3.2). Purified LPCR-amplified product was directly sequenced with the original amplification primers (NV1817F and NV1813R) using dye terminator sequencing (Section 2.7). Three hundred and fifty-two nucleotides of the plus strand and 402 nucleotides of the minus strand were determined. The regions sequenced exhibited 100% homology with previously determined sequence data for this region of the LSU rRNA gene (Figure 6.2).

### 3.3.4 Yeast sequence analysis

The 5' and 3' ends of the yeast LPCR-amplified product were sequenced (Figure 3.2). Purified LPCR-amplified product was directly sequenced with the original amplification primers (SCrDNAF and SCrDNAR) using dye terminator sequencing (Section 2.7). Four hundred and fourteen nucleotides of the plus strand and 451 nucleotides of the minus strand were determined (Figure. 3.2). This region extends from helix 24 to helix 46 of the LSU rRNA (eg. Figure 6.4a) but excludes 137 nucleotides located about the two primer annealing sites. These nucleotides were either within the 49 bp amplification gap just downstream from the primers or could not be resolved from the chromatogram.

Within the GenBank database (Section 2.8), the yeast *C. neoformans* returned the highest sequence similarity score of 96.89 and 88.33% for the plus and minus strand respectively. In comparison, the *S. cerevisiae* sequence has a similarity score of 92.01 and 64.42% respectively. The variation in sequence similarity between the first and second fragments is attributed to the region amplified. The second fragment overlaps the hypervariable D2 region (Figure 3.2, positions 57 - 299) and the variable helices 28, 29, 30, and 31 (Figure 3.2, positions 340 - 426) of the D3 region contained within the LSU rRNA.

A comparison of the similarity scores between *C. neformans* and the yeast, and, *S. cerevisiae* and the yeast, demonstrates a closer taxonomic relationship between *C. neoformans* and the yeast. Therefore, the high sequence similarity score of *C. neoformans* suggests that the undescribed yeast either belongs to or is closely related to the genus *Cryptococcus*.

## 3.4 Summary

Long PCR was used to determine the arrangement of the rDNA repeat units in *N. vespula* and *N. apis*. In the case of *N. vespula*, the result was highly successful. The identify of the fragment amplified from *N. vespula* DNA template was confirmed by both restriction digest mapping and direst sequencing. It can be concluded that the rDNA repeat units of *N. vespula* occur in tandem arrays similar to that of most other eukaryotes. The origins of the three amplified fragments from *N. apis* were not determined. However, at least one fragment was equivalent in length to that determined elsewhere for *N. apis*. In light of the unusual arrangement of rRNA genes observed in the Microsporidia, it seems likely that the amplified fragments of *N. apis* could represent dispersed rDNA operons. Two experiments could verify or discredit this possibility: 1) the LPCR fragments could be probed with specific rDNA probes; and 2) if these bands are shown by probing experiments to be of rDNA origin, the LCPR fragments should then in turn be used to probe the chromosomes of *N. apis*.

An extension of this study was the amplification of an rDNA repeat unit from an undescribed yeast using primers designed from conserved fungal sequences. This aspect was intended to show the utility of LPCR in molecular studies. Here, it was demonstrated that the primers required for LPCR could be derived from a consensus sequences of related species if only highly conserved regions of the target gene where considered. Also, it was demonstrated that it is possible to directly sequence LPCR product using the original amplification primers. Therefore, with the exception of hypervariable regions, LPCR and subsequent direct sequencing obviates the need to clone or subclone LPCR fragments before sequencing and thereby accelerating the process of data acquisition.

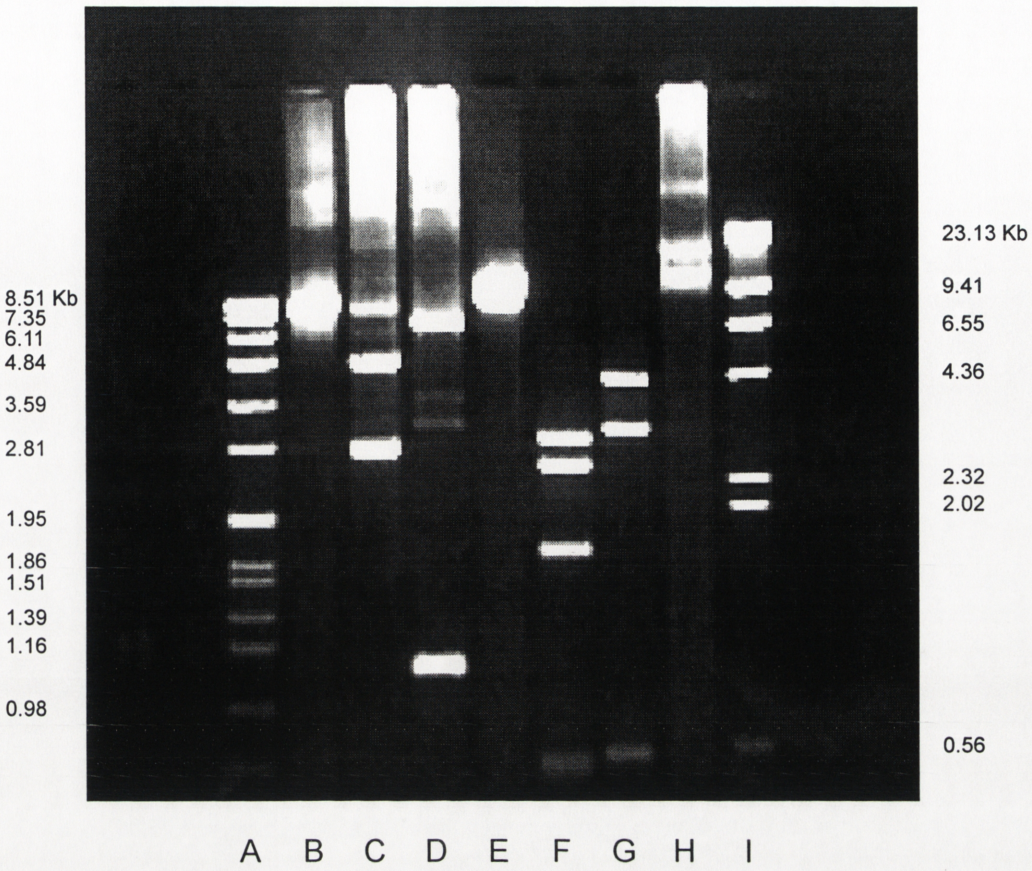


Figure 3.1

The restriction endonuclease mapping of LPCR fragments from *N. vespula* and an unidentified yeast, including the undigested LPCR products from *N. apis*. The LPCR products were amplified using rDNA operon-specific oligonucleotide primers for each species. The same primers were used for *N. vespula* and *N. apis* DNA.

A. Lambda/*Spp* I/*Eco* RI size markers - individual fragment sizes to the left of the gel.

B. *N. vespula* uncut LPCR product

C. *N. vespula* LPCR product digested with *Eco* RI

D. *N. vespula* LPCR product digested with *Hind* III

E. Yeast uncut LPCR product

F. Yeast LPCR product digested with *Dra* I

G. Yeast LPCR product digested with *Kpn* I

H. *N. apis* uncut LPCR product

I Lambda/*Hind* III size markers - individual fragment sizes to the right of the gel.

| Restriction_Enzyme | Expected Length | Observed Approximate Length |
|---|---|---|
| Uncut | 7292 bp | 7300 bp |
| Eco RI | 6286 | 6300 |
|  | 886 | 900 |
|  | 121 | not visible |
| Hind III | 4604 | 4600 |
|  | 2689 | 2600 |

Table 3.1

Restriction fragment lengths of endonuclease digested *N. vespula* LPCR- amplified product.

| Restriction_Enzyme | Expected Length | Observed Approximate Length |
|---|---|---|
| Uncut | not known | 9100 bp |
| *Dra* I | not known | 3100 |
|  |  | 2600 |
|  |  | 1600 |
|  |  | 500 |
| *Kpn* I | not known | 4300 |
|  |  | 3100 |
|  |  | 560 |

Table 3.2

Restriction fragment lengths of endonuclease digested yeast LPCR-amplified product.

Figure 3.2

Comparative sequence alignments for the unidentified yeast LPCR sequence, *C. neoformans* 'CryNeo' and *S. cerevisiae* 'SacCer'. Position 1 corresponds to the 771$^{st}$ nucleotide position of the *C. neoformans* sequence (GenBank accession number L14067) and the 482$^{nd}$ position of the *S. cerevisiae* sequence (J01355). Position 1 also corresponds to the 1$^{st}$ nucleotide position of helix 24 of the LSU rRNA (eg. Figure 6.4a).

```
LPCR Seq : ATGAAAAGCA CTTTGGAAAG AGAGTTAAAC AGTACGTGAA ATTGTTGAAA  50
CryNeo   : .......... .......... .......... .......... ..........
SacCer   : ........A. .....A.... .....G...A .......... ..........


LPCR Seq : GGGAAACGAT TGAAGTCAGT CATG-CTCTT AGGATTCAGC C-GT---TCT  100
CryNeo   : .......... .......... .G..T..A.. G..-...... .A..---...
SacCer   : .....GG.CA .TTGA....A ....-G.G.. TT.-.G.CCT .T.CTCC.TG


LPCR Seq : ---GC-GGTG TATTTCCTTT -GAGTGGGGT CAACATCAGT TTTGATCGAT  150
CryNeo   : ---..T.... ....C..... A..C--.... .......... .C......G.
SacCer   : TGG.TA..G. A..C..GCA. TTCACT...C ..G....... ....G.G.CA


LPCR Seq : GGATAAAGGC ACGAGGAAGG TAGCACTCT- CGG-GTGAAC TTATAGCCTC  200
CryNeo   : ......G... TG......T. .G.......T ...G...-TG ..........
SacCer   : .......TC. .T-.....T. ....TTG.CT ...TAA.-TA .........G


LPCR Seq : GCGTCATATA CATTGATTGG GACTGAGGAA CGCAGCATGC CTTTATGGCC  250
CryNeo   : CT...GC... ..C..G.... .......... T.....TC.. ..........
SacCer   : -T.GG.ATAC TGCCAGC... .........C T..GA.---- ----------


LPCR Seq : GGGATTCGTC CACGTACATG CTTAGGATGT TGACATAATG GCTTTAAACG  300
CryNeo   : ...G....C. .....T.GA. .......... .....A.... ..........
SacCer   : ---------- ---...--A. TCA......C ..G....... .T.A..TG.C


LPCR Seq : ACCCGTCTTG AAACACGGAC CAAGGAGTCT AACATATCTG CGAGTATTTG  350
CryNeo   : .......... .......... .......... .......... .....G....
SacCer   : G......... .......... :......... ...G.C.A.. .....G....


LPCR Seq : GGTGTCAAAC CCGAGTGCGC AATGAAAGTG AACGTAGGAG CGATCC--GC  400
CryNeo   : A......... T....C...A .......... ..T....... G.....--..
SacCer   : .....A.... ..ATAC...T .......... ........TT G.GG..TC..


LPCR Seq : AA--GGTGCA GCTTCGACCG ATCTGG---- ---------- ----------  450
CryNeo   : ..--..A... C......... ...C..ATCT TCTGTGATGG ATTTGAGTAA
SacCer   : ..GA...... CAA....... ...CT.ATGT CTTCGGATGG ATTTGAGTAA
                                          <-SCrDNAR


LPCR Seq : ---------- ---------- ---------- ---------- ----------  500
CryNeo   : GAGCATATAT GCTGGGACCC GAAAGATGGT GAACTATGCC TGAATAGGGC
SacCer   : GAGCATAGCT GTTGGGACCC GAAAGATGGT GAACTATGCC TGAATAGGGT
                                                   SCrDNAF->


LPCR Seq : ---------- ---------- ---------- ---------- ----------  550
CryNeo   : GAAGCCAGGG GAAACTCTGG TGGAGGCTCG TAGCGATTCT GACGTGCAAA
SacCer   : GAAGCCAGAG GAAACTCTGG TGGAGGCTCG TAGCGGTTCT GACGTGCAAA


LPCR Seq : ---------- ---TTGGGTA TANGGGCGAA AGANTAATCG AACCATCTAG  600
CryNeo   : TCGATCGTCG AAT....... ..G....... ...C...... ..........
SacCer   : TCGATCGTCG AAT....... ..G....... ...C...... ..........

LPCR Seq : TAGCTGGTTC CTGCCGAACT TTCCCTCAGG ATAGCAGAAA CTCGCATCAG  650
CryNeo   : .......... .......... .......... .......... ..........
SacCer   : .......... ........GU UU........ .........G ....T.....
```

```
LPCR Seq : TTTTATGAGG TAAAGCGAAT GATTAGAGGC CTTGGGGATG AAACATCCTT  700.
CryNeo   : .......... .......... .......... ........C. ....G.....
SacCer   : .......... .......... .........T TCC....TC. ...TGA....

LPCR Seq : AACCTATTCT CAAACTTTAA ATATGTAAGA AGTCCTTGTT ACTTAATTGA  750
CryNeo   : .......... .......... ..G....... ..CA.....C .........G
SacCer   : G......... .......... .......... .......... ..........

LPCR Seq : ACGTGGACAT GCGAATGA-G AGTTTCTAGT GGGCCATTTT TGGTAAGCAG  800
CryNeo   : ...A.CG... ........-. .......... .......... ..........
SacCer   : .......... TT......A. ..C..T.... .......... ..........

LPCR Seq : AACTGGCGAT GCGGGATGAA CCGATCGTGA GGTTAAGGTG CCGGAATATA  850
CryNeo   : .......... .......... .......... .......... ........C.
SacCer   : .......... .......... ....A...AG A......... ........C.

LPCR Seq : CGCTCATCAG ACACCACAAA AGGTGTTAGT TCATCTAGAC AGCAGGACGG  900
CryNeo   : .......... .......... .......... .......... ..........
SacCer   : .......... .......... .......... .......... ...C......

LPCR Seq : TGGCCATGGA AGTCGGAATC CGCTAAGGAG TGTGTAACAA CTCACCTGCC  950
CryNeo   : .......... .......... .......... .......... ..........
SacCer   : .......... .......... .......... .......... ......G...

LPCR Seq : GAATGAACTA GCCCTGAAAA TGGATGGCGC TCAAGCGTAT TACCCATACC 1000
CryNeo   : .......... .......... .......... ........G. ..........
SacCer   : .......... .......... .......... ........G. ....T....T

LPCR Seq : TCACCGTCAG CGTT  1014
CryNeo   : .......... ....
SacCer   : CT........ G...
```

# CHAPTER 4

# *Nosema apis* - INTRASPECIFIC COMPARISON

## 4.1 INTRODUCTION

*N. apis* is a host specific microsporidian parasite of the adult European honeybee *A. mellifera* (Bulla and Cheng, 1977), and is an endemic pest of honeybees throughout the world (Matheson, 1993). The ultrastructure and life cycle of *N. apis* have been extensively studied using electron microscopy (Fries, 1993), but these studies have revealed nothing about the genetic diversity of this organism. More specifically, it is generally assumed that all microsporidians obtained from honeybee and observed using the light microscope are *N. apis*. This, however, may not be so. General disagreement about the classification of Microspora based on ultrastructure and life cycles pervade the literature (Baker *et al.,* 1994; Baker *et al.,* 1995; and Fries *et al.,* 1996). A case in point is a microsporidian found to parasitise honeybee larvae (Buys, 1972, 1977). Based on ultrastructure inferred from electron micrographs, this microsporidian was identified as a new species. However, Clark (1980) reported that the spore size and developmental stages of this microsporidian indicated that it was a unique strain of *N. apis* rather than a new species. Therefore, the use of the light or the electron microscope for identifying closely related species or inferring phylogenetic relationships is of limited use.

Other examples of incorrect diagnosis of microsporidial species have been documented. Vossbrinck *et al.* (1993) discuss the case of the incorrect identification of three isolates of a microsporidial species obtained from several AIDS patients. From morphological studies these isolates were initially identified as *E. cuniculi*. Later, it was demonstrated by SDS-PAGE and Western blot analysis that these isolates, while identical to each other, were not *E. cuniculi.* They were subsequently assigned the name *E. hellem.* Vossbrinck *et al.* conclude that SDS-PAGE and Western blot analysis may not even be sensitive enough to distinguish between closely related species and that molecular techniques offer the most reliable method of identification.

In the absence of definitive morphological characters for species identification other techniques using molecular markers may greatly assist in the classification of the Microspora.

Riboprinting is one such technique. This technique uses restriction fragment length polymorphisms (RFLP's) of PCR-amplified ribosomal rRNA genes as molecular markers to reveal inter- and intraspecific variation in the target species. The successful application of this technique to determine the phylogenetic relationships among 13 microsporidian species clearly demonstrated its usefulness to determine relationships at the species level (Pomport-Castillon *et al.,* 1997). However, no intraspecific genetic variation was detected in the region amplified. Therefore, to detect intraspecific genetic diversity in the Microsporidia, it may be necessary to extend the riboprinting technique to other regions of the genome. Alternately, it may be necessary to use more discriminatory techniques such as DNA sequencing.

In this chapter I report a partial sequence of the rDNA operon for nine geographically distinct isolates of *N. apis* and compare these data to sequences of the Microsporidia *N. vespula* and *V. lymantriae* to infer phylogenetic relationships, using evolutionary tree building methods.

## 4.2 Materials and methods

### 4.2.1 Isolation of genomic template

Spores of *N. apis* and *N. vespula*, obtained *in vivo,* were recovered from host tissue and purified (Section 2.2). Genomic DNA was extracted from the spores of a Canberra isolate of *N. apis* and from *N. vespula* by the germination method (Section 2.3.1a, Section 2.3.2). Genomic DNA was extracted from another eight isolates of *N. apis* by the mechanical disruption method (Section 2.3.1b). In some cases, the sample of *N. apis* spores donated was insufficient to allow for a purification step before extraction of the genomic DNA.

The *V. lymantriae* sequence data (accession number L13330) was obtained from the GenBank database using the "ENTREZ" proprietary software (National Centre for Biotechnology Information). The *V. lymantriae* sequence data was aligned by eye to the *N. apis and N. vespula* sequences as previously described (Section 2.8).

### 4.2.2 Selection of primers for PCR amplification

In designing the primers for this experiment, a target sequence was chosen that contained regions likely to be evolving at different rates, a phenomenon known to occur in the cytoplasmic rRNA genes (Van De Peer *et al.,* 1993; De Rijk *et al.,* 1995). Regions that are highly conserved aid in the correct alignment of sequences, while regions that evolve more quickly are likely to show genetic variation useful for comparative analysis.

The target sequence chosen extended from nucleotides position 1151 to 1851 of the *N. vespula* sequence shown in figure 6.2. It encompasses the 3' end of the SSU rRNA gene, the ITS, and the 5' end of the LSU rRNA gene. Within the target region are the hypervariable ITS, helices 15 to 18, and helix 25 (Figure 5.3a). The target region is flanked by nucleotide sequences that are sufficiently conserved within the Microsporidia to allow Microsporidia-specific PCR amplification from heterogeneous genomic DNA.

The primers, used to amplify the target sequence, were based on the *N. vespula* sequence (Figure 6.2). Primers NV1161F and NV1851R (Section 2.4), were used to amplify the target region. These primers also contained *Eco*RI restriction sites to facilitate cloning of PCR product.

### 4.2.3 Amplification of target sequence

The target sequence was amplified using the long-PCR (LPCR) method (Section 3.2.3). This method was found to be more robust when amplifying from template extracted by the mechanical disruption method (Section 2.3.1b). When the mechanical disruption method was used to obtain template genomic DNA there was no attempt made to remove contaminates such as cell components and bacteria from the template. It was necessary to use this method of DNA extraction from spores when only a limited number of spores were available for some isolates. Despite the potential for the presence of contaminating host genomic DNA (Section 2.2), only the target sequences were amplified in all cases.

### 4.2.4 Visualisation of the LPCR amplification products

Visualisation of LPCR products was performed as described previously (Section 2.6).

### 4.2.5 Procedures for molecular cloning

The buffer containing the LPCR-amplified fragments was transferred from under the oil overlay to an Eppendorf tube. A 200 ul aliquot of sterile water was added to the buffer. The mixture was phenol/chloroform treated, chloroform extracted and DNA in the aqueous phase ethanol precipitated (Sambrook *et al.,* 1989). The LPCR-amplified fragments were digested with *Eco*RI (Promega) and ligated into the *Eco*RI cloning site of the plasmid vector Bluescript® SK$^+$ (Stratagene) and screened for inserts (Section 2.7).

### 4.2.6 Procedures for sequencing

Vector DNA was prepared by the alkaline lysis method (Sambrook *et al.,* 1989), and the insert was sequenced using a combination of dye primer sequencing (Applied Biosystems) and direct sequencing using ABI Prism® Dye Terminator Cycle Sequencing (Applied Biosystems) (Section 2.7). The primers used for direct sequencing (NV1584RD, NV1690RD, NV1629FD, and NV1373FD (Section 2.4)) were determined manually from the rDNA sequence of *N. vespula* (Figure 6.2). The accuracy of the sequence data was confirmed by sequencing both strands of each clone.

### 4.2.7 Sequence alignments and phylogenetic analysis

DNA sequences obtained from the ABI sequencer chromatograms were visually compared (Section 2.8) and the sequences aligned (Section 2.9). Unambiguous sequence positions were determined (Section 2.9) and analysed using maximum parsimony, neighbor joining (Jukes and Cantor distance), and maximum likelihood (Section 2.9).

## 4.3 Results and discussion

Nine geographically distinct isolates of *N. apis* were used in this study on the premise that genetic variation would be most likely be detected using isolates from different localities. However, this does not exclude the possibility that genetic diversity could also exist in one population of *N. apis* obtained from one colony of honeybees. This possibility was not investigated in this work.

### 4.3.1 Sequence length and variation

A region of approximately 665 to 676 basepairs (bp) was amplified, cloned and sequenced. Figure 4.1 depicts the cloned fragment and the position and direction of the primers used to sequence the fragment. The aligned sequences from the nine *N. apis* isolates, one *N. vespula* (NV) and one *V. lymantriae* (VL) isolate are shown in figure 4.2. A consensus sequence containing the predominant nucleotide character-state at each alignment position was also included in the alignment. By comparison to the consensus sequence in figure 4.2, genetic variation is present in all taxa. These results reveal the presence of sequence length variation amongst the PCR fragments obtained from the nine *N. apis* isolates. This length variation ranges from 663 bp for Kangaroo Island (KI) to 666 bp for Canberra (CN). The PCR fragment length for NV is 675 bp. Only 519 nucleotides of sequence were available in Genbank for this region of VL. Among the nine isolates of *N. apis*, most of the length variation was restricted to a region within the ITS and the hypervariable region of helix 16 (Figure 4.2 and 4.3). The length variation between NV and the nine *N. apis* isolates occurs in the ITS and helices 10, 16, and 28. More specifically, two sets of three nucleotide insertions and three single nucleotides insertions are present. The three-nucleotide insertions form internal loops in helices 10 and 28, as indicated in figure 4.3. The single nucleotide insertions occur in helices 10 and 18, and the hairpin of helix 20 (Figure 4.3). As shown in figure 4.2, NV and VL share most of the insertion and deletion events.

### 4.3.2 Nucleotide variation including transition and transversion analysis

In addition to the insertion/deletion events discussed, there are a number of other alignment positions that differ from the consensus sequence. For example, amongst the *N. apis* isolates, this variation ranges from no substitutions and two deletions in the Western Australian (WA) isolate to eight substitutions and three deletions in the *A. cerana* (AC) isolate. The variation between the consensus sequence and the sequences of NV and VL is far greater then that observed for the consensus sequence and the sequences of the nine *N. apis* isolates. For example, there are 50 positions (Figure 4.2) in which NV and VL share character-states but are different to the consensus sequence.

Amongst the sequences of the *N. apis* isolates there are 31 substitution events. Table 4.1 lists these events, table 4.2 summaries the character-state replacements, while figure 4.3 displays their location on the LSU rRNA secondary structure model for the region sequenced. Of these, 12 occur within secondary structure helices of the LSU, 14 occur in loops or hairpins of the LSU, and five occur in the ITS. Five of the 12 helical substitutions disrupt canonical bonds resulting in the formation of one A-A, one A-C, and three U-C base pairs. Four of the five disruptive substitutions are transversions (purine to pyrimidine and pyrimidine to purine). The remaining seven helical substitutions convert canonical bonds (3 C:G + 4 A:U) to either U:G or G:U bonds. These substitutions all represent transition events (pyrimidine to pyrimidine and purine to purine). The five substitutions within the ITS are three A to T transversions and two transitions, one C to T, and one A to G. The remaining 14 substitutions within loops and hairpins involve six transversion and eight transition events. In summary, these data demonstrate a 2:1 transition to transversion rate within the helices sequenced, a 4:3 transition to transversion rate within the loops and hairpins, and a 2:3 transition to transversion rate within the ITS. For character-state substitution possibilities in a 'random' system, transversions are twice as likely to occur as transitions. These data suggest that the process of substitution in the rDNA sequence may not be random. These data also suggest that the process of substitution may vary across the rDNA sequence. Finally, these data also indicate a possible preference of character-state for nucleotide replacement in the substitution process.

Two interesting comparisons to the presented results include riboprinting experiments performed on members of the genus *Trypanosoma* (Clark *et al.,* 1995) and *Entamoeba* (Clark and Diamond, 1997). First, analysis of 20 isolates of trypanosomes resulted in eight distinct patterns that corresponded to the recognised species. In instances where multiple isolates for one species were analysed no evidence of intraspecific variation was detected. However, intraspecific variation had previously been observed in isoenzyme patterns. Second, analysis of 87 isolates of amoebae of the genus *Entamoeba* resulted in 24 distinct patterns corresponding to previously described species. Intraspecific variation was found in three species and was such that it led to the grouping of like riboprints into 'ribodemes' ("populations of amoeba that share the same riboprint patterns"). Phylogenetic analysis based on parsimony and distance methods of the data suggested a deep divergence in this genus.

### 4.3.3 Phylogenetic analysis

Genetic variation, as demonstrated among the *N.* apis isolates, and the single NV and VL isolates, can be used as the basis for phylogenetic analysis. The phylogenetic analysis of these data is restricted to 501 positions because of the inclusion of VL data (Figure 4.2). Of the 501 positions, only 58 positions are phylogenetically informative of which 50 are shared exclusively by NV and VL. Therefore, the results obtained from the limited number of informative positions within the nine *N. apis* isolates should be interpreted with caution. It is difficult to determine which commonalities in the alignment data are due to homoplasy (convergence, parallelism, and reversals) or are plesiomorphs (ancestral character state). In these data, a sequence position is informative if two but not more than nine sequence positions share the same character-state, while among the remaining sequences, at least two share a different character-state.

The cladogram in figure 4.4a shows a consensus of the most parsimonious trees and requires 82 steps. Results of parsimony analysis suggest that the taxa can be divided into four distinct clades: 'Australia', consisting of KI, BB, CN, WA; 'Java', consisting of AC and JV; 'Vairimorpha', consisting of VL, NV and CA; 'International', consisting of SW and NZ. Bootstrap (Bs) analysis (1000 replicates) shows weak support for these clades. These grouping of taxa do however reflect some natural association.

The Australian clade for example consists of geographically distinct isolates from the mainland Australia and Kangaroo Island. Honeybees (*A. mellifera)* are an exotic species to Australia. The introduction of honeybees (and presumably *N. apis*) occurred about 1810 (Warhurst and Goebel, 1995). The number of introductions of bees since that time has been limited because of geographical isolation and quarantine regulations designed to keep Australia free of the many exotic pests and disease of bees. It is plausible (and most parsimonious) that all these *Nosema* isolates have descended from one original founder population of *N. apis,* although support for this is weak (Bs 7%). The few phylogenetic informative events that are seen would take place in a stepwise fashion: the divergence of the KI-BB group and the CN-WA groups from a common ancestor, and then the divergence of the sister taxa. The taxa of this group has since independently acquired lineage dependant substitutions. Alternatively, all the informative position could be homoplasies caused by some common selective pressure exerted on the Australian population of *N. apis*.

Similarly, honeybees are only a relative recent introduction to New Zealand, with the first introduction occurring in 1839 (Matheson, 1993b). New Zealand also has very strict quarantine controls and so the number of likely introductions of honeybees, and hence *N. apis,* would have been very limited, perhaps even more so than Australia. Therefore, given the geographic separation of Sweden and New Zealand, the sister relationship of SW and NZ in the International clade is almost certainly the results of homoplasy in the data. Their sister taxa relationship is supported only by the presence of a common T at position 142 (Table 4.1) that is not shared with any other taxa analysed.

The Java clade (AC and JV) indicates a true phylogenetic relationship between these two isolates despite only moderate support (Bs 75%). It is likely that the ancestor of these two isolates was a parasite of *A. mellifera* that subsequently 'species-jumped' and is now a parasite of *A. cerana*. This is not surprising as *A. mellifera* and *A. cerana* are very closely related and are believed to be in an immature stage of speciation (Ruttner and Maul, 1983). There are reports of *N. apis* being found in *A. cerana* (Singh, 1975; Yakobson *et al.,* 1992) but the accuracy of these reports, which were based on light microscopy, has been questioned as it may have been *N. ceranae* (Fries *et al.,* 1996). This is the first report based on molecular data that proves the presence of *N. apis* as a parasite of *A. cerana.*

The remaining clade presents two interesting results. First, the close taxonomic relationship (Bs 100%) of NV to VL instead of to *N. apis.* This sister-taxa relationship has been observed in phylogenetic trees constructed from SSU rRNA sequence data. Therefore, it is not surprising then that the sequence data from the LSU rRNA also supports this relationship. The data in figure 4.2 demonstrates 50 phylogenetically informative positions exclusive to NV and VL. Second, and somewhat more puzzling, is the closer association of CA to the NV-VL group (Bs 65%). This is likely to be the consequence of homoplasy as evident by the substitution events relative to the consensus sequence (plesiomorphic state) as shown in table 4.1. In this table, CA, NV and VL share two informative positions (Table 4.1, positions 343 and 460) plus a deletion (position 92). This compares to AC, NV and VL that share two informative positions (positions 153, 168) and BB, NV and VL that share one informative position (position 70). Other than position 360, that involve the transition from a C:G bond to a U:G bond, all the remaining positions are unpaired in the secondary structure and are in regions that appear to be evolving more quickly than adjacent sequence. For example, nucleotides at sequence positions 142 to 171 (Figure 4.3 helices 6 and 7) in these taxa have undergone a number of deletion events relative to the eukaryotic standard model.

Analysis of these data using a distance method (Jukes and Cantor), neighbor joining (Figure 4.4b) and maximum likelihood (Figure 4.4c) demonstrates the same branching pattern as seen in the parsimony analysis (Figure 4.4a) including similar bootstrap support. In particular, both of these analyses provide statistical support for the sister branching of NV and VL (100%). Additionally, these analyses indicate the relative distance of the NV-VL group from *N. apis* (maximum likelihood 0.1188 and jukes cantor 0.0566). This supports other findings based on SSU rRNA sequence comparisons that NV is more closely related to the genus *Vairimorpha* than *Nosema* (Baker *et al.,*1995; Malone and McIvor, 1996).

## 4.4 Summary

The data presented in this chapter demonstrates the presence of intraspecific variation in *N. apis* isolates and between *N. apis, N. vespula,* and *V. lymantriae*. Moreover, excluding the Swedish and New Zealand isolates, phylogenetic analysis grouped the remaining *N. apis* isolates and other taxa in apparently natural groupings based on their origin. It is also apparent that the *N. apis* isolates formed a coherent grouping distinct from *N. vespula* (NV) and *V. lymantriae* (VL).

The phylogenetic analysis supported findings by others that *N. vespula* is more closely related to the genus *Vairimorpha* than *Nosema.* This result was also strongly supported by comparisons in the sequence alignments. The Canadian isolate contained sufficient phylogenetically informative positions to cause it to group with the NV-VL group during tree construction. This result, based on sequence alignment comparisons, is an artefact caused by the limited number of phylogenetic informative positions. The sequence alignments and tree constructions support previous findings that *N. apis* has species jumped from *A. mellifera* to *A. cerana* (Singh, 1975; Yakobson *et al.,* 1992).

Excluding the NV-VL grouping the observed bootstrap for the maximum parsimony, neighbor joining and maximum likelihood analyses are low. These low bootstrap values are indicative of the small number of phylogenetically informative positions among the *N. apis* isolates. This region is not evolving at a rate that provides enough discriminatory information to allow for the positive identification of strains or isolates within *N. apis*. Therefore, other molecular markers are required. Perhaps, the sequence of the non-transcribed spacer situated between the LSU and SSU rRNA genes would be more suitable. Alternatively, mini- or microsatellites (small repetitive sequences $\leq$ 20 base pairs) may prove to be more informative and reliable as molecular markers.

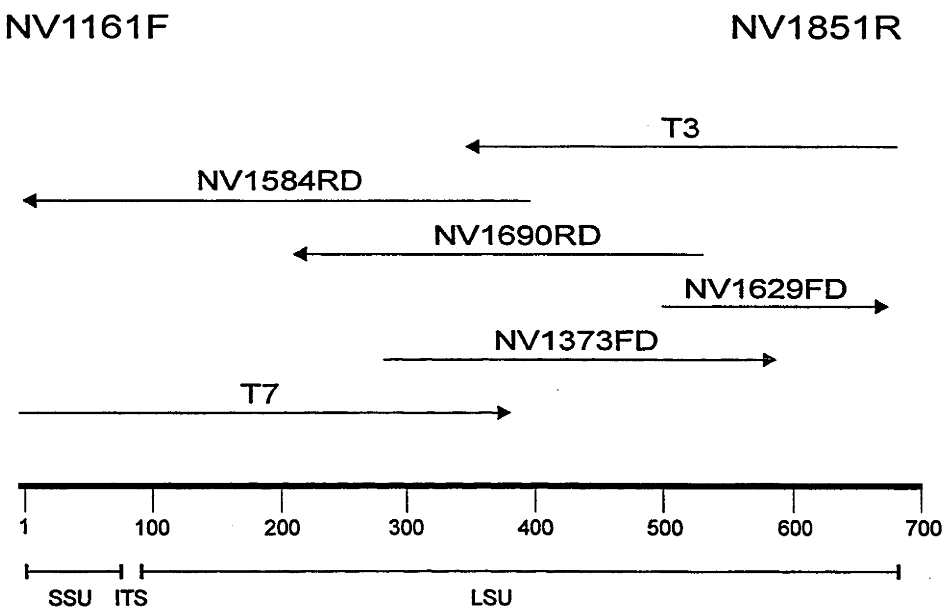


Figure 4.1

A map showing a typical clone used to determine the partial sequence of the SSU rDNA, the ITS and the LSU rDNA of nine isolates of *Nosema apis* and one isolate of *N. vespula.* The primers used to amplify the region (section 2.4) are shown above the map. The scale bar indicates the position of each primer. Each region determined by direct sequencing (section 2.7) is identified by the code of primer used (section 2.4). The regions determined by dye primer sequencing (section 2.7) of the parent clone (pBSII SK$^+$ plus insert) are indicated by their respective primer, either T7 or T3. Each arrow and its direction represent the size and the direction of the region sequenced. Position "1" indicates the first nucleotide position of the sequence in figure 4.2.

Figure 4.2

Comparative sequence alignments of a 676 base pair region from the rDNA operon of nine *N. apis* isolates, one *N. vespula* (NV) isolate, and one *V. lymantriae* (VL) isolate (*Vossbrinck* et. al., 1993). The first nucleotide position shown corresponds to the 1175th nucleotide position of the SSU of *V. necatrix* (Vossbrinck *et al.,* 1987). The alignments cover a region from the 3' end of the SSU rRNA gene, the ITS, and the 5' end of the LSU rRNA gene. The primers (Section 2.4), and their positions used to amplify and sequence the region, are also shown. Note, the region shown for primer NV1161F indicates the 3' half of the primer only. The region for the ITS, helices 10, 16 and 28, and their compliments helices 10', 16' and 28' are shown by the letter H.

The abbreviations for the isolates are: AC - Bogor, Java, Indonesia; JV - Semarang, Java, Indonesia; KI - Kangaroo Island, Australia; BB - Batemans Bay, Australia; CN - Canberra, Australia; WA - Perth, Western Australia; SW - Uppsala, Sweden; NZ - Auckland, New Zealand; CA - Dawson Creek, Canada; NV - *N. vespula* (ex D. Anderson); and VL - *V. lymantriae* (Acc. No. L13330). *Note that:* the JV, KI, BB, CN, WA, SW, NZ, and CA isolates were obtained from European honeybee colonies (*A. mellifera*), while the AC isolate was obtained from the Asian hive bee (*A. cerana*). The consensus sequence was obtained as described in the text (Section 4.3.1).

*Nosema apis*

```
AC        :  .......... .......... .......... .......... ..........50
JV        :  .......... .......... .......... .......... ..........
KI        :  .......... .......... .......... .......... ..........
BB        :  .......... .......... .......... .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .......... .......... .......... .......... ..........
NV        :  .......... .......... .......... .......... ..........
VL        :  .......... .......... .......... .......... ..........
Consensus :  ACAATATGTA TTAGATCTGA TATAAGTCGT AACATGGTTG CTGTTGGAGA
             NV1161F

AC        :  .......... .......... .......... ..--...... ..........100
JV        :  .......... .......... .......... ..--...... ..........
KI        :  .......... .......... .......... ..--...... ..........
BB        :  .......... .........T .......... ..--...... ..........
CN        :  .......... .......... .......... ..T......G ..........
WA        :  .......... .......... .......... ..--...... ..........
SW        :  .......... .......... .......... ..--...... ..........
NZ        :  .......... .......... .......... ..--...... ..........
CA        :  .......... .......... .......... .TT....... .-........
NV        :  .......... .........T ...T..AA.A .C.....G.. .-.....A..
VL        :  .......... .........T ...T..AA.A .T.....G.. .-.....A..
Consensus :  ACCATTAGCA GGATCATAAC GAAGAATTAC AAATTTTTTA GAATTAGTTT
                3′ end of SSU ←  Internal Transcribed Spacer

AC        :  .......... .......... .......... .......... ..........150
JV        :  .G........ .......... .......... .......... ..........
KI        :  .......... .......... .......... .......... ..........
BB        :  .......... .......... .......... .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... .T........
NZ        :  .......... .......... .......... .......... .T........
CA        :  .G........ .......... .......... .......... ..........
NV        :  A.AT...... .......... .........A .......... .T....G...
VL        :  -..T...... .......... .........A .......... .T....G...
Consensus :  TATATTTGCC CACACATGGG ATCAATAGGG TACCATAACG AGGAAGATCG
             → 5′ end of LSU

AC        :  ..A....... .......... .......... .......... .---......200
JV        :  .G........ .......G.G A......... .......... .---......
KI        :  .......... .......... .......... .......... .---......
BB        :  .......... .......... .......... .......... .---......
CN        :  .......... .......... .......... .......... .---......
WA        :  .......... .......... .......... .......... .---......
SW        :  .......... .......... .......... .......... .---......
NZ        :  .......... .......... .......... .......... .---......
CA        :  .......... .......... .......... .......... .---......
NV        :  ..AAA..... .......G.. ..T.C..... TA..AT..T. ......T...
VL        :  ..ATA..... ...NN..G.. ..T.C....N TA..AT.NT. ......T...
Consensus :  TAGCGGAATA CGAAAGATTA TTGATCGAAT ATATTAATAT ATATATAGAT
                                              H10        H10'
```

```
AC        :  .......... .......... .......... .......... ..........250
JV        :  .......... .......... .......... .......... ..........
KI        :  .......... .......... .......... .......... ..........
BB        :  .......... .......... .........G .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .......... .......... .......... .......... ..........
NV        :  A......... .......... .......... ......A... ..........
VL        :  A......... .......... .......... ......A... ..........
Consensus :  TACCCTTTGA ACTTAAGCAT ATCATTAAAA GGAGGAGAAG AAACTAACTA
                  NV1373FD

AC        :  .......... .......... .......... .......... ..........300
JV        :  .......... .......... .......... .......... ..........
KI        :  .......... .......... .......... .......... ..........
BB        :  .......... .......... .......... .......... .A........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .......... .......... .......... .......... ..........
NV        :  .......... ......G... .......... ..T....... ..........
VL        :  .......... ......G... .......... ..T....... .........C
Consensus :  GGATTTCTTT AGTAGCAGCG AGTGAACAAG AAACAACCCT TGATTGTAAT


AC        :  .T....-..- .......... .......... .......... ..........350
JV        :  .T....-..- .......... .......... .......... ..........
KI        :  ....-.-... .......... .......C.. .......... ..........
BB        :  ....-.-... .......... .......C.. .......... ..........
CN        :  ....-.-... .......... .......... .......... ..........
WA        :  ....-.-... .......... .......... .......... ..........
SW        :  .-....-... .......... .......... .......... ..........
NZ        :  ......-... .......... .......... .......... ..........
CA        :  ....-.-... .......... .......... .......... ..A.......
NV        :  ....A..... ..A....... ...T....C. ...AT....A ..A.....A.
VL        :  ....A..... ..A......C ...T....C. ...AT....A ..A.....A.
Consensus :  CCTTTACTGG AGCTGTAAAT CATATATTTT ATTTCTTATT TCGTAGAGGA
                H16

AC        :  ...-...... .......... ...-...... .......... ..........400
JV        :  ...-...... .......... ...-...... .......... ..........
KI        :  ...-...... .......... ...-...... .......... ..........
BB        :  ...-...... .......... ...-...... .......... ..........
CN        :  ...-...... .......... ...-...... .......... ..........
WA        :  ...-...... .......... ...-...... .......... ..........
SW        :  ...-...... .......... ...-...... .......... ..........
NZ        :  ...-...... .......... ...-...... .......... ..........
CA        :  ...-...... .......... ...-...... .......... ..........
NV        :  .T...A..T. ....GTTGAT ......G.AT ..TT.CT... ..........
VL        :  .T......T. ....CTTGAT ......G.AT ..TT.CT... ..........
Consensus :  TGTTATATCC GTTATAAATG AGAATATATA AAAGTAATTG AGTAGGGCTG
                                          H16'          NV1584RD
```

```
AC        :  .......... .......... .......... ....G..... ..........450
JV        :  .......... .......... .......... .......... ..........
KI        :  .......... .......... .......... .......... ..........
BB        :  .......... .......... .......... .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .......... .......... .......... .......... ..........
NV        :  .......... .......... .C........ .......... ..........
VL        :  .......... .........N .C........ .......... ..........
Consensus :  CTTGGTAGTG CAGTTTGAAT ATAGGTAGAA TGAGATATCT AAGGTTAAAT

AC        :  .......... .......... .......... .......... ..........500
JV        :  .......... .......... .......... .......... ..........
KI        :  .......... ...-...... .......... .......... ..........
BB        :  .......... .......... .......... .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .........T .......... .......... .......... ..........
NV        :  .........T .......... .......... .........T ..........
VL        :  .........T .......... .......... .........N ..........
Consensus :  ATAATGGTAC ACCGATAGCA AATAAGTACT GCGAAGGAAC TTGTGAAAAT
                        NV1629FD

AC        :  .....T.... .......... .......... .......... ..........550
JV        :  .....T.... .......... .......... .......... ..........
KI        :  .AT....... .......... .......... .......... ..........
BB        :  .......... .......... .......... .......... ..........
CN        :  .......... .......... .......... .......... ..........
WA        :  .......... .......... .......... .......... ..........
SW        :  .......... .......... .......... .......... ..........
NZ        :  .......... .......... .......... .......... ..........
CA        :  .......... .......... .......... .......... ..........
NV        :  ........GT .......... .......... .........G ..........
VL        :  ........GT ........C. .
Consensus :  GTGTGGGTTA TAGCCTTATT TTTAAGGACC CGTCTTGAAA CACGGACCAA

AC        :  .......... .......... ...---.... .......... ..........600
JV        :  .......... .......... ...---.... .......... ..........
KI        :  .......... .......... ...---.... .......... ..........
BB        :  .......... .......... ...---.... .......... ..........
CN        :  .......... .......... ...---.... .......... ..........
WA        :  .......... .......... ...---.... .......... ..........
SW        :  .......... .......... ...---.... .......... ..........
NZ        :  .......... .......... ...---.... .......... ..........
CA        :  .......... .......... ...---.... .......... ..........
NV        :  .......... ........A. .........T ..T..A.... .....AT...
VL        :
Consensus :  GGAGATTATA ATTATAGCGA GATAAAAACA ATGTAGTCGT TATTAGCTTG
                                      H28         H28'
```

```
AC        : .......... .......... .......... .......... ..........650
JV        : .......... .......... .......... .......... ..........
KI        : .......... .......... .......... .......... .........A
BB        : .......... .......... .......... .......... ..........
CN        : .......... .......... .......... .......... ..........
WA        : .......... .......... .......... .......... ..........
SW        : .......... .......... .......... .......... ..........
NZ        : .......... .......... .......... .......... ..........
CA        : .......... .......... .......... .......... ..........
NV        : .......... .......... .......... .......... ..........
VL        :
Consensus : ATAAGTTATA ATTATAAGAC CCGAAACACA GTGAACTATA CATGTTCTGG

AC        : .......... .......... ......676
JV        : ....G..... .......... ......
KI        : .......... .......... ......
BB        : .......... .......... ......
CN        : .......... .......... ......
WA        : .......... .......... ......
SW        : .......... .......... ......
NZ        : .......... .......... ......
CA        : .......... .......... ......
NV        : .......... .......... ......
VL        :
Consensus : TTGAAGATAA GCAACAGTTT ATTGGA
                      NV1851R
```

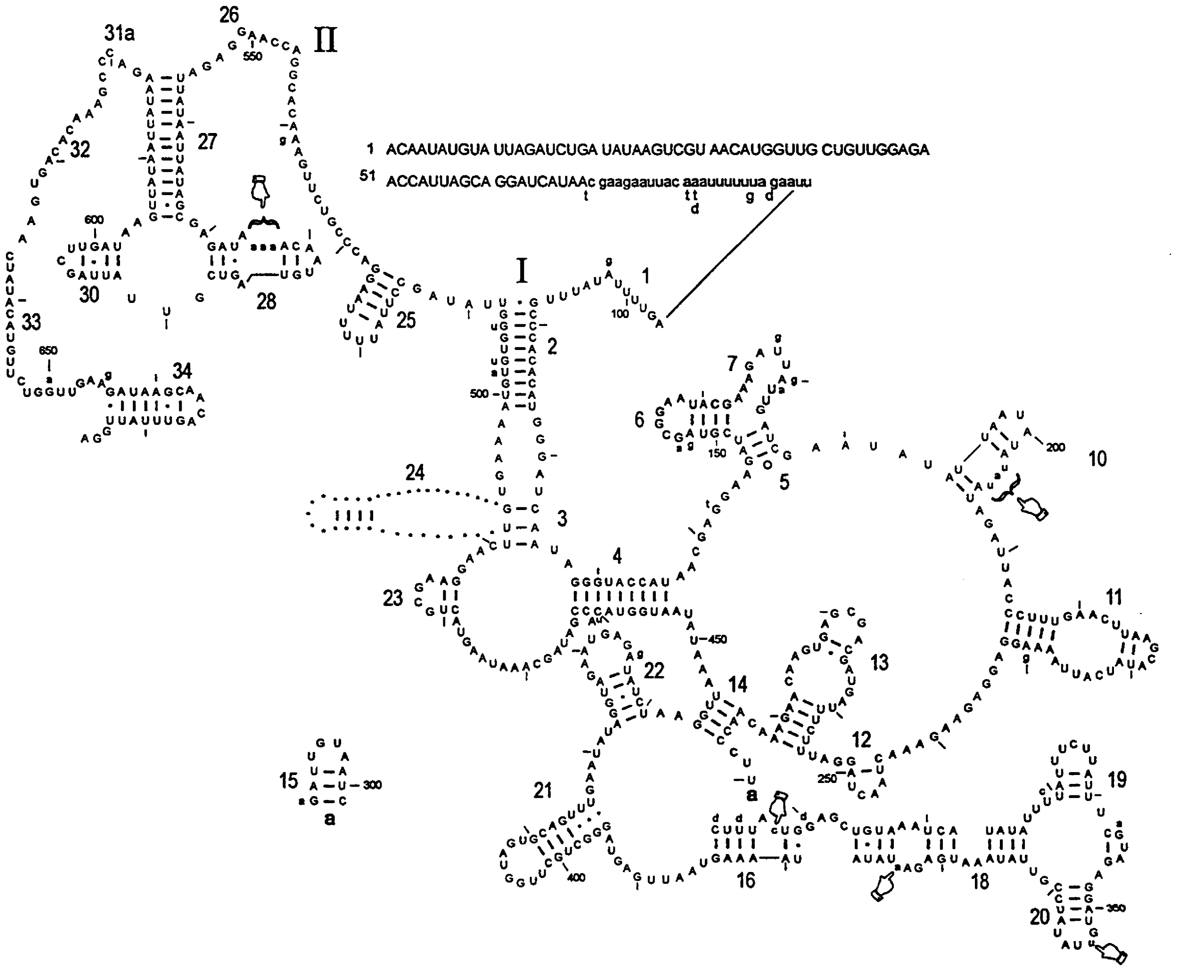

26
II
31a
32
27
I
1
2
25
28
30
33
34
1 ACAAUAUGUA UUAGAUCUGA UAUAAGUCGU AACAUGGUUG CUGUUGGAGA
51 ACCAUUAGCA GGAUCAUAAc gaagaauuac aaauuuuuua gaauu
7
6
5
10
24
3
4
23
11
22
14
13
12
15
a
21
19
16
18
20
550
600
650
500
450
150
100
200
250
300
350
400

Figure 4.3

A consensus partial secondary structure of the microsporidian LSU rRNA based on the alignment of nine *N. apis* isolates, one *N. vespula* isolate, and one *V. lymantriae* isolate. Nucleotide insertions (pointer), base substitutions (lower case), and deletions (d) are shown. Numbering of nucleotide positions is as for figure 4.1. Every 50th nucleotide is numbered while every 10th nucleotide has a stroke mark. Helix 15 (indicated by **a**) is drawn separate to the remaining consensus secondary model. The actual position of helix 15 within the secondary structure model is also indicated by **a**. Helix 24 is highly conserved in the eubacteria, archaebacteria and eukaryotes, and absent in these Microsporidia. In this diagram, the nucleotide positions normally occurring in helix 24 of eukaryotes are marked by *.

Table 4.1

Nucleotide positions within a region of 676 nucleotides of the rDNA operon from nine *N. apis* isolates, one *N. vespula* isolate (NV), and one *V. lymantriae* isolate (VL) are compared. Only nucleotide positions that have undergone a substitution event or loss within at least one of the nine isolates are considered in conjunction with the other two species. The abbreviations are : 'Seq Pos' - sequence position relative to figure 4.2; 'LSU Pos' - sequence position relative to figure 5.2; 'ITS' - indicates a nucleotide position within the internal transcribed spacer of the rDNA operon; 'consensus nucleotide' - indicates dominant nucleotide character state at a particular position; 'paired nucleotide' - indicates a bonded nucleotide within secondary model (Figure 5.3a); 'M' - indicates an absent nucleotide relative to the consensus nucleotide; and A, C, T, G - indicates nucleotide character state.

Abbreviations for the isolates as are as shown in figure 4.2.

Table 4.1

| Seq | LSU | Consensus | Paired | *Nosema apis* Isolate | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pos | Pos | Nucleotide | Nucleotide | AC | JV | KI | BB | CN | WA | SW | NZ | CA | NV | VL |
| 70 | ITS | C | | | | | T | | | | | | T | T |
| 82 | ITS | A | | | | | | | | | | T | C | T |
| 83 | ITS | A | | M | M | M | M | T | | | | T | | |
| 84 | ITS | T | | M | M | M | M | | M | M | M | | | |
| 90 | ITS | A | | | | | | G | | | | | | |
| 92 | ITS | A | | | | | | | | | | M | M | M |
| 102 | 8 | A | U | | G | | | | | | | G | | |
| 142 | 48 | G | | | | | | | | T | T | | | |
| 152 | 58 | A | U | | G | | | | | | | | | |
| 153 | 59 | G | | A | | | | | | | | | A | A |
| 168 | 74 | T | | G | | | | | | | | | G | G |
| 170 | 76 | A | | G | | | | | | | | | | |
| 171 | 77 | T | | A | | | | | | | | | | |
| 230 | 133 | A | U | | | | G | | | | | | | |
| 292 | 195 | G | C | | | | A | | | | | | | |
| 302 | 205 | C | G | T | T | | | | | M | | | | |
| 305 | 208 | T | | A | A | M | M | M | M | | | M | | |
| 310 | 212 | G | | M | M | | | | | | | | | |
| 328 | 230 | T | | | | C | C | | | | | | | |
| 343 | 245 | G | | | | | | | | | | A | A | A |
| 435 | 335 | A | | G | | | | | | | | | | |
| 460 | 360 | C | G | | | | | | | | | T | T | T |
| 501 | 402 | T | A | | | A | | | | | | | | |
| 502 | 403 | G | C | | | T | | | | | | | | |
| 505 | 406 | G | C | T | T | | | | | | | | | |
| 650 | 547 | G | | | | A | | | | | | | | |
| 655 | 552 | A | | | G | | | | | | | | | |

| Consensus Nucleotide | Replacement Nucleotide | Occurrence |
|---|---|---|
| A | G | 8 |
|  | T | 3 |
|  | C | 0 |
| G | T | 5 |
|  | A | 4 |
|  | C | 0 |
| C | T | 4 |
|  | A | 0 |
|  | G | 0 |
| T | A | 4 |
|  | C | 2 |
|  | G | 1 |

Table 4.2

A comparison of nucleotide substitution events based on the consensus sequence of figure 4.2.

Figures 4.4a to c

Phylogenetic trees constructed from a partial rDNA operon sequence alignment (Figure 4.2) of nine isolates of *N. apis*, one *N. vespula* (NV) isolate and one *V. lymantriae* (VL).

a. Maximum parsimony consensus tree.
501 sites (58 informative), 82 steps, 1000 bootstrap replicates

b. Neighbor joining tree using Jukes and Cantor distance measure.
501 site, 1000 bootstraps

c. Maximum likelihood tree.
501 sites, 200 bootstraps, maximum likelihood ln(L) -1084.573

Abbreviations for the isolates are as shown in figure 4.2.

Figure 4.4a

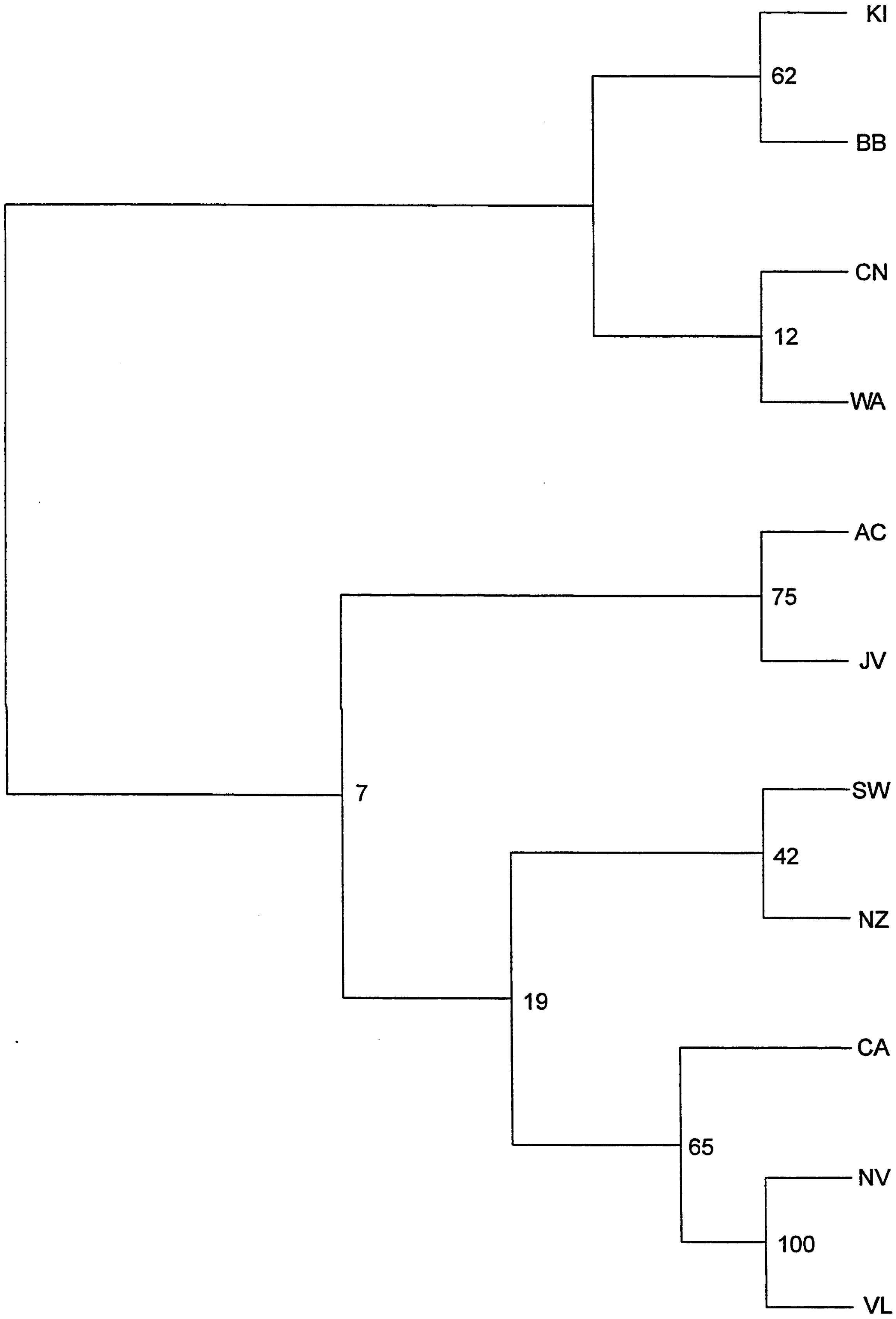
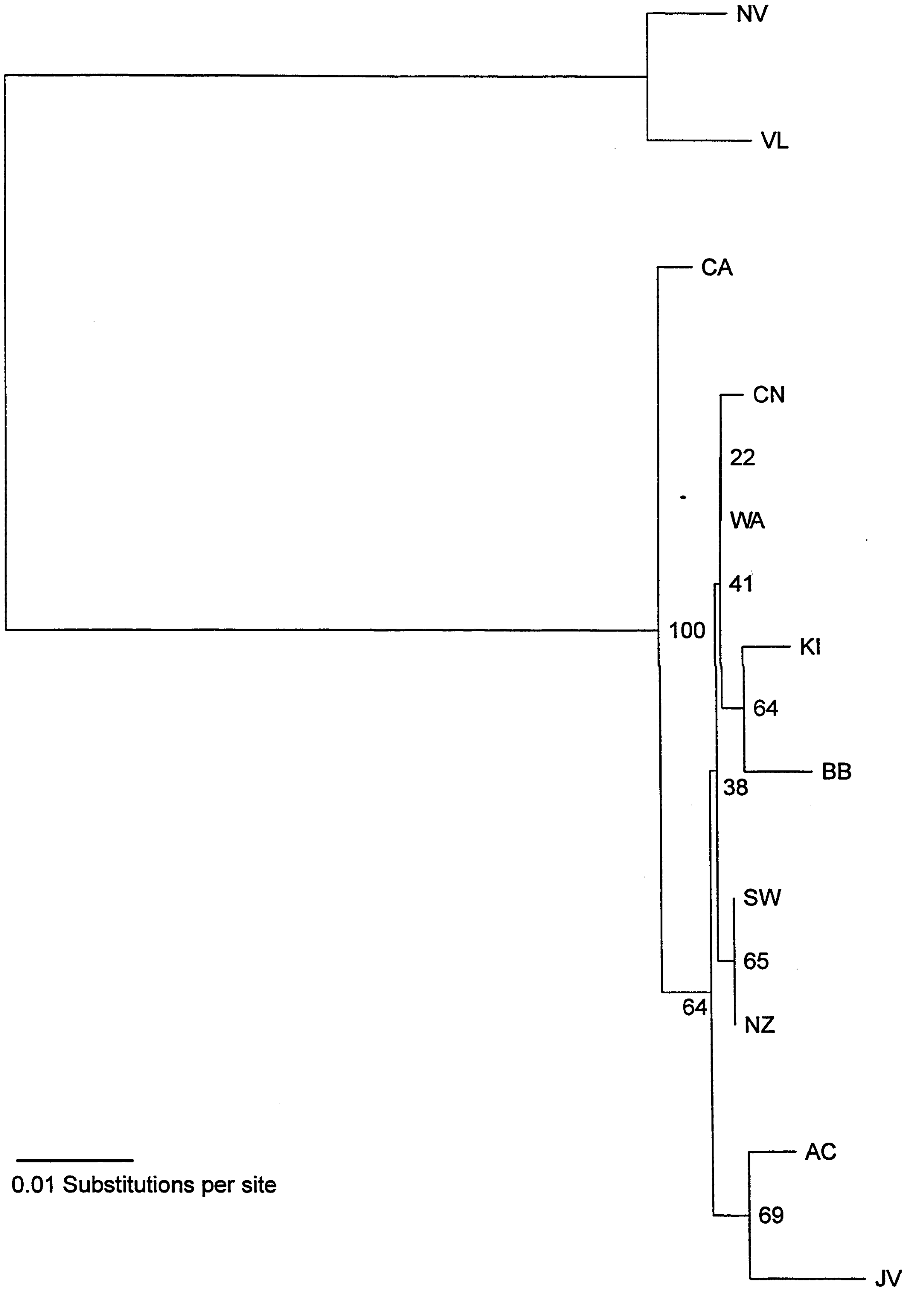


Figure 4.4b


NV
VL
CA
CN
22
WA
41
100
KI
64
BB
38
SW
65
NZ
64
AC
69
JV
0.01 Substitutions per site

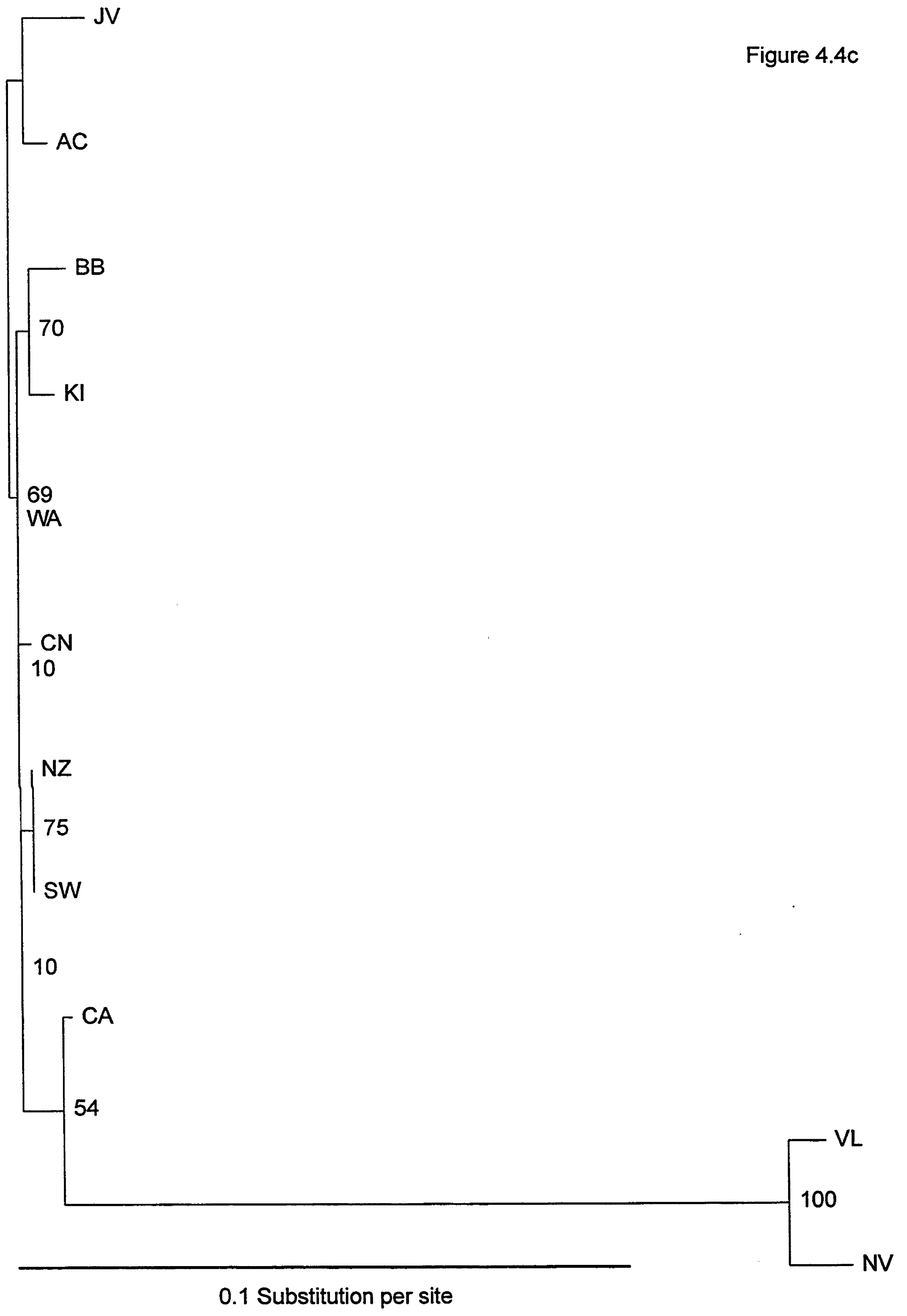

JV
AC
BB
70
KI
69
WA
CN
10
NZ
75
SW
10
CA
54
VL
100
NV
0.1 Substitution per site

Figure 4.4c

# CHAPTER 5

# *Nosema apis* - LARGE RIBOSOMAL SUBUNIT: SEQUENCE AND SECONDARY STRUCTURE

## 5.1 INTRODUCTION

In recent years, the view that the rRNAs serve merely as scaffolding ensuring the correct spatial assembly of ribosomal proteins has change significantly. Santer (1963) was the first to suggest that the rRNAs were involved directly in ribosome function. Woese (1980) subsequently put forward several arguments that involved the rRNAs as being the principal functional constituent of the ribosome. However, since these initial reports, knowledge about the highly conserved, functional activities of the rRNAs has rapidly increased (reviewed by Raué *et al.,* 1990; and Noller, 1991). Central to the function of the ribosome is the maintenance of the correct biochemical and biophysical interactions within the ribosome and between the ribosome and intracellular components. These interactions determine the functional integrity of the rRNA molecule that in turn is dependant on the conservation of the secondary and tertiary structure. In this context, models of the rRNA secondary structures are necessary as a basis for the determination of structure/function relationships in the ribosome (Ehresmann *et al.,*1990; Hill *et al.,* 1990).

Comparative sequence analysis (Fox and Woese, 1975), also known as the phylogenetic approach (Brimacombe, 1984), assumes that functionally equivalent regions of an RNA molecule will exhibit the same secondary and tertiary structures in all organisms despite primary sequence differences (Schnare *et al.,* 1996). Initially, secondary structure elements were determined by detecting compensatory changes in sequence alignments indicative of helical structures. For this purpose only canonical base-pairs (A:U and G:C) and U:G elements were considered. As available data has grown, so has the comparative justification for other nucleotide pairs, including those that do not exhibit typical Watson-Crick interactions. Comparative analysis has also identified tertiary interactions (Gutell *et al.,* 1994b). Direct techniques have been used to resolve the three dimensional structure and function of the rRNA subunits. Examples of these techniques include affinity labelling, cross-linking and footprint assays, chemical protection assays, directed mutagenises, and immunoelectron microscopy

(Ehresmann *et al.,* 1990; Hill *et al.,*1990).

Molecular biological techniques have been used since the 1960's to answer questions about the structure of the rRNA molecules. These techniques have also been used to study the evolutionary interrelationships of these molecules. Consequently, there is now an extensive collection of primary sequence data. The application of computer-aided comparative sequence analysis to these data (Schnare *et al.*, 1996 and references therein) has seen the number of defined interacting nucleotide positions grow. Furthermore, this approach has led to the detection of positional covariance (a nucleotide substitution event at one position leading to a compensational substitution event at a distant position) in alignments independent of the ability of the partner nucleotide to form canonical base-pairs in a helix.

The first, LSU rRNA secondary structure models proposed were for the 23S rRNA of *E. coli* (Glotz *et al.*, 1981; Branlant *et al.*, 1981; Noller *et al.*, 1981). These models were based on experimental as well as comparative data. The first secondary structure model published for the eukaryotic 26S rRNA was for yeast (Veldman *et al.*, 1981; Georgiev *et al.*, 1981). The yeast sequence was substantially longer than its *E. coli* counterpart but had potential to form a core secondary structure much like that of the bacterial 23S rRNA subunit (Veldman *et al.*, 1981). A series of studies published on the 23S-like rRNAs of yeast (Hogan *et al.*, 1984), rat (Hadjiolov *et al.*, 1984), mouse (Michot *et al.*, 1984) and *X. laevis* (Clark *et al.*, 1984) concluded that eukaryotic and bacterial 23S-like subunit share a common secondary structure core. Also, that the observed differences in sequence lengths of the 23S-like rRNAs could be accounted for as discrete variable blocks localised to specific regions within the structure.

Leffers *et al.* (1987) proposed a model for the 23S rRNA secondary structure of *Desulfurococcus mobilis* Stetter and Zillig based on comparisons with other archaebacteria. From this model and its nomenclature, a standard model for the 23S-like subunit was proposed based on phylogenetic sequence comparisons for eukaryotes (9), archaebacteria (7), and eubacteria/chloroplasts (12) (Höpfl *et al.*, 1989). Subsequently, as more phylogenetically diverse LSU rRNA sequences have become available, the model has undergone continuous refinement (Gutell and Woese, 1990, 1992a, 1993). There now exists a compendium of LSU rRNA secondary structure models. This compendium contains: LSU rRNA secondary structure models for a number of specific taxa; a universal LSU rRNA standard model; specific models for either eukaryote or eubacteria. (Gutell *et al.*, 1993).

This chapter records the sequence of the LSU rRNA of *N. apis* and fits this sequence to the established LSU model for eukaryotes.

## 5.2 Materials and methods

Sequence data for the LSU rRNA gene of *N. apis* was obtained from the Canberra isolate (CN) by the molecular techniques described in sections 2.3 to 2.7. This sequence was aligned to existing pre-aligned sequences, and subsequently a model for the secondary structure of the *N. apis* LSU rRNA was developed and drawn as described in section 2.8.

## 5.3 Results and discussion

### 5.3.1 LSU rRNA gene sequence

As depicted in figures 5.1a and b, two DNA fragments were amplified and sequenced. From the two overlapping clones of 667 and 1785 base-pairs (bp), a region of 2304 bp was determined (Figure 5.2). This however does not represent the entire LSU rRNA sequence of *N. apis* as depicted in figure 5.3a and b. The initial 93 bp of sequence defines 69 bp of the 3' end of the SSU rRNA gene and 24 bp of the internal transcribed spacer (ITS). The remaining 2211 bp of sequence includes all structural elements of the LSU rRNA up to and including helix 92. The secondary structure model presented in figures 5.3a and b is a composite of two *N. apis* LSU rRNA sequences; the sequence presented here for the region up to helix 92 and a partial sequence including the sequence for helix 93 onwards (Gatehouse and Malone, 1998). A length of 2554 bp for the LSU rRNA gene was determined from the putative LSU rRNA secondary model. No introns were identified in the LSU rRNA gene.

A comparison of the LSU rRNA gene sequenced here (2212 bp), with that recently published by Gatehouse and Malone (1998) shows the presence of 7 substitutions and 2 deletion events. These events have mostly occurred in unpaired structural loops or hairpins. The exception is a substitution event at position 1739 (Figure 5.3b) where the A in the A:U pair has been replaced by a G giving a G:U pair. Overall, the published sequence is 99.59% similar to the region of the LSU rRNA gene sequenced in this study.

Table 5.1 examines the nucleotide composition of the whole region sequenced including the additional rRNA gene sequence used to complete the LSU rRNA secondary model. The proportion of G + C nucleotides in the *N. apis* LSU rRNA gene is 35.31%. This is comparable to the G + C content determined for the SSU rRNA gene of other *Nosema* and *Vairimorpha* species; these varying from 33.9 to 37.3% (Gatehouse and Malone, 1998).

### 5.3.2 Internal transcribed spacer sequence

The LSU rRNA gene contains a covalently-linked 5.8S-like sequence in its 5' region (up to and including helix 10 of Figure 5.3a). Therefore, the rDNA operon of *N. apis* contains only one ITS similar to that seen in prokaryotes. The presence of a single ITS in the rDNA operon of the microsporidia was first observed in *V. necatrix* (Vossbrinck and Woese, 1986). The length of the ITS is estimated to be 24 bp. This was determined by estimating the putative 3' termini of the SSU and the putative 5' termini of LSU rRNA models for *N. apis.* Estimation of the 3' termini was by sequence comparison to the *V. necatrix* SSU rRNA model (Neefs *et al.*, 1991).

In comparison to higher eukaryotes, the ITS of *N. apis* is short. This feature is common in the amitochondrial protozoa. However, while the ITS of the amitochondrial protozoa is short, it exhibits a diverse nucleotide composition (Katiyar *et al.,* 1995). Usually the ITS of the amitochondrial protozoa exhibits one or two prominent nucleotides, for example, A in *Entamoeba and Trichomonas,* T in *Encephalitozoon*, and C in *Giardia.* As shown in Table 5.1, A (10) and T (9) are the two prominent nucleotides in *N. apis* representing 79.16% of the nucleotides present. Katiyar *et al.* (1995) examined the ITS sequences of amitochondriate protozoa for the presence of an RNA processing signal as might be indicated by the presence of a secondary structural element; no such element was apparent. They proposed a model in which the processing of ITS sequences occurred, because in contrast to mature RNA sequences, they were not protected by secondary structure or bound ribosomal proteins. Also, that they may possess RNases that preferentially recognise the predominant nucleotide(s) found in their ITS.

An examination of the aligned sequences for the ITS of nine *N. apis* isolates, and the single *N. vespula*, and *V. lymantriae* isolates (Figure 5.4a) demonstrates the presence of a potential secondary structure (Figure 5.4b) that may act as the signal for ITS processing in these species. These species are closely related phylogenetically (Baker *et al.,* 1995) and may share a common ITS processing mechanism. This putative secondary structure model is supported by the presence of a non-disruptive substitutions in the helix (G:U to U:G paired transversions) in *N. vespula* and *V. lymantriae,* while the remaining substitutions (including a two nucleotide insertion and one deletion) occurs in the closing loop and adjoining 5' and 3' sequences of the

ITS.

### 5.3.3 Comparison of the *N. apis* LSU rRNA

The rRNA subunits of the Microsporidia have been described as prokaryotic in size (Ishihara and Hayashi, 1968). The LSU rRNA from *E. coli* is 2,902 nts long (Guttel *et al.,* 1994b). At a length of 2554 nts, the *N. apis* LSU rRNA is the shortest eukaryotic LSU rRNA yet determined and is 348 nts shorter than the LSU rRNA of *E. coli.* Previously, the eukaryotic cytoplasmic LSU rRNA was known to vary in size from 2811 nts to 5185 nts for *G. muris* and *Homo sapiens* Linnaeus respectively.

As discussed previously, the cytoplasmic rRNAs of the prokaryotes and eukaryotes can be superimposed as they contain a conserved core interspersed by specific regions of variation. However, in the SSU of the microsporidian *V. necatrix,* conserved structural elements have been truncated or deleted (Neefs *et al.,* 1991). Cavalier-Smith (1993) suggested that the loss of these SSU structural elements in *V. necatrix* is a secondary loss perhaps due to an obligate parasitic lifestyle and that similar deletions may have occurred in the LSU. It was also suggested that similar losses or reductions may have occurred in the microsporidial LSU. Table 5.2 presents 15 examples from the *N. apis* LSU in proof of Cavalier-Smith's belief that secondary shortening of the LSU was likely to have occurred. This table, indicating mean and standard deviation (SD), demonstrates the degree of variability for 15 specific regions across the three evolutionary domains. In each case presented, the homologous region in *N. apis* is either absent, or shorter than the mean minus the SD for all the equivalent regions in the eukaryotes and almost all equivalent regions in the prokaryotes.

Schnare *et al.* (1996) have presented a comprehensive analysis of the eukaryotic cytoplasmic LSU including the identification of many eukaryotic-specific features. These features include non-canonical base-pairing (bp), nucleotide insertions, and nucleotide deletion. With respect to this publication and the current eukaryotic LSU model (Gutell, 1994a; Appendix 2a and 2b) I make the following comparisons to the putative LSU model (Figures 5.3 a and b) of *N. apis.*

### 5.3.4 A comparison of the *N. apis* and generic eukaryotic LSU rRNA secondary structure

#### *a Non-canonical base-pairs and other eukaryotic signature events*

Table 5.3 is a comparison of eukaryote characteristics and their equivalents in the LSU rRNA of *N. apis*. There are three variations that occur in *N. apis*: the non-canonical base-pair (usually C<>A) at positions 1636<>1642 has been replaced with a U-G pair; the insertion after position 1983 is absent reducing this loop from 7 to 6 nts as in the eubacteria; and a deletion is absent after position 1706 extending the junction loop to 13 nts as seen in the eubacteria.

*b Secondary structure comparisons*

**Domain 1**

*Positions 68 - 79*

This domain contains several interesting structural elements (Figure 5.3a). In eukaryotes, the 5.8S subunit base pairs with the LSU at helices 2, 4, and 10. Helix 7 (H7) in eukaryotes is located within the central core of 5.8S subunit. Helix 7 is slightly smaller in the eukaryotes than the prokaryotes, but is of a more constant size and structure in the prokaryotes (eukaryote mean 22.1 SD 1.4, archaebacterial mean 23.9 SD 0.3, and eubacterial mean 24.4 SD 0.5). Helix 7 of *N. apis* has been reduced to 11 nt that may base-pair to form a helix of 2 bp closed by a loop of 5 nt and adjoined to adjacent helices by single nucleotides.

*Positions 84 - 88*

Typically, helices 8 and 9 occupy this region of the prokaryotic and eukaryotic LSU. Exceptions occur in the LSU rRNA of the microsporidians including *N. apis* and *V. necatrix*, the diplomonads *G. muris* and *G. ardaea*, and the eubacteria *Pirella marina* Schlesner (Keulen *et al.*, 1992). Helix 9 is absent from some archaebacteria including the euryarchaeota *Halobacterium halobium* Petter, *H. marismortui*, and *Halococcus morrhuae* Elazari-Volcani (Gutell *et al.*, 1992).

*Positions 89 - 99*

In all eukaryotes, except the Microsporidia, this region is a discontinuous helix, helix 10. It is formed by base-pairing of nucleotides from the 3' end of the 5.8S subunit and from the 5' beginning of the LSU. The formation of a continuous helix/loop structure is otherwise specific to the prokaryotes that lack a separate 5.8S rRNA gene.

*Positions 192 - 290*

This is a region of hypervariability in which two small but conserved helical elements occur in both the prokaryotes and eukaryotes, helix 19 (H19) and 20 (H20). Helix 19 in both the pro- and eukaryotes usually consists of 5 bps closed by a loop of variable length. In addition, in eukaryotes there is a single nucleotide bulge between the 1st and 2nd nucleotides of the helix.

In both the prokaryotes and eukaryotes, H20 consists of 3 bps closed by a loop of 7 nts. In *N. apis* however H19 has been reduced to 3 bps closed by a loop of 5 nts. In comparison to the pro- and eukaryotes, helix 20 has been extended from 3 to 4 bps but is closed by a tetraloop instead of the usual 7 nts. By comparison to the eukaryotic model, this represents one of only two putative extensions found in the LSU of *N. apis*. The other likely extension occurs in helix 57 as will be discussed later.

*Positions 398 - 399*

In figure 5.3a there is a region marked by asterisks. The asterisks indicate the usual location of the highly conserved helix 24. The loss of H24 is of particular interest because of its usual extreme conservation in all three evolutionary domains. It is of an almost constant size (eukaryote mean 35.8 SD 0.6, archaebacterial mean 35.9 SD 0.9, and eubacterial mean 36.2 SD 0.8) and contains many highly conserved nucleotides. According to standardised eukaryotic and eubacterial models (Gutell, 1994a), 19 of the eukaryotic and 20 of the eubacterial positions exhibit character-state conservation in more then 90% of species for which sequence data is available. Yet in *N. apis* and other microsporidians H24 is absent. The absence of such a highly conserved structure and its effects on the tertiary rRNA-rRNA or rRNA-protein interactions would be a rich area for future research. This deletion is reminiscent of the deletions of helices 10, 11, and 44 in the *V. necatrix* SSU rRNA (Neefs *et al.,* 1991).

*Positions 413 - 428*

Helix 25 is a hypervariable region in eukaryotes of sometimes considerable length (mean 307.7, SD 171.0). However, in *N. apis* it is quite small at 5 bps closed by a tetraloop.

**Domain II**

*Positions 470 - 500*

This region is smaller in eubacteria than archaebacteria and smaller in prokaryotes than eukaryotes (eubacteria mean 34.0 SD 4.1, archaebacteria mean 63.7 SD 4.4, eukaryotes 94.3 SD 21.2). Within each evolutionary domain, this region contains three conserved helices, helices 28, 29, and 30. Among the eukaryotes, an additional, helix 31, may be present 3' to helix 30. Helix 28 is slightly more variable in the eukaryotes than the prokaryotes and does contain variable, internal, and opposing loops flanked on the 5' and 3' sides by 4 and 3 bps respectively. The closing loop of H28 is also variable in length and nucleotide composition. Helix 28 in *N. apis* is smaller consisting of 5 bps and a pair of opposing As between the 3rd and 4th nucleotide pair. Helix 29 is however highly conserved in the prokaryotes and eukaryotes consisting of 2 bps and a tetraloop. In the LSU of *N. apis,* helices 29, 30, and 31 have been replaced by a single helix loop structure. This structure is unlike H29 as it contains either 3 or 4 bps and is closed by a hairpin of either 3 or 5 nts respectively. The apparent replacement of H29, 30, and 31 with a single helix is another good example of degenerate evolution of the microsporidial LSU.

*Positions 685 - 800*

Within the three evolutionary domains this region forms helix 38 (H38) and in the eukaryotes an additional helix, helix 38a (H38a). Helix 38a is derived from the extension of an internal loop found in the prokaryotes, is usually ~37 to 38 nts long, and contains a single hairpin of approximately 8 to 10 bp. Exceptions to this number and arrangement of nucleotides in protists are found in the LSU of *Dictyostelium* (61 nts) and *Euglena* (84 nts). This region contains the discontinuity between *Euglena* rRNA species five and six (Schnare and Gray, 1990). Most of the unpaired nucleotides of H38a are located on the 3' side of the hairpin. Opposing the unpaired nucleotides of H38a is another internal loop of variable length and composition that occurs in all evolutionary domains. Typically, H38a and the opposing internal loop in the eukaryotes is preceded by a helix of 7 bps. To the 5' side of H38a and the 3' side of the opposing loop (positions 700 -755 on *N. apis* model, Figure 5.3a), H38 extends outwards consisting of approximately 30 bps, five small internal loops containing 2 to 4 nts, and at least one bulged nucleotide. In *N. apis,* H38a consists of 39 nts including a helix of 10 bps. However, the usual unpaired nucleotides to the 3' side of H38a have paired with the nucleotides of the opposing loop extending the base helix from 7 bps to 14 bps and reducing the opposing loop to a single nucleotide bulge. Also, it appears from a comparison of conserved nucleotide positions that H38 has been truncated by 7 bps. A similar pairing between typically unpaired nucleotides of the opposing loops is seen in the *Giardia* species. Here the base helix is extended from 7 to 12 bps with a corresponding reduction in the size of the opposing loops. However, the remainder of H38 has not been truncated in these species.

*Positions 1024 - 1036*

This region is divergent in length (mean 45.7 SD 40.8) and nucleotide composition in the eukaryotes, however, it usually forms a single helix (of at least 5 bps) and hairpin structure, helix 45. Helix 45 of *N. apis* is short and contains a helix of either 3 or 4 bps. There is also an additional nucleotide pair between the 2$^{nd}$ and 3$^{rd}$ bps that may form a non-canonical pairing of the form U<>U. The helix is closed by a loop of either 3 or 5 nts respectively.

*Positions 1125 - 1134*

This region forms a highly conserved helix/hairpin structure (helix 47) in prokaryotes (Table. 5.2). This structure is more variable in length in the eukaryotes (mean 21.0 nts SD 6.2 nts). For example, in *E. gracilis* the structure contains 48 nts while in *G. muris* there is only 12 nts. Helix 47 of the *N. apis* is truncated to 3 bps closed by a tetraloop.

**Domain III**

*Positions 1152 - 1167*

The nucleotide arrangement in this region is highly conserved in both the prokaryotes and eukaryotes. The conserved model for this region (Helix 50) consists of 6 bps, two opposing internal loops of 3 (5' side) and 1 (3' side) nts, 2 bps and a closing loop of 5 nts. In *N. apis,* the terminal 2 bps and the closing loop have been replaced by a tetraloop while the base helix of 6 bps remains.

*Positions 1178 - 1181*

This region has a conserved base structure (helix 52). This structure consists of two helices each containing 2 bps. These helices are separated by two opposing loops: the first (5' side) contains 3 nts and the second (3' side) containing 4 nts. The region then extends outwards forming two helix/hairpin structures at right angles to each other. The first of these structures usually contains approximately 3 to 6 bps, while the second is hypervariable in length. The number of nucleotides contained within this variable structure ranges from 41 to 147 nts. The largest variation seen in the variable structure occurs in the Protista. The *Giardia* species are included among the protista that have more than 41 nts within this variable region, in contrast to sharing similar reductions in many regions noted for the LSU of *N. apis.*

*Positions 1215 - 1349*

The character-state of the nucleotides within this region is hypervariable. This region however, can usually be modelled around a set of core helical elements, helices 54 to 59. Of these helices, H57 is the most highly conserved in the eukaryotes. The usually arrangement of this helix is 6 bps in which the 3$^{rd}$ nucleotide pair is not bonded. The helix is usually closed by a loop of 9 nts. Comparative evidence of paired nucleotide replacement (Figures 5.5a,b and 5.5c,d) for *N. apis* and *N. vespula* suggests that the number of paired nucleotides in H57 has been extended to 8, while the loop contains 10 nts (Figures 5.5b,d). This is the second example of a putative extension of a helix-hairpin structure that can be attributed to base-pairing in the microsporidial LSU. Extension of this helix also occurs in the crenarchaeota, but the closing loop usually contains fewer nucleotides. For example, *S. solfataricus* has a helix of 7 bp closed by a loop of 4 nts while *D. mobilis* has a helix of 11 bp with an internal loop and is closed by a loop of 4 nts.

**Domain IV**

*Positions 1456 - 1478*

This region (helix 63) is hypervariable in length yet in many cases can be modelled as one or two helices. This region in eukaryotes has a mean length of 215 nts and a SD of 149.7. It is the second most variable region within the eukaryotic LSU after H25. This helix in the *N. apis* LSU is 8 bps long (including a eukaryotic-specific non-bonded C<>A pair) and is closed by a tetraloop. The *N. apis* helix is dwarfed by the same region in the LSU of *Homo sapiens* (663 nt) but is similar in length and structure to that of the *Giardia* species.

*Positions 1540 - 1661*

This region is extremely conserved in the prokaryotes and eukaryotes. In the eukaryotes, 87% of the nucleotide positions maintain the same character-state in at least 90% of all sequenced eukaryotic LSU rRNAs. Helix 68 occurs within this region. In the *N. apis* LSU, this helix has been truncated from 55 to 46 nts. Helix 68 (Figure 5.3b) is structurally identical to the standard eukaryotic model up to and including bp 13 (positions 1563:1576). At this point there is an additional single base-pair and a closing loop of 11 nts. In the eukaryotic model, after base-pair 13, two opposing internal loops are present, each containing 4 nts. Beyond the internal loops is a helix of 5 bps closed by a tetraloop. The first 3 (5' side) and the last 2 (3' side) nts of the closing loop in *N. apis* possess the same character-state as the first 3 (5' side) and the last 2 (3' side) nts of the opposing internal loops in the eukaryotic model. Therefore, it appears that a deletion event(s) has removed the remaining conserved helix and tetraloop. The remainder of the region is identical to the eukaryotic model.

**Domain V**

*Positions 1812- 1813 and 1848 - 1856*

Both of these are variable in length in the eukaryotes but usually form helical-loop structures identified as helices 78 (mean 15.7, SD 15.9) and 79 (mean 119.8, SD 68.0). Helix 78 has been deleted in the *N. apis* LSU while helix 79 has been truncated to a loop of 7 nts.

The remaining secondary elements of the *N. apis* LSU conform to the eukaryotic model. The eukaryotic model also contains several known tertiary interactions most of which appear in the putative *N. apis* model.

*c Tertiary interactions comparisons*

The LSU models of Gutell *et al.* (1993) incorporate comparatively inferred tertiary interactions (Leffers *et al.*, 1987; Haselman *et al.*, 1989; Guttel and Woese, 1990; Larsen, 1992; Gutell *et al.,* 1994). Several of these have been experimentally confirmed (references cited: Ryan and Draper, 1991; Kooi *et al.*, 1993; Aagaard and Douthwaite, 1994; Rosendahl *et al.,* 1995). These interactions are indicated in figure 5.3a and b of the *N. apis* rRNA secondary structure. The nucleotide(s) involved in tertiary interactions are linked by a line. Of the possible interactions, two are apparently absent in *N. apis*. These interactions normally occur between two pair of nucleotides located approximately at positions 243 and 257 (Figure 5.3a) and between single nucleotides located approximately at positions 1802 and 1823 (Figure 5.3b).

## 5.4 Summary

The LSU of *N. apis* contains many features that identify it as eukaryotic. It is also comparatively very small when all three evolutionary lineages (eubacteria, archaebacteria and eukaryotes) are considered. This can, in many instances, be accounted for by the loss of, or reduction in, the hypervariable regions. These losses by themselves however, do not account for the total reduction in size of the *N. apis* LSU. The evolution of the LSU and SSU of the microsporidians are apparently degenerate, with the loss also of universally conserved features found in other cytoplasmic rRNAs. It also appears that at least two tertiary interactions have been dispensed with compared to the eukaryotic model. It cannot be ascertained from the data presented here if these losses are a consequence of an obligate parasitic lifestyle, as has been suggested. Further work is required to determine this question and the implications of such conserved losses on the function of the ribosome. These results do however present an interesting paradox. Is this data evidence for the very early divergence of the Microsporidia, in which the deletions present have occurred slowly over time, or are they the result of punctuated evolution? The remaining chapters of this thesis investigate this question further and in doing so extend the analysis of the microsporidian rRNA.

Figure 5.1 a and b

Maps showing the overlapping clones used to determine the partial sequence of the LSU rRNA gene from *N. apis*. The primers used to amplify each region are shown above each map. The scale bar indicates the position of each sub-clone or each region determined by direct sequencing. Each region determined by direct sequencing is identified by the primer used (Section 2.4) or is otherwise identified by the name of the sub-clone. Each arrow and its direction represent the size and the direction of the LSU rRNA gene insert sequenced as a sub-clone or the results of direct sequencing. Position 1 indicates the first nucleotide of the LSU rRNA gene.

a. A region of the SSU rRNA gene, ITS and LSU rRNA gene encompassing the 1150th SSU to the 500th LSU rRNA gene nucleotides.

b. A region of the LSU rRNA gene encompassing the 400th to 2200th nucleotides.

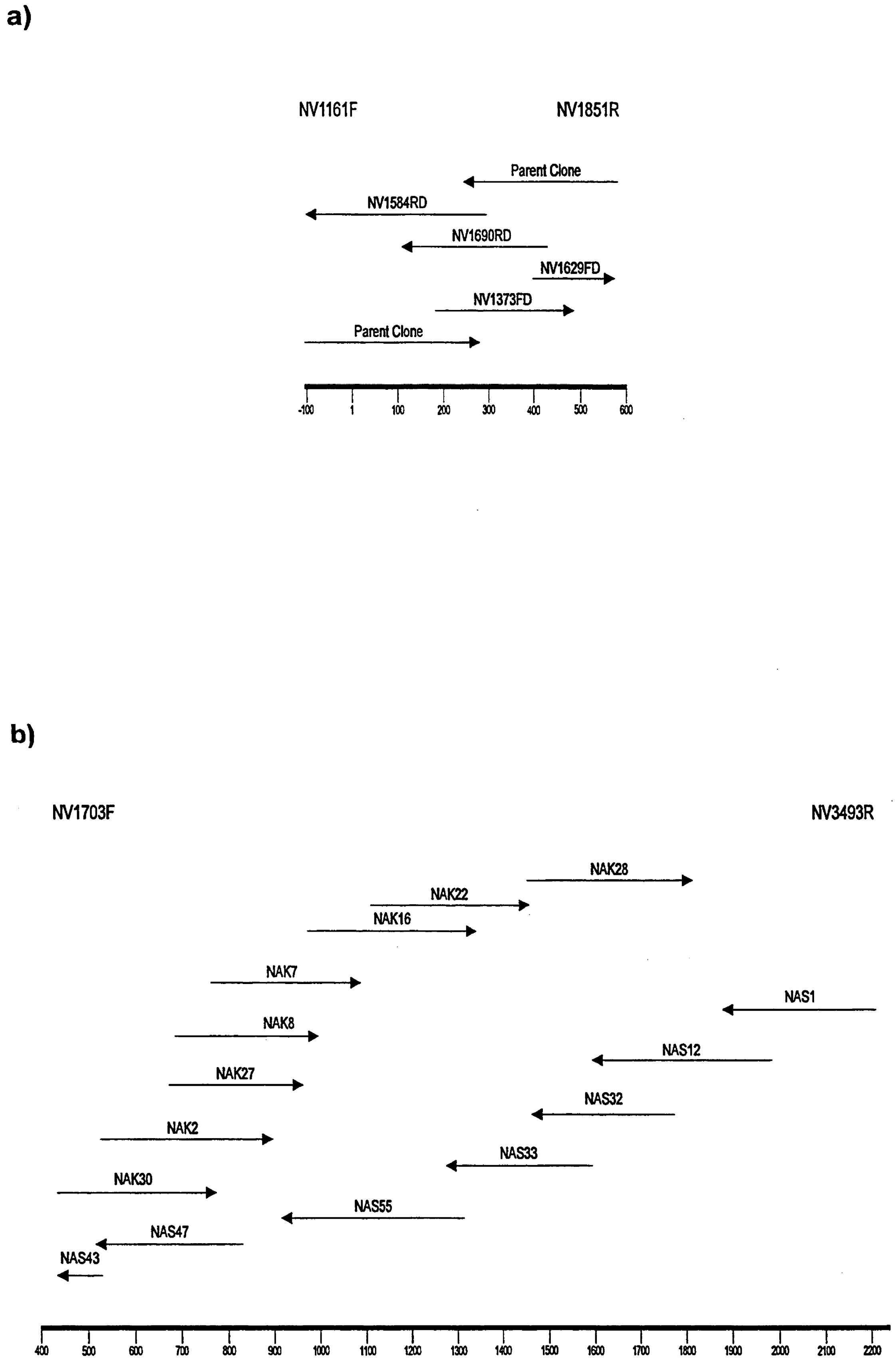

a)
NV1161F
NV1851R
Parent Clone
NV1584RD
NV1690RD
NV1629FD
NV1373FD
Parent Clone
-100
1
100
200
300
400
500
600
b)
NV1703F
NV3493R
NAK28
NAK22
NAK16
NAK7
NAS1
NAK8
NAS12
NAK27
NAS32
NAK2
NAS33
NAK30
NAS55
NAS47
NAS43
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200

Figure 5.2

The sequence of LSU rRNA gene including the 3' end of the SSU rRNA gene and the ITS of *N. apis*. As determined from the secondary structure models: Position 1 corresponds to the 1175th nucleotide position of the SSU rRNA sequence of *V. necatrix* (Vossbrinck *et al.*, 1987). Position 70 corresponds to the 1st nucleotide position of the internal transcribed spacer rDNA sequence. Position 96 corresponds to the 1st nucleotide position of the LSU rRNA gene.

*Nosema apis*

```
   1  ACAATATGTA  TTAGATCTGA  TATAAGTCGT  AACATGGTCG  CTGTTGGAGA
  51  ACCATTAGCA  GGATCATAAC  GAAGAATTAC  AATTTTTTTG  GAATTAGTTT
 101  TATATTTGCC  CACACATGGG  ATCAATAGGG  TACCATAACG  AGGAAGATCG
 151  TAGCGGAATA  CGAAAGATTA  TTGATCGAAT  ATATTAATAT  AATAGATTAC
 201  CCTTTGAACT  TAAGCATATC  ATTAAAAGGA  GGAGAAGAAA  CTAACTAGGA
 251  TTTCTTTAGT  AGCAGCGAGT  GAACAAGAAA  CAACCCTTGA  TTGTAATCCT
 301  TTATGGAGCT  GTAAATCATA  TATTTTATTT  CTTATTTCGT  AGAGGATGTA
 351  TATCCGTTAT  AAATGAGATA  TATAAAAGTA  ATTGAGTAGG  GCTGCTTGGT
 401  AGTGCAGTTT  GAATATAGGT  AGAATGAGAT  ATCTAAGGTT  AAATATAATG
 451  GTACACCGAT  AGCAAATAAG  TACTGCGAAG  GAACTTGTGA  AAATGTGTGG
 501  GTTATAGCCT  TATTTTTAAG  GACCCGTCTT  GAAACACGGA  CCAAGGAGAT
 551  TATAATTATA  GCGAGATAAC  AATGTAGTCG  TTATTAGCTT  GATAAGTTAT
 601  AATTATAAGA  CCCGAAACAC  AGTGAACTAT  ACATGTTCTG  GTTGAAGATA
 651  AGCAACAGTT  TATTGGAAGA  CCATAATCAT  TCTGACGTGC  AAATCGATGA
 701  TTTAAGATGT  GTATAGTGGC  GAAAGGCCAA  TCGAACTGTG  TGGTAGCTGG
 751  TTCACAGCGA  AATGTCTCTA  AGGACAGCAG  TCATTTTTTA  GGACATAGAT
 801  GTAGGACACT  GTTATACTAT  TTATAGTATG  AGAAATTACG  AATTCTATGG
 851  AACATTGTAA  ATTAAGTTTA  CGATGTGTAT  CTAAGAGTAT  GACTAGTGGG
 901  CACATGATTG  TAAGAATGAT  GTGCAAAAAG  GGATGAACCT  TATGAAACAT
 951  TAAATATTCT  AAATAGTAGA  CACTATACCA  TAAATATGAT  GAATACATTG
1001  AGACAGTAGG  GCGGTTGTTA  TGGAAGTAGA  AATCCGCTAA  GAAACGTGTT
1051  ACAACGTACC  TACCGAATGT  ATTATTGTAT  AAAATGGAAG  AAGATTACTA
1101  CTTTTATGAG  ATGTTTCTGT  ATAGTATTCA  GGTAGCTGTG  CAATTTGTTT
1151  GTATTGAAGT  ATGCATGTGA  GTGTGTATTG  AAGAAACAAA  TGAGCCGATC
1201  TTGGAGGCAG  TAACAATATA  TTTGTATAGA  TATTAGACTA  GGGTTTTCAA
1251  TTACTATTGA  AGTGAATCGG  AGTTATGTTA  AAAACAAAGA  AGATTAATAA
1301  TTCTTCATTT  TACTACGACA  AGGCGACTTA  ATTATGACGG  TATATTTTTT
1351  GCATAAGAAA  AATGAATTAT  GTGATGTTTA  GCTATGGATT  GTCATGATAA
1401  GAAATAGCTT  TTTATATATG  CCTGATAAAA  AAGACGTAGT  GAATCCGTAC
1451  CTATACCGCA  TCAGGTGTCA  ATGTTTACAA  ACAAATATTT  TAAAATAACG
1501  TAAGTAAGGG  AATTCGGCAA  ATTAGATCTG  TAACTTTGGG  ATAAAGATTG
1551  GCTCTAGCAT  GCTAGAACTT  TTACTATGTA  AGGAATCTGA  CTGTTTATTA
1601  AAAACATAGC  TTTTTGTATT  TACAAGAAGT  GAATTCTGCC  CAGTGCATTT
1651  ATTGTTAAAG  TTGTGTAAGC  GAATGTAAAC  GGCGGGAGTA  ACTATGACTC
1701  TCTTAAGGTA  GCCAAATGCC  TCGTCATTTA  ATTGGTGACG  CGCATGAATG
1751  GAGCAACGAG  ATTCCTACTG  TCCCTACTTA  CAATTTTGTG  AAACCACGAA
1801  ACAAGGGAAC  GGGCTTGTTA  TAAATCAGCG  AGGAAAGAAG  ACCCTGTTGA
```

```
1851  GCTTGACTTT  AGTATGTCCT  AATGAATATT  CGATATATTG  TAGCGAGGTG
1901  GGAGGAAAGT  GTGAAACCAC  TAGTATATTT  GAATATTTGT  TTACAATGGA
1951  TATATGGGGA  GTTTGGCTGG  GGCGGTACAG  CTGTTAAAAA  GTAACACAGC
2001  TGTCCTAAGG  TAGATGAATG  ATGGATGGTA  ACCATCAGTT  TATTATAAGG
2051  GAATAAATCT  GCTTTACTTT  ATATAGATGA  GTATATTTAG  TAGGGAAACC
2101  TTGGCCTTGC  GATCCCAATT  ACATTCATTA  TGTATTGGGT  GTTTGAAAAG
2151  TTACCACAGG  GATAACTGGC  TTGTAGCAGG  CAAGCGATCA  TAGCGACTCT
2201  GCTTTTTGAT  TCTTCGATGT  CGTCTCTTCT  GAACATCGTA  GTGTATATGT
2251  TACGAAGTGT  TGGATTGTTC  ACCCGGTAAT  GAGGAACGTG  AGATGGGTTT
2301  AGAC
```

Figure 5.3 a and b

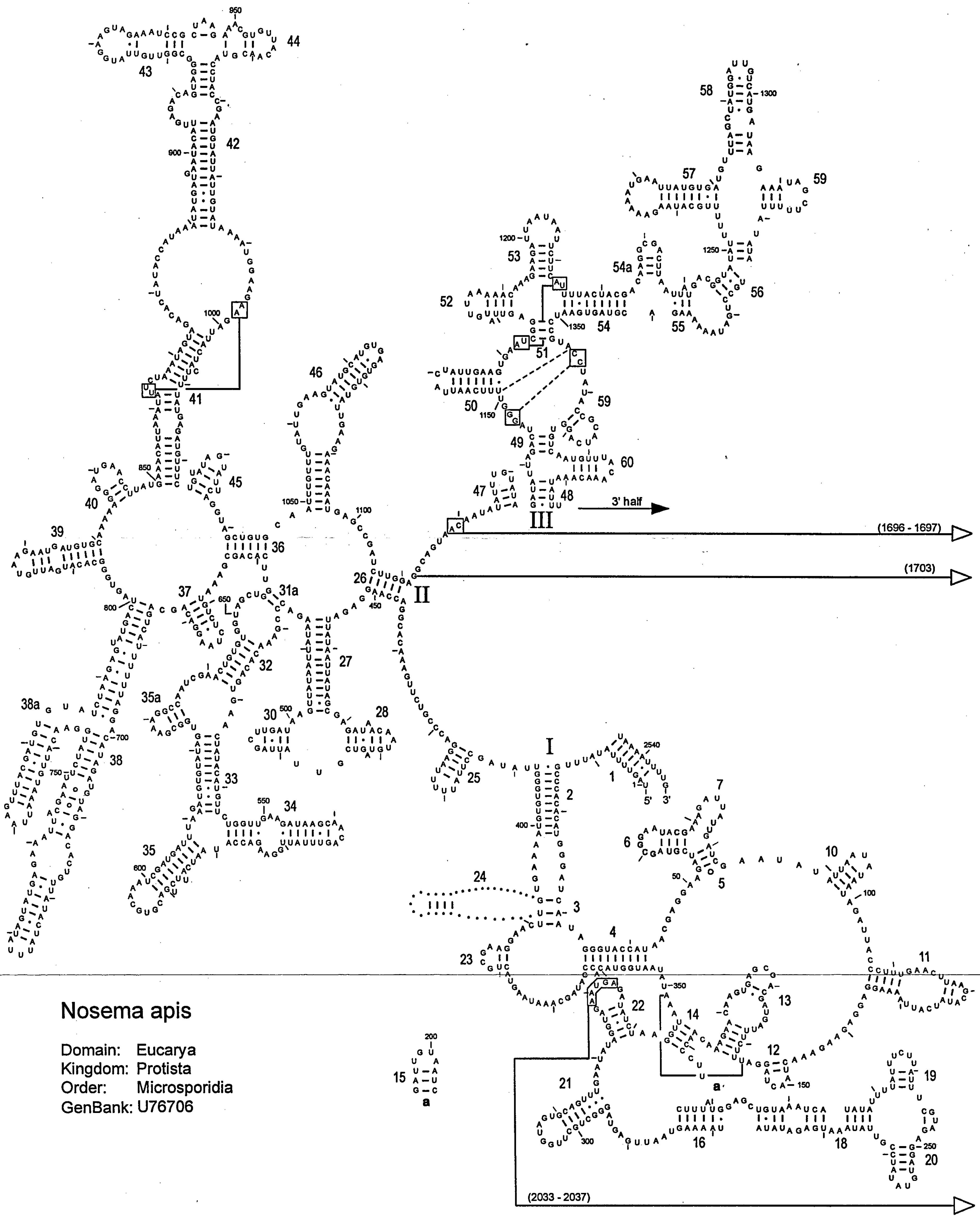
Putative secondary structure model for the *N. apis* LSU rRNA.

a. The 5' region of the putative LSU rRNA model.

b. The 3' region of the putative LSU rRNA model.

The abbreviations in both figures are as for figure 4.3 and further details are discussed in the text.

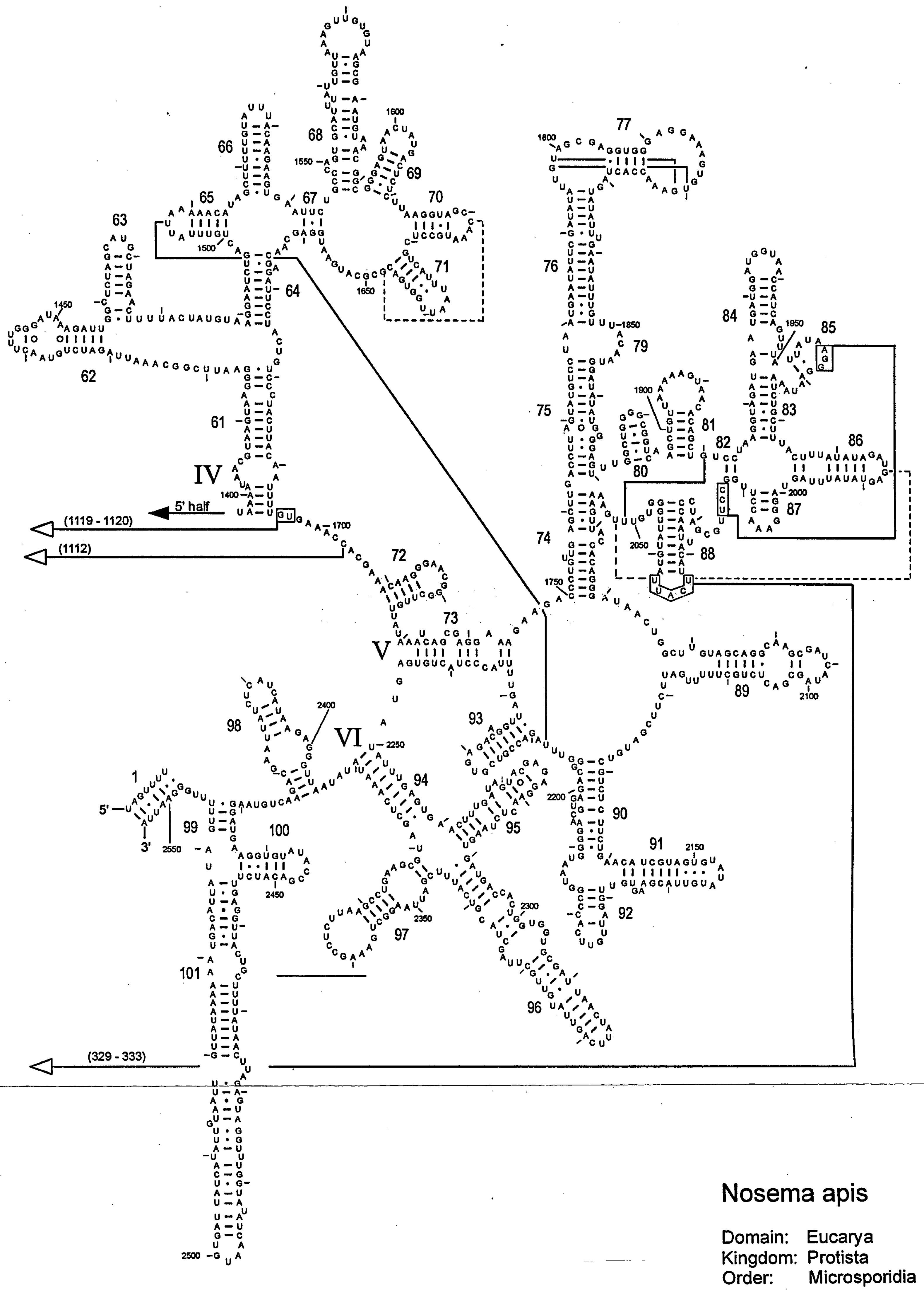
# Secondary Structure: large subunit ribosomal RNA - 5' half



## Nosema apis

Domain: Eucarya  
Kingdom: Protista  
Order: Microsporidia  
GenBank: U76706

# Secondary Structure: large subunit ribosomal RNA - 3' half



Nosema apis

Domain: Eucarya
Kingdom: Protista
Order: Microsporidia
GenBank: U76706

| Region | Nucleotide | Total | Percentage Composition | A + T% | G + C% |
|---|---|---|---|---|---|
| LSU gene | A | 841 | 32.92 | | |
| | C | 339 | 13.27 | | |
| | G | 562 | 22.02 | | |
| | T | 812 | 31.79 | | |
| Subtotal | | 2554 | | 64.69 | 35.31 |
| ITS rDNA | A | 10 | 41.67 | | |
| | C | 2 | 8.33 | | |
| | G | 3 | 12.50 | | |
| | T | 9 | 37.50 | | |
| Subtotal | | 24 | | 79.16 | 20.84 |
| 3' SSU gene | A | 23 | 33.33 | | |
| | C | 13 | 18.84 | | |
| | G | 17 | 24.64 | | |
| | T | 16 | 23.19 | 56.52 | 43.48 |
| Subtotal | | 69 | | | |
| Over all | A | 874 | 33.02 | | |
| | C | 354 | 13.37 | | |
| | G | 582 | 21.99 | | |
| | T | 837 | 31.62 | | |
| Grand Total | | 2647 | | 64.64 | 35.36 |

Table 5.1

Nucleotide composition by region for the portion of the *N. apis* rDNA operon sequenced plus the sequence from helix 92 onwards (from U97150). This data includes the 3' portion of the SSU rRNA gene, the ITS and the complete LSU rRNA gene as determined from the secondary models of the *N. apis* SSU and LSU rRNAs.

```
                                  70
                                   |
AC         :  .......... .......... .......... ..--...... ..........
JV         :  .......... .......... .......... ..--...... ..........
KI         :  .......... .......... .......... ..--...... ..........
BB         :  .......... .........T .......... ..--...... ..........
CN         :  .......... .......... .......... ..T......G ..........
WA         :  .......... .......... .......... ..--...... ..........
SW         :  .......... .......... .......... ..--...... ..........
NZ         :  .......... .......... .......... ..--...... ..........
CA         :  .......... .......... .......... .TT....... .-........
NV         :  .......... .........T ...T..AA.A .C.....G.. .-.....A..
VL         :  .......... .........T ...T..AA.A .T.....G.. .-.....A..
Consensus  :  ACCATTAGCA GGATCATAAC GAAGAATTAC AAATTTTTTA GAATTAGTTT
```

Figure 5.4a

Comparative sequence alignments of 9 *N. apis* isolates, *N. vespula* (NV), and *V. lymantriae* (VL) (Vossbrinck *et al.*, 1993). Position 70 corresponds to the putative 1$^{st}$ nucleotide position of the ITS rDNA.

Abbreviations for the isolates are as shown in figure 4.2

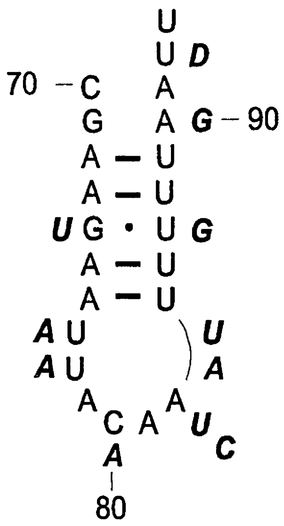


Figure 5.4b

Secondary structure model of a putative helix and loop that may act as the ITS RNA processing signal. Nucleotide substitutions (A,C,U,G), deletions (D), and insertions (U,A) are shown adjacent to the consensus nucleotides in Figure 5.4a. Nucleotide numbering also reflects that shown figure 5.4a.

| Helices | Region | Eubacteria | | Archaebacteria | | Eukaryote | | *N. apis* |
|---|---|---|---|---|---|---|---|---|
| | Between | | | | | | | |
| | Nucleotides | Mean | Stdev | Mean | Stdev | Mean | Stdev | Actual |
| Helix 7 | 68 - 79 | 24.4 | 0.5 | 23.9 | 0.3 | 22.1 | 1.4 | Absent |
| Helix 9 | 87 - 88 | 25.7 | 25.7 | 14.1 | 6.5 | 20.8 | 4.2 | Absent |
| Helix 24 | 392 - 393 | 36.2 | 0.8 | 35.9 | 0.9 | 35.8 | 0.6 | Absent |
| Helix 25 | 413 - 428 | 28.3 | 3.6 | 78.7 | 7.4 | 307.7 | 171.0 | 14 |
| Helices 29 - 31 | 485 - 502 | 34.0 | 4.1 | 63.7 | 4.4 | 94.3 | 21.2 | 16 |
| Helix 45 | 1023 - 1037 | 19.0 | 6.7 | 22.2 | 3.0 | 45.7 | 40.8 | 13 |
| Helix 47 | 1124 - 1135 | 19.0 | 0.0 | 19.0 | 0.4 | 21.0 | 6.2 | 10 |
| Helix 52 | 1177 - 1192 | 31.8 | 0.6 | 35.2 | 0.7 | 68.5 | 16.5 | 14 |
| Helix 54a | 1223 - 1238 | 19.1 | 30.9 | 16.2 | 8.0 | 35.5 | 14.7 | 14 |
| Helix 58 | 1283 - 1307 | 39.7 | 17.0 | 45.3 | 0.5 | 48.4 | 7.3 | 23 |
| Helix 59 | 1307 - 1320 | 21.5 | 8.7 | 16.2 | 6.1 | 26.2 | 1.6 | 12 |
| Helix 63 | 1456 - 1478 | 33.2 | 11.6 | 28.2 | 6.5 | 215.0 | 149.7 | 20 |
| Helix 78 | 1811 - 1812 | 35.0 | 0.3 | 35.8 | 2.2 | 15.7 | 15.9 | Absent |
| Helix 79 | 1851 - 1852 | 33.1 | 3.1 | 23.9 | 2.4 | 119.8 | 68.0 | Absent |
| Helix 98 | 2394 - 2431 | 15.1 | 6.7 | 16.6 | 16.0 | 139.3 | 37.9 | 36 |

Table 5.2

A comparison of 15 LSU variable regions for the three evolutionary domains (eubacteria, archaebacteria and eukaryotes) and *N. apis* in which the region is absent or contains at least 10 nucleotides in *N. apis*. Means and standard deviations (Stdev) were calculated from the aligned sequences of 41 eubacteria, 15 archaebacteria, and 34 eukaryotes obtained from the DCSE database home page at URL http://www-rrna.uia.ac.be/~peter/dcse/index.html.

| Feature | Position | Eukaryote | *N. apis* |
|---|---|---|---|
| Non-canonical base-pair | 626:632 | C-A | C-A |
| | 1636:1642 | usually C-A | U-G |
| Non-bonded nucleotides | 1460:1475 | C<>A | C<>A |
| | 2172:2180 | A<>C | A<>C |
| Insertion after | 585 | + | + |
| | 587 | + | + |
| | 1556 | + | + |
| | 1819 | + | + |
| | 1884 | + | + |
| | 1983 | + | - |
| Deletion after | 573 | + | + |
| | 816 | + | + |
| | 853 | + | + |
| | 1706 | + | - |
| | 2030 | + | + |

Table 5.3

Structural features of the LSU that distinguish eukaryotes from eubacteria (Gutell *et al.,* 1993; Schnare *et al.,* 1996) and their occurrence in the *N. apis* LSU. In each case, the eubacteria have a state opposite to that indicated for the eukaryotes. Positions indicated are relative to those of the *N. apis* LSU rRNA secondary structure model, figure 5.3a and b.

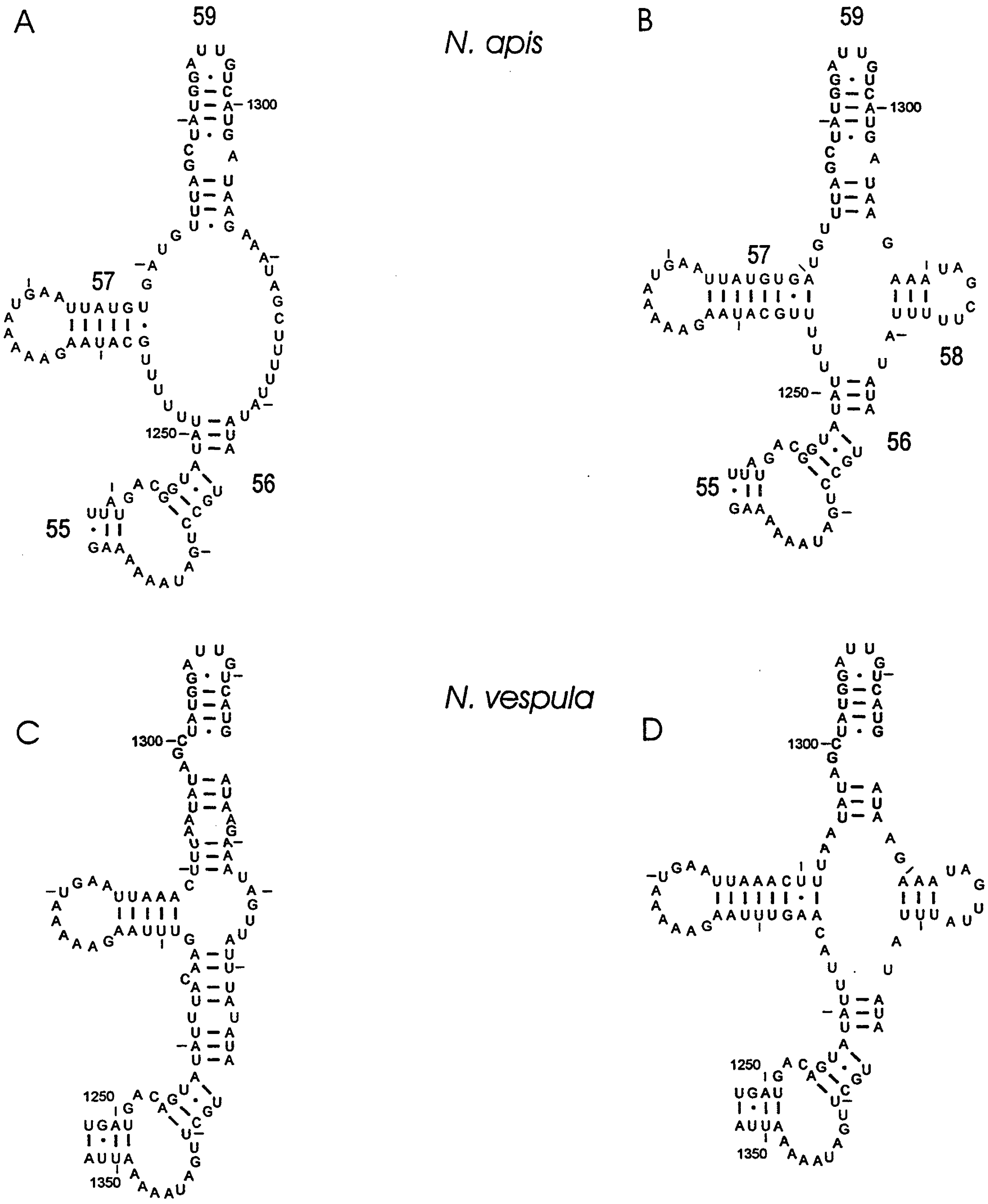


Figure 5.5

Alternative secondary structural arrangements for the region containing helices 56 to 59 inclusive of the *N. apis* and *N. vespula* LSU rRNAs.

# CHAPTER 6

# *Nosema vespula* rDNA REPEAT UNIT: SEQUENCE, GENE ARRANGEMENT AND SUBUNIT SECONDARY STRUCTURES

## 6.1 Introduction

Gerbi (1985) eloquently defined the nature of the translation apparatus and its central role in evolution:

> "Darwinian evolution requires selection operating on the phenotype to favor which changes in the genotype are to be maintained and perpetuated. Since proteins are generally responsible for the phenotype, most Darwinian evolution as we think of it today could not occur until the establishment of protein synthesis. It is primarily the translation apparatus that allows expression of the genotype at the level of the phenotype."

The evolution of the translation apparatus is far from understood. However, evidence now exists supporting a common ancestor for the translation apparatus of extant cellular life forms (Gerbi, 1985; Gray and Schnare, 1990; Noller, 1991). In particular, support is derived from the analysis of the primary sequence, secondary structure, tertiary interactions, and rRNA/ribosomal-protein interactions of the rRNA subunits. This analysis includes taxa from the three primary lineages: eubacteria, archaebacteria and eukaryotes (Gerbi, 1985; Appels and Honeycutt, 1987; Raué *et al.,* 1990; Hill *et al.,* 1990, and Noller, 1991, Baranov *et al.,* 1998).

Collectively, such studies have revealed that the rRNA subunits have been subjected to considerable functional and evolutionary constraint. For example, regions approximately equivalent to one-third of the total *E. coli* rRNA secondary structures are globally conserved in the three primary lineages (Gray *et al.,* 1984; Cedergren *et al.,* 1988).

Even before any primary sequences of rRNAs were available, Woese and Fox (1977) recognised their potential as a means of inferring a phylogenetic framework that could include all cellular life forms. Due to the high sequence and structural similarity of rRNA sequence amongst all organisms, the rRNAs remain the most important genes for resolving deeply branched relationships including the divergence of the early eukaryotes. Subsequently, and from the efforts of many, there now exists a substantial amount of taxonomic information describing inferred phylogenetic relationships based on rRNA sequence alignments. These analyses suggest that the earliest diverging eukaryotes are the Microsporidia and the Diplomonadida, although the relative branching order has been the subject of much debate (Chapter 1). In an attempt to resolve this debate, phylogenetic inferences have been suggested based on protein encoding genes. The question of respective divergence order however still remains unresolved.

Perhaps by extending the phylogenetic analysis of rRNA genes to include their arrangement, secondary structure, tertiary and macro-molecular interactions, additional evidence may be revealed that answers the question of divergence order. That is, it may be possible to infer the divergence order of the earliest eukaryotes based on the presence or absence of plesiomorphic states (structures or interactions) in the translation apparatus of the Microsporidia and Diplomonadida. For example, this type of analysis has revealed the presence of eukaryotic-specific elements within the secondary structure of the LSU (Michot and Bachellerie, 1987).

Until recently, only the primary sequence data for the SSU rRNA gene of several microsporidian species was available. In this chapter I present: the entire sequence of a microsporidial rDNA repeat unit; the arrangement of the rRNA genes within the operon; putative secondary structure models for the rRNA subunits including tertiary interactions; and phylogenetic inferences based on gene arrangement and secondary structure elements.

## 6.2 Materials and Methods

Sequence data for the SSU, LSU and 5S rRNA genes of *N. vespula* were obtained by molecular techniques as described in sections 2.3 to 2.7. The SSU and LSU rRNA gene sequences were aligned to existing pre-aligned sequences as described in section 2.8. Secondary structure models for the SSU, LSU and 5S rRNA were developed and drawn as described in section 2.8.

## 6.3 Results and discussion

### 6.3.1 rDNA repeat unit sequence

As depicted in Figures 6.1a to d, four DNA fragments were amplified and sequenced. From these four overlapping clones of 802, 681, 1790, and 4255 bps a total of 7292 bps were determined (Figure 6.2). The approximate lengths of each rRNA gene were determined from the putative secondary structure models for the SSU (Figure 6.3), LSU (Figures 6.4a & b), and 5S (Figure 6.5) rRNAs. From these models, the respective sizes of the SSU, LSU, and 5S rRNAs genes are 1245, 2549, and 120 bps.

The base composition across the rDNA repeat unit is region dependent (Table 6.1). The rDNA repeat unit contains 29.0% G + C. The base composition for individual regions ranges from 21.94% for the non-transcribed spacer to 45.53% for the 5S rRNA gene. The G + C content of the SSU and LSU rRNA genes respectively is 36.76% and 33.74%. This is comparable to the G + C content determined for the SSU rRNA gene of other *Nosema* and *Vairimorpha* species. These vary from 33.9 to 37.3% G + C (Gatehouse and Malone, 1998).

### 6.3.2 Gene arrangement

Most cellular organisms have multiple copies of the rRNA genes that code for the rRNA subunits. Typically, the prokaryotic arrangement of the rRNA genes within the operon (Figure 6.6a) is 5'-SSU rRNA gene ⇒ internal transcribed spacer (ITS) ⇒ LSU rRNA gene ⇒ ITS ⇒ 5S rRNA gene- 3'. The genes are transcribed as a single unit by RNA polymerase I and post-transcriptionally cleaved into the rRNA subunits. When multiple copies are present in the genome, the cistrons exist scattered across the genome. In contrast, the eukaryotic arrangement of rRNA genes within the operon (Figure 6.6a) is 5'- SSU rRNA gene ⇒ ITS1 ⇒ 5.8S rRNA gene ⇒ ITS2 ⇒ LSU rRNA gene -3'. The eukaryotic 5S rRNA gene usually occurs at other loci within the genome. The SSU and LSU rRNA genes are transcribed by RNA polymerase I whereas the 5S rRNA gene is transcribed by a separated RNA polymerase, RNA polymerase III. The eukaryotic rDNA operons usually occur as tandem arrays separated by a non-transcribed spacer (Gerbi 1985, Appels and Honeycutt, 1987).

Derivations from the typical pro- and eukaryotic gene arrangements do occur. The operon of the archaebacteria may be interrupted such that they are separate but still transcribed individually by RNA polymerase I (Figure 6.6a) (Appels and Honeycutt, 1987). A covalently linked 5.8S-LSU equivalent to the prokaryotic 23S gene occurs in the Microsporidia (Figure 6.6b) (Vossbrinck and Woese, 1986). The 5S rRNA gene in some eukaryotes occurs in the same orientation and repeating unit as the other rRNA genes. However, transcription in the opposite direction has been noted (Kawai *et al.,* 1995; Kawai *et al.,* 1997). These are the so called "linked 5S" rRNA genes that are found in the Euglenophyceae, Phaeophyceae, some fungi, protozoa and metazoa (Kawai *et al.,* 1995).

The largest PCR fragment cloned (Figure 6.1d) spans the non-transcribed spacer separating two contiguous rDNA operons. This PCR fragment demonstrates that the rDNA repeats in *N. vespula* occur in a head-to-tail tandem fashion equivalent to that typically found in eukaryotes. This arrangement of rDNA repeating units has also been demonstrated for the Microsporidian *N. apis* (Gatehouse and Malone 1998). However, it has been shown that the rDNA operons of some Microsporidia do not occur as tandem repeats but rather are scattered across all or some of the chromosomes (Kawakami *et al.*, 1994; Biderre *et al.*, 1994; Peyretaillade *et al.,* 1998).

Estimates of PCR product length, and confirmation of template by partial sequencing, indicate that the rDNA repeating unit of *N. apis* is approximately 18 kilobases. This is more than twice the length of the *N. vespula* repeat unit of 7292 bps. A comparison between the primary sequence of *N. vespula* and *N. apis* for the region encompassed by the SSU and LSU rRNA genes demonstrates that most of this length variation is localised to the non-transcribed spacer. By comparison, the rDNA repeat unit length of *N. vespula* is similar to that reported for the *Giardia* species which range from 5.6 and 7.6 kilobases (Van Keulen *et al.,* 1991).

The arrangement of genes in the rDNA operon of *N. vespula* (Figure 6.6b) was found to be identical to that observed for the prokaryotes—a covalently linked 5.8S-like and LSU rRNA genes and the 5S rRNA gene located downstream from the 3' termini of the LSU rRNA gene. The covalently linked 5.8S-like and LSU rRNA genes was expected. However, the presence of the 5S rRNA gene within the repeat unit was not. The 5S rRNA gene is located 279 base-pairs downstream of the LSU rRNA gene. Evidence suggests that this is a functional 5S rRNA gene. The length of the gene sequence is approximately 120 nucleotides which is equivalent in length to that of the typical 5S rRNA gene (Wolters and Erdmann, 1986; Barciszewska *et al.,* 1996). A search of the GenBank database demonstrates homology of this sequence to that of *N. bombycis* (Acc No. D14631) and to a range of other eukaryotic 5S RNA gene sequences including those of protozoans and advanced eukaryotes. The RNA polymerase III termination signal (TTTTT) is also present in the 3' flanking region of the 5S rRNA gene. Comparative analysis to the standard models of the 5S rRNA gene (Wolters and Erdmann, 1986; Barciszewska *et al.,* 1996) demonstrates homology at the sequence level for conserved nucleotides and for a putative secondary structure. Interestingly, the 5S rRNA gene of the microsporidian *E. cuniculi* has found not to be linked to the rDNA operon rather occurring else where within the genome (Peyretaillade *et al.,* 1998).

Gerbi (1985) hypothesised that the typical prokaryotic rRNA gene arrangement is the plesiomorphic state. However, putative phylogenetic relationships between the prokaryotes and eukaryotes based on rRNA gene arrangement remains controversial because of the phylogenetic distribution of eukaryotic taxa (lower and upper) that possess a 'linked' 5S rRNA gene (Kawai *et al.,* 1997). Additionally, other genes have been shown to be linked to the 5S rRNA genes in some organisms.

A histone gene has been reported as linked to the 5S rRNA gene in the crustacean *Artemina* (Cruces *et al.,* 1989). More recently, the trans-spliced leader has been shown to be linked with the 5S rRNA gene in some protozoa (trypanosomatids) (Aksoy *et al.,* 1992; *Euglena*, Keller *et al.,* 1992; *Toxoplasma* Guay *et al.,* 1992). Maslov *et al.* (1993) consider that the linkage of the 5S rRNA gene and the trans-spliced leader in some protozoa (*Trypanosoma, Herpetomonas, Bodo, and Euglena*) to be an evolutionary primitive state within the lineage. In contrast, it has been suggested that the linkage of the 5S rRNA gene to the rDNA and other genes is a secondary evolutionary state in eukaryotes (Drouin and Moniz de Sá, 1995). Drouin and Moniz de Sá site as evidence for this conclusion: the occurrence of 5S-rDNA linkage with transcription of the 5S rRNA gene in both directions; the occurrence of this 5S-rDNA linkage in primitive and advance eukaryote; and the linkage of the 5S rRNA gene to other genes. Kawai *et al.* (1997), have studied 5S-rDNA linkages in the Chromophyta, Dinophyceae, and Euglenophyceae. They consider the 5S-rDNA linkage in eukaryotes has occurred multiple times within the various phylogenetic groups. They also suggest that the linkage of the 5S rRNA gene to other genes may explain why the 5S rRNA gene of the eukaryotes is transcribed by RNA polymerase III. This would facilitate the relative ease of translocation for the 5S rRNA gene to other loci.

The arrangement of the rRNA genes within the rDNA operon suggests a plesiomorphic link to the prokaryotes. However, the possession of the RNA polymerase III termination signal at the termini of the 5S rRNA gene links this species to the eukaryotes. In a phylogenetic context, these results demonstrate an intermediate state in ribosomal gene arrangement between prokaryotes and eukaryotes. This suggests that the microsporidians diverged after the prokaryotes and before other extant eukaryotes including the Diplomonadida.

### 6.3.3 Sub-repeats of the intergenic spacer

Sub-repeat sequences occur in the non-transcribed spacers of the rDNA of most eukaryotes (Morton *et al.,* 1995). Specific examples include: *X. laevis* (De Winter and Moss, 1987), *Drosophila melanogaster* Meigen (Grimaldi and Dinocera, 1988), *A. thaliana* (Doelling *et al.,* 1993), *Oryza sativa* Linnaeus (Cordesse *et al.,* 1993), and *Leishmania amazonensis* Lainson and Shaw (Uliana *et al.,* 1997). It has been shown that these elements alter the transcription rate of the rRNA genes (Reeder, 1990).

A contiguous block of five repeats was identified in the intergenic spacer of *N. vespula.* The repeats start 660 nucleotides downstream from the 5S rRNA gene and involve a region of 583 bps (Figure 6.2, nucleotide 4881 onwards). The repeats vary in length from 113 to 120 bps (Figure 6.7). Most of the length variation is accounted for by a 5 bps insert in repeats two and four. The similarity of the sequences to each other varies from 70% to 83% (Table 6.2). There are two regions within the repeats that are highly conserved. The first has the sequence TTAGTTAAGTGATGTTTATAATATTA, and the second, the sequence TAGGTTAGTTGTTTAAATATGATTAATAAAA.

### 6.3.4 Secondary Structure Models

#### *a. Small subunit ribosomal RNA*

A putative rRNA secondary structure model has been determined from the SSU rRNA of *N. vespula* (Figure 6.3). This model is similar to that proposed initially by Neefs *et al.,* (1991) and subsequently modified by Gutell (1994a) for the SSU rRNA of *V. necatrix.* A comparison with the standard eukaryotic SSU rRNA model (Appendix 1) reveals the absence of helices 10, 11, 44 as previously noted for *V. necatrix* (Neefs *et al.,* 1991). In contrast, the microsporidian *E. cuniculi* possesses helix 10 while helices 17 and 43 are absent (Hartskeerl *et al.,* 1993). These structural similarities between *V. necatrix* and *N. vespula* concur with the phylogenetically inferred trees that place *N. vespula* and *V. necatrix* as sister taxa relative to the *Encephalitozoon* genera (Baker *et al.,* 1995; Malone and McIvor, 1996). These types of structural similarities/differences may prove useful for the classification of microsporidian species into their appropriate genera. Such additional phylogenetically informative characters would clarify some of the apparent confusion that currently exists based on ultrastructure and the reproductive cycle (Chapter 1).

In comparison to the standard eukaryotic SSU rRNA model, the Microsporidia (as represented by *N. vespula*, and *V. necatrix, Nosema necatrix* Kramer (Fries *et al.*, 1996), and *E. cuniculi* (Hartskeerl *et al.*, 1993)) possess secondary structure elements that are prokaryotic-like. Three of these structural elements are also shared by the Diplomonadida (as represented by *G. muris, G. ardeae,* and *G. intestinalis.*

*Positions 111 - 145*

The distal half of helix 9 in *N. vespula* (Figure 6.8a) is formed by the pairing of 14 nucleotides closed by a hairpin. The eukaryotic structure includes a non-bonded nucleotide pair with a hairpin containing four or five nucleotides. In the eubacteria and archaebacteria, all the nucleotide pairs form bonds. The helix is closed by a tetraloop of highly conserved nucleotides (>95% conserved) of the form GAAA. The structure of this region is prokaryotic-like in the Microsporidia and Diplomonadida. All 14 paired nucleotides form bonds and their respective helices are closed by a tetraloop. The character-state of the tetraloop in the Diplomonadida is CAAC and in Microsporidia UAAA except for *E. cuniculi* that is GAAA.

*Positions 329 - 371*

Structural models of helix 18 have been proposed for *N. vespula* (Figure 6.3), *E. cuniculi* and diplomonads (Gutell, 1994a). The structure of helix 18 in the diplomonads parallels that of other eukaryotes particularly regarding the size of the closing hairpin and its two closing nucleotide pairs. The sequence for helix 18 is identical in *V. necatrix,* and *N. necatrix.* These sequences differ from that of *N. vespula* at two positions within proposed non-bonded regions.

In the proposed model (Figure 6.8b), helix 18 can be drawn as a long helix similar to that found at the base of helix 18 in the prokaryotes. The length of this helix differs between *E. cuniculi* and the other Microsporidia by 5 nucleotide pairs—*E. cuniculi* being the shorter helix. Despite the difference in helix length, this region in the Microsporidia appears to be an intermediate structure between the prokaryotic and eukaryotic states. It appears somewhat more similar to the structure of helix 18 in the archaebacteria which appears to have lost the hypervariable region beyond the first opposing loop structure shown in figure 6.8b for the eubacteria.

*Positions 372 - 374*

The junction loop that joins helix 18 and 19 (Figure 6.8b) contains two conserved nucleotides in prokaryotes. The second of these nucleotides is also conserved in eukaryotes. The character-state of these nucleotides is AA in the prokaryotes and A in the eukaryotes. The preceding nucleotide in the archaebacteria is also conserved (approximately 85%) and is usually a U. *E. cuniculi* has the arrangement UGA, while the other Microsporidia possesses the archaebacterial arrangement of UAA. In contrast, the diplomonads possess the conserved eukaryotic nucleotide but the preceding nucleotides are neither eubacterial or archaebacterial in character state. The character state of these 3 nucleotides is either GGA or CGA.

*Positions 587 - 628*

The second nucleotide pair at the base of helix 25 (Figure 6.8c) is conserved within the eubacteria, archaebacteria and eukaryotes, although some differences are present between the prokaryotes and eukaryotes. In eukaryotes the character state of this nucleotide pair is U:A whereas in the prokaryotes, Microsporidia and *Giardia* species this nucleotide pair is of the form C:G.

*Positions 763 - 773 and 995 - 1007*

The fifth nucleotide pair of helix 32 (Figure 6.8d) in eukaryotes is usually non-bonded and of the form C<>C or C<>U. In the prokaryotes this nucleotide pair forms a bond with the conserved character state of A:U in eubacteria and C:G in the archaebacteria. The Microsporidia all possess a bonded nucleotide pair but of the form G:C. The diplomonads have the eukaryotic arrangement, with *G. muris* and *G. ardeae* sharing the C<>U character state and *G. intestinalis* the C<>C state.

*Positions 809 - 848*

Structurally, helix 35 (Figure 6.8e) is very similar in the eukaryotes and the archaebacteria but differs from the eubacteria in structure and in nucleotide conservation. Furthermore, within these three lineages, the structures are well conserved. The structure of this helix in Microsporidia is neither that of the eubacteria nor eukaryote/archaebacteria. In fact, the microsporidial structure can be drawn in a manner similar to either alternative. The divergence of the archaebacteria prior to the microsporidia suggest that the eubacterial alternative is either incorrect or a reversion to the plesiomorphic state. Interestingly, the prokaryotic-like structure is the only possible consensus structure for all the Microsporidia represented (data not shown).

*Positions 926 - 950*

The prokaryotes and eukaryotes differ structurally in helix 42 (Figure 6.8f). In eukaryotes the usual arrangement is a helix of seven or eight nucleotide pairs closed by a hairpin of five nucleotides. This helix is linked to helix 41 by two nucleotides and to helix 40 by three or more nucleotides. Only the two nucleotides joining helices 41 and 42 are conserved in the eukaryotic model. In contrast, the initial 3 nucleotide pairs observed in helix 42 of the eukaryotic model are present in the prokaryotic model but as non-bonded pairs. Furthermore, the character state of the first 2 nucleotide pairs contained in helix 42, its closing hairpin, and the junction loop between helix 41 and 40 are conserved (≥ 90%) in the prokaryotes. Of particular note are the nucleotides at positions 930 and 944 (*N. vespula* numbering). These are a highly conserved (≥ 95%) G<>A pair characteristic of prokaryotes.

Except for *E. cuniculi,* the microsporidians, and the diplomonads share structural homology with the prokaryotes in this region. *E. cuniculi* differs from the other microsporidians in that helix 41 is absent. Also, helix 42 of *E. cuniculi* contains an extra bonded nucleotide pair and the closing hairpin is a tetraloop. Despite this variation within the microsporidians, most of the conserved nucleotides seen in the prokaryotes are also present. Conservation of prokaryotic nucleotides was also observed in the diplomonads.

The nucleotide character-state differences between *N. vespula* and the eubacterial model involve a bonded nucleotide substitution (C:G to A:U) at the base of the helix (positions 931:943, *N. vespula* numbering) and a single nucleotide substitution at position 929 (U to A). The first substitution event does not alter the helical structure. The second substitution supports the non-bonded arrangement of nucleotides at this position in the eubacteria—the substitution results in the formation of a non-bonded A<>G pair (positions 929-945). In contrast, the second substitution event as seen in *Giardia* has replaced the U with a C that could potentially bond with the opposing G resulting in a C:G base pair. Additionally, a non-bonded G at position 948 has been substituted by an A that does not alter the structure of this region in respect to the prokaryotic models. The prokaryotic characteristic G<>A non-bonded pairing at positions 930 and 944 (NV, 930-944) are present in both the microsporidians and the diplomonads.

### *b. Large subunit ribosomal RNA*

A putative secondary structure model has been determined from the LSU rRNA of *N. vespula* (Figure 6.4a, b). This model is very similar to that proposed for *N. apis* in chapter five. Consequently, the discussion will be limited initially to structural differences between the putative *N. vespula* and *N. apis* models and then to a discussion on prokaryotic-like structural elements.

### *N. vespula* and *N. apis* LSU structural variation

Structurally, there are two major and six minor differences between the LSU rRNAs of *N. vespula* and *N. apis* (Figure 5.3a, b and 6.4a, b). The first of the major differences can be seen in the hypervariable region encompassed by helices 14 and 19. This region can usually be modelled around several helical elements (Gutell, 1993). *N. vespula* and *N. apis* contain helices 15, 16, and 18, however only helix 15 is strictly conserved. Many of the nucleotides of helices 16 and 18 are conserved between these two species, yet the structure of the helices could only be described as similar. It would appear that 'slippage' of structural elements has occurred. This is apparent when nucleotide conservation occurs in only one strand and at adjacent locations in the organisms compared. In comparing the two putative models, it appears that the nucleotides of the 5' region of helix 16 in *N. apis* form the junction loop between helices 15 and 16 of *N. vespula.* This event is even more apparent when the nucleotides are aligned as below:

| | |
|---|---|
| *N. vespula* junction loop | `5'-CUUAACUGGAGA-3'` |
| *N. apis* 5' strand of helix 16 | `5'-CUUUA-UGGAGC-3'` |

The second major structural difference occurs in helix 101. By constraining the helical structures in both microsporidians such that it is similar to that found in other eukaryotes, helix 101 of *N. apis* is longer. Helix 101 of *N. apis* contains 86 nucleotides which can be arranged to form 31 bonded pairs and several internal loops and bulged nucleotides. In comparison, this region in *N. vespula* contains 67 nucleotides resulting in potentially 20 bonded nucleotide pairs, several loops and bulged nucleotides. It is likely the shorter helix of *N. vespula* is the consequence of a deletion event(s). This is considered to be the most likely cause for two reasons: the first 20 nucleotides of this helix are conserved in both organism; and deletions events, as seen by the absence of conserved structure, are common in microsporidians.

There are also several minor structural differences, of these the most interesting occurs in helix seven. As discussed in chapter five, helix 7 of the microsporidians has been truncated with respect to the standard eukaryotic model. In *N. vespula,* this helix potentially exists by the base-pairing of three A:U nucleotide pairs. In *N. apis* however, the first A:U nucleotide-pair of the helix has been replaced by an A<>G pair. It is possible that these nucleotides may also form a bond. The A:G arrangement of bonded nucleotides occurs several times in both models of the eukaryotic SSU and LSU (Gutell 1994a; Gutell *et al.*, 1993) and is often the first or last nucleotide pair of helical elements—for example helices 5, 38, 62, and 75 of the *N. vespula* and *N. apis* LSU models.

The remaining structural differences between the LSUs of *N. vespula* and *N. apis* respectively are: the extension of helix 28 by a nucleotide pair and an additional nucleotide in the closing hairpin; the deletion of a bulged nucleotide from the 3' region of helix 45; the deletion of two nucleotides in the junction loop between helices 48 and 49; an additional nucleotide pair in helix 52; an additional nucleotide in the closing hairpin of helix 52; and an additional nucleotide pair in helix 54a.

Prokaryotic-like features of the *N. vespula* LSU rRNA

The Microsporidia (as represented by *N. vespula* and *N. apis*) possess secondary structure elements that are prokaryotic-like. There are eight secondary structure elements in the putative *N. vespula* and *N. apis* LSU rRNA models that are prokaryotic in form. One of these structural elements is also shared by the Diplomonadida (as represented by *G. muris, G. ardeae, G. intestinalis).*

*Positions 1 - 7 and 2493 - 2499*

A feature of prokaryotes, helix 1 (Figure 6.9a) results from base pairing between nucleotides of the 5' and 3' termini of the LSU rRNA. In LSU models for both microsporidians (Figures 5.3a and 6.4a) , the formation of helix 1 is possible and supported by the presence of conserved nucleotide pairs and paired nucleotide substitutions. However, additional support based on comparative analysis of other microsporidian sequence is required to confirm the presence of this helix.

*Positions 51 - 54 and 80 - 83*

At the base of helix 5 is a universally conserved A:G bonded, nucleotide pair (Figure 6.9b). The remainder of the helix consists of 3 more bonded, nucleotide pairs. The only other conserved nucleotide in this helix of the eukaryotes is a G immediately 3' to the conserved A. In the prokaryotes however, all but the fourth nucleotide pair are conserved; the second and third pair both being a G:C. Both the microsporidians and the diplomonads possess the initial G:C pair. The second G:C pair is only found in *N. vespula* and *G. ardeae.* This nucleotide pair in *G. muris* is G:U while in *N. apis* and *G. intestinalis* is A:U. Those results suggest that the prokaryotic state observed in *N. vespula* and G. *muris* may be homoplasies.

*Positions 89 - 102*

Helix 10 (Figure 6.9c), a site of discontinuity in eukaryotes, is formed by base pairing between the 3' termini of the 5.8S subunit and 5' termini of the LSU. The internal transcribed spacer that separates a 5.8S rRNA gene and LSU rRNA gene of the eukaryotes is absent in prokaryotes. That is, the 5.8S subunit of eukaryotes and the 3' region up to and including helix 10 of the prokaryotic LSU are structurally homologous. This prokaryotic feature of eukaryotes has thus far been shown to be unique to the microsporidians.

*Positions 533 - 538 and 650 - 655*

The first nucleotide pair of helix 32 is highly conserved in the prokaryotes and eukaryotes but differs in character state in the three primary lineages (Figure 6.9d). In the eukaryotes this nucleotide pair is of the form G:C. However, in the prokaryotes and the Microsporidia is of the form C:G.

*Positions 678 - 691*

Helix 37 is capped by a tetraloop. This structure is universally conserved in both the prokaryotes and eukaryotes except for the second nucleotide pair (Figure 6.9e). In the eukaryotes this nucleotide pair is highly conserved, of the form U<>U, and remains non-bonded. In contrast, while of variable composition in the prokaryotes, this nucleotide pair form a bond. In the microsporidians this base pair is of the form G:C but variable in the diplomonads: *G. muris* G:U, *G. ardeae* U<>U, and *G. intestinalis* G:C. The other notable exception in the eukaryotes occurs in *P. falciparum* and is of the form A:U (Gutell *et al.*, 1994b).

*Positions 694 - 810*

Structurally, helix 38 (Figure 6.9f) is similar in the prokaryotes and eukaryotes with the exception that one of the two hypervariable regions contained therein is a eukaryotic 'signature' element, helix 38a. Helix 38a is also present in the Microsporidia and the Diplomonadida. Nucleotide conservation in helix 38 occurs in the region below the opposing hypervariable regions and the closing hairpin-loop of the helix. Within this region (positions 724 to 728 and 749 to 752, *N. vespula* numbering ) are two conserved, opposing internal loops flanked by base-pairs. In the eubacteria, this region consists of the highly conserved nucleotides 5'-C A C U G-3' opposed by 3'-N A A A-5' (where 'N' means positional conservation). In the eukaryotes this region consists of 5'-N ^ g a u G-3' opposed by 3'-C a A A-5' (where lower case is 90 - 95% conservation and '^' has variable positional conservation). Of the diplomonads represented in this discussion the nucleotide sequence of *G. intestinalis* (5'-

C G C U G-3' and 3'-C A A A-5') is most similar to that of the eubacteria. The microsporidian sequences are 5'-C A C U G-3' and 3'-G A A A-5'. In the archaebacteria the structure is similar to that of the eubacterial but does not have the same level of nucleotide conservation.

*Positions 1707 - 1719*

The junction loop between helix 61 and 72 (Figure 6.9g ) contains 12 nucleotides in the eukaryotes and 13 nucleotides in prokaryotes and *N. vespula*. The single G at position 1717 (*N. vespula* numbering) in the microsporidia displaces a conserved A seen in the eukaryotic model. The corresponding nucleotide in prokaryotes is present but not conserved.

*Positions 2302 - 2356*

The character state of positions 2307 to 2314 and 2342 to 2349 (Figure 6.9h, *N. vespula* numbering) of helix 96 in the eubacteria are 5'-c C N c U N G U-3' and 5'-N N U a G C N A-3' (where underlined regions base pair to form a helix). In eukaryotes, this region is of variable nucleotide composition and of the form 5'-N N N u g G N-3' and 5'-N N N N N N a-3', with the second region containing one less nucleotide than in the eubacteria. In the Microsporidia these regions closely resemble that of the eubacteria and are of the form 5'-CCACUGGU-3' and 5'-CUUAGCUA-3'. The equivalent regions in the archaebacteria are very similar to that of the eubacteria being 5'-CCNCNNGU-3' and 5'-NNCAGCNA-3'.

*Tertiary Interactions*

Analysis of microsporidian rRNA tertiary interactions may reveal that some of these are prokaryotic-like. An example of such a potential interaction involves coaxial stacking. It has been suggested that the first half of helix 95 and 96 potentially form a coaxial stack (Figure 6.9h) (Gutell *et al.,* 1994b). The combine length of these two helical regions is 13 nucleotide pairs. In the eubacteria these two helices are 7 and 6 nucleotide pairs in length, whereas in the archaebacteria and the eukaryotes the corresponding lengths are 8 and 5 nucleotide pairs. The microsporidians share the eubacterial arrangement.

*c. 5S sub-unit ribosomal RNA*

A putative secondary structure model has been determined from the 5S rRNA of *N. vespula* (Figure 6.5). This model is similar to that of the standard eukaryotic 5S rRNA model (Wolters and Erdmann, 1986; Barciszewska *et al.,* 1996). However, several differences in nucleotide conservation patterns and secondary structure are apparent.

There are 74 conserved nucleotide positions in the standard eukaryotic model for the 5S rRNA (Wolters and Erdmann, 1986). Of these positions, nine have alternative character states and one is absent in the proposed *N. vespula* model (Figure 6.5). More significantly, 6 of these 9 positions are universally conserved across the three primary lineages (positions 22, 34, 36, 41, 43, and 100).

The other feature that differentiates the 5S sunubit of *N. vespula* from that of other eukaryotes is the possible extension of helix D by two nucleotide pairs and the deletion of a conserved U in the d' loop. Helix D of the eukaryotic model contains five bonded nucleotide pairs. This helix starts with the a conserved nucleotide pair at positions 68:109 (C:G) and ends with the conserved nucleotide pair at positions 72:105 (G:C). The character states of the nucleotides at positions 73:104 (U:G) and 74:103 (U:A) favours the formation of bonds that would extend the helix. The extension of helix D is also a feature of the archaebacterial model determined for halophilic and methanogenic archaebacteria (Wolters and Erdmann, 1986).

## 6.4 . Summary

The size of microsporidian rRNA subunits and the absence of the 5.8S rRNA gene were the first physical indicators the these organisms may be very ancient eukaryotes. It has now been demonstrated that other physical features are present, adding additional support for the early divergence of the Microsporidia. These features include: the prokaryotic arrangement of the rRNA genes within the rDNA operon and the presences of a linked 5S rRNA gene; the presence of highly conserved nucleotide positions that have maintained a prokaryotic character state; and a number of secondary structural elements in the rRNA of the Microsporidia that are prokaryotic-like.

Figure 6.1 a-d

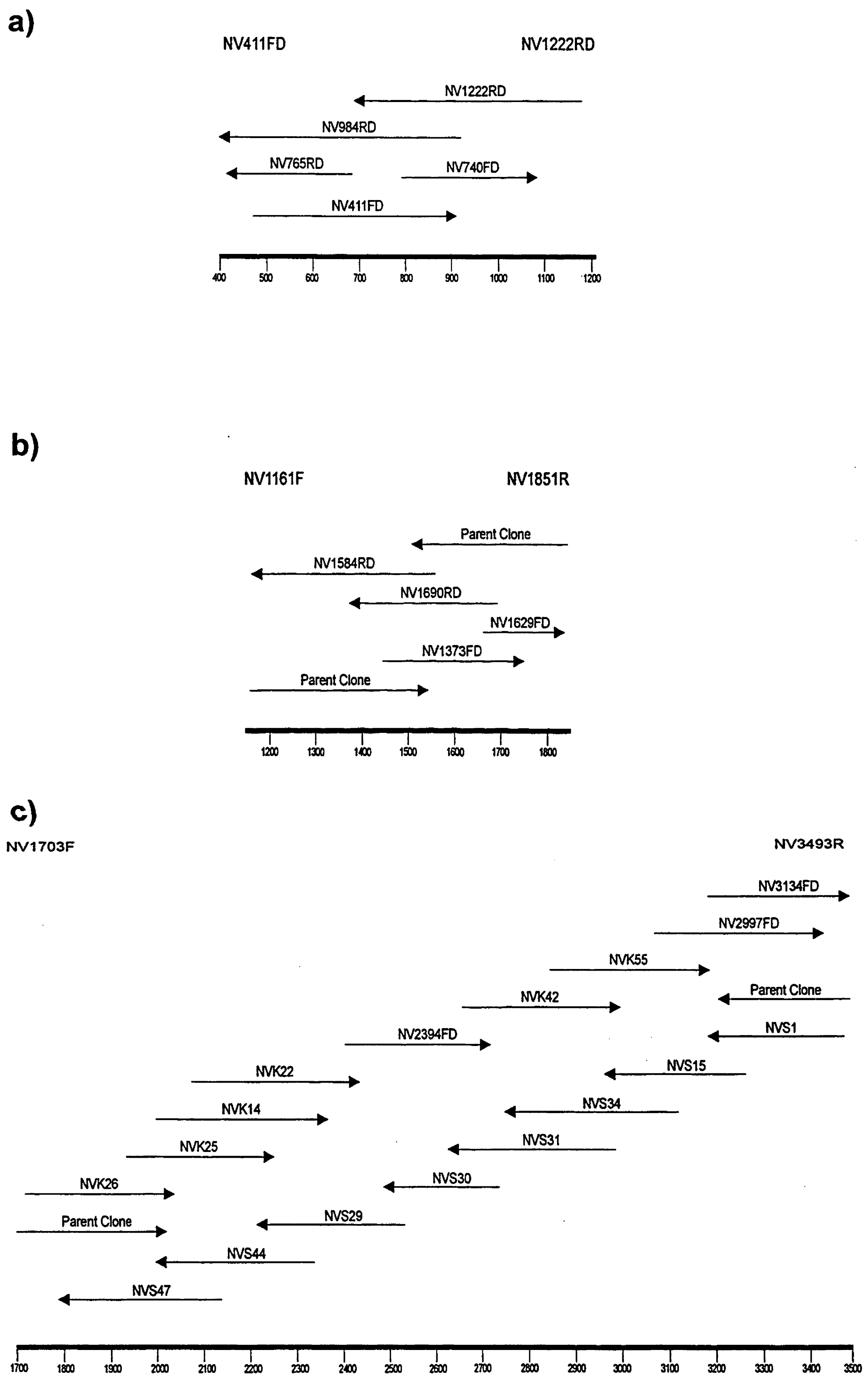
Maps showing the overlapping clones used to determine the sequence of the rDNA repeat unit of *N. vespula*. The primers used to amplify each region are shown above each map. The scale bar indicates the position of each sub-clone or each region determined by direct sequencing. Each region determined by direct sequencing is identified by the primer used (Section 2.4) or is otherwise identified by the name of the sub-clone. Each arrow and its direction represent the size and the direction of the rDNA insert sequenced as a sub-clone or the results of direct sequencing. Position 1 indicates the first nucleotide of the SSU rRNA gene.

a. A region of the SSU rRNA gene from its 400th to the 1200th nucleotides.

b. A region of the SSU rRNA gene, internal transcribed spacer, and LSU rRNA gene encompassing the 1150th SSU to the 500th LSU rRNA gene nucleotides.

c. A region of the LSU rRNA gene encompassing the 400th to 2200th nucleotides.

d. A region of the LSU rRNA gene, non-transcribed spacer, 5S rRNA gene and SSU rRNA gene encompassing the 2000th LSU to the 450th SSU rRNA gene nucleotides.


a)
NV411FD
NV1222RD
NV1222RD
NV984RD
NV765RD
NV740FD
NV411FD
400
500
600
700
800
900
1000
1100
1200
b)
NV1161F
NV1851R
Parent Clone
NV1584RD
NV1690RD
NV1629FD
NV1373FD
Parent Clone
1200
1300
1400
1500
1600
1700
1800
c)
NV1703F
NV3493R
NV3134FD
NV2997FD
NVK55
Parent Clone
NVK42
NVS1
NV2394FD
NVS15
NVK22
NVS34
NVK14
NVS31
NVK25
NVS30
NVK26
NVS29
Parent Clone
NVS44
NVS47
1700
1800
1900
2000
2100
2200
2300
2400
2500
2600
2700
2800
2900
3000
3100
3200
3300
3400
3500

d)

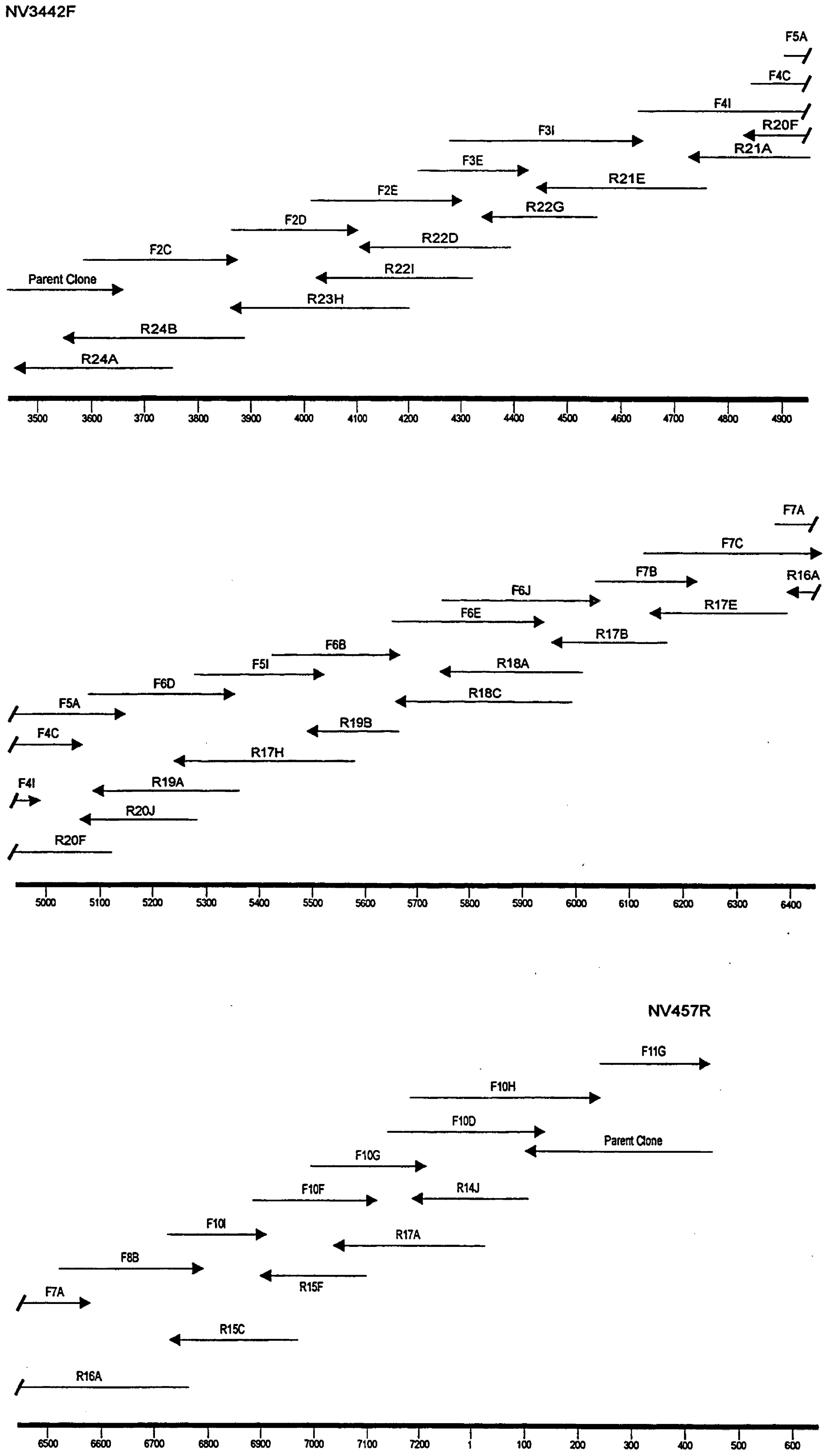

NV3442F
F5A
F4C
F4I
R20F
F3I
R21A
F3E
R21E
F2E
R22G
F2D
R22D
F2C
R22I
Parent Clone
R23H
R24B
R24A
3500
3600
3700
3800
3900
4000
4100
4200
4300
4400
4500
4600
4700
4800
4900
F7A
F7C
F7B
R16A
F6J
F6E
R17E
R17B
F6B
F5I
R18A
F6D
R18C
F5A
R19B
F4C
R17H
F4I
R19A
R20J
R20F
5000
5100
5200
5300
5400
5500
5600
5700
5800
5900
6000
6100
6200
6300
6400
NV457R
F11G
F10H
F10D
Parent Clone
F10G
F10F
R14J
F10I
R17A
F8B
R15F
F7A
R15C
R16A
6500
6600
6700
6800
6900
7000
7100
7200
1
100
200
300
400
500
600

Figure 6.2

The entire rDNA repeat unit of *N. vespula.* The approximate location of the first nucleotide for rRNA genes and spacer regions was determined from the secondary structure models. The first nucleotide in the sequence is the 1st nucleotide position of the SSU rRNA gene. Position 1244 is the 1st nucleotide position of the ITS rDNA sequence. Position 1270 is the 1st nucleotide position of the LSU rRNA gene. Position 3818 is the 1st nucleotide position of the first non-transcribed rDNA sequence. Position 4098 is the 1st nucleotide position of the 5S rRNA gene. Position 4217 is the 1st nucleotide position of the second non-transcribed rDNA sequence. Positions 4881, 4996, 5116, 5229, and 5348 are the 1st nucleotide positions of five rDNA sub-repeat sequences.

*Nosema vespula*

```
   1 CACCAGGTTG ATTCTGCCTG ACGTAGACGC TATTCCCTAA GATTAACCCA
  51 TGCATGTTTT TGACATTTGA AAAATGGACT GCTCAGTAAT ACTCACTTTA
 101 TTTTATGTAC ATTTGAAACT AACTACGTTA AAGTGTAGAT AAGATGTGTA
 151 CAGTAAGAGT GAGACCTATC AGCTAGTTGT TAAGGTAATG GCTTAACAAG
 201 GCAGTGACGG GTAACGGTAT TACTTTGTAA TATTCCGGAG AAGGAGCCTG
 251 AGAGACGGCT ACTAAGTCTA AGGATTGCAG CAGGGGCGAA ACTTGACCTA
 301 TGGATTTTAT CTGAGGCAGT TATGGGAAGT AATATTATAT TGTTTCATAT
 351 TTTAAAAGTA TATGAGGTGA TTAATTGGAG GGCAAATCAA GTGCCAGCAG
 401 CCGCGGTAAT ACTTGTTCCA AGAGTGTGTA TGATGATTGA TGCAGTTAAA
 451 AAGTCCGTAG TTTATATTTA AGAAGCAATA TGAGGTGTAC TGTATAGTTG
 501 GGAGAAAGAT GAAATGTAAC GACCCTGACT GGACGAACAG AAGCGAAAGC
 551 TGTACACTTG TATGTATTTT TTGAACAAGG ACGTAAGCTG GAGGAGCGAA
 601 GATGATTAGA TACCATTGTA GTTCCAGCAG TAAACTATGC CGACGATGTG
 651 ATATGATATA TTTTGTATTA CATAATAGAA ATTAGAGTTT TTTGGCTCTG
 701 GGGATAGTAT GATCGCAAGA TTGAAAATTA AAGAAATTGA CGGAAGAATA
 751 CCACAAGGAG TGGATTGTGC GGCTTAATTT GACTCAACGC GAGGTAACTT
 801 ACCAATATTT TATTATTTTG AGACGATTTT TAATCAGAGA ATGATAATAG
 851 TGGTGCATGG CCGTTTTCAA TGGATGCTGT GAAGTTTTGA TTAATTTCAA
 901 CAAGACGTGA GACCCTTTTA TTATAGACAG ACACAATCAG TGTAGGAAGG
 951 AAAGGATTAA AACAGGTCCG TTATGCCCTC AGACATTTTG GGCTGCACGC
1001 GCAATACAAT AGATATATAA TCTTTATGGG ATAATATTTT GTAAGAGATA
1051 TTTGAACTTG GAATTGCTAG TAAATTTTAT TAAATAAGTA GAATTGAATG
1101 TGTCCCTGTT CTTTGTACAC ACCGCCCGTC GCTATCTAAG ATGATATATG
1151 TTGTGAAATT AGTGAAAACT ACTTGAACAA TATGTATTAG ATCTGATATA
1201 AGTCGTAACA TGGTTGCTGT TGGAGAACCA TTAGCAGGAT CATAATGAAT
1251 AAAAAAACAT TTTGTAGATT AGATTAAATT TTGCCCACAC ATGGGATCAA
1301 TAGGATACCA TAACGATGAA GGTCGTAAAA GAATACGAAA GAGTATTTAC
1351 CGAATTAATA TATTTATATA TTGATAACCC TTTGAACTTA AGCATATCAT
1401 TAAAAGGAGG AAAAGAAACT AACTAGGATT TCTTTAGTAG CGGCGAGTGA
1451 ACAAGAATCA ACCCTTGATT GTAATCCTTA ACTGGAGATG TAAATCATTT
1501 ATTCTATTAT TTATATCATA GAGAATTTTA AATTCGTTAG TTGATAGAAT
1551 AGAATAATTT CTTTGAGTAG GGCTGCTTGG TAGTGCAGTT TGAATACAGG
1601 TAGAATGAGA TATCTAAGGT TAAATATAAT GGTATACCGA TAGCAAATAA
1651 GTACTGCGAA GGAATTTGTG AAAATGTGTG GGTGTTAGCC TTATTTTTAA
1701 GGACCCGTCT TGAAGCACGG ACCAAGGAGA TTATAATTAT AGCAAGATAA
1751 AAACTATTTA ATCGTTATTA ATTTGATAAG TTATAATTAT AAGACCCGAA
1801 ACACAGTGAA CTATACATGT TCTGGTTGAA GACAAGCAAC AGTTTGTTGG
```

```
1851 AAGACCATAA TCATTCTGAC GTGCAAATCG ATGATTTAAG ATGTGTATAG
1901 TGGCGAAAGA CCAATCGAAC TGTGTGGTAG CTGGTTCACA GCGAAATGTC
1951 TCTCAGGACA GCAGTCATTT TATATAGGAC ATAGATGTAG GACACTGTTA
2001 TAGTATTTAT ATTATGAGAA ATCTTTAATT CTATGGAAGT CTATAAATTA
2051 ATTTTATAGG GTGAATCTAT TGTATGACTA GTGGGCACAT GATTGTAAGA
2101 ATGATGTGCA AAAAGGGATG AACCTTATGA AACATTAAAT ATTCTAAATA
2151 GTAGATATTA TACCATAAAT ATGATGAATA CATTGAGACA GTAGGGCGGT
2201 TGTTATGGAA GTAGAAATCC GCTAAGAAAC GTGTTACAAC GTACCTACCG
2251 AATGTATTAT TGTATAAAAT GGAAGAAGAT TACTACTTTT ATGAGATGTT
2301 TCTATAAATT ATTTAGGTAG CTGTGCAATT TGTTTGTATT GAAGTATGTA
2351 TGTGAGTACA TATAGAAGAA ACAAATGAGC CGATCTTGGT GGCAGTAACA
2401 ATATATTTGT ATAGATATAG ACTAGGGTTT TCAATTATTA TTGAAGTGAA
2451 TCGGTGTTAT TTATAAAAAC AAAGAAGATT AATAATTCTT CATTAAATTA
2501 CTGTAAGGTG ACTTGATGAT GACAGTATAT TTACAAGTTT AAGAAAAATG
2551 AATTAAACTT TAATATAGCT ATGGATTGTC ATGATAAGAA ATAGTTATTT
2601 ATATATGCTT GATAAAAATT ATGTAATTTT AGCCGTACCT ATACCGCATC
2651 AGGTGTCTTT GTATCATAAC AAATATTTTA AATTAAAGTA AGTAAGGGAA
2701 TTCGGCAAAT TAGATCTGTA ACTTTGGGAT AAAGATTGGC TCTAGCATGC
2751 TAGAACTTTT ACTATGTAAG GAATCTGACT GTTTATTAAA AACATAGCTT
2801 TTTGTATTTA CAAGAAGTGA ATTCTGCCCA GTGCATTTAT TGTTAAAGTT
2851 GTGTAAGCGA ATGTAAACGG CGGGAGTAAC TATGACTCTC TTAAGGTAGC
2901 CAAATGCCTC GTCATTTAAT TGGTGACGCG CATGAATGGA GCAACGAGAT
2951 TCCTACTGTC CCTACTTACA ATTTTGTGAA ACCACGAGAC AAGGGAACGG
3001 GCTTGTTGAA TATCAGCGGG GAAAGAAGAC CCTGTTGAGC TTGACTTTAG
3051 TATGTTCTAA TAATTATTCG GTATATTGTA GAGAGGTGGG AGATGTATTG
3101 TGAAACCACT AGTATGTTTG AATATTTATT TATTATGAAT ATATGGGGAG
3151 TTTGGCTGGG GCGGTACAGC TGTTAAAAAG TAACACAGCT GTCCTAAGGT
3201 AGATGAATGA TGGATGGTAA CCATCAGTTT ATTATAAGGG AATAAATCTG
3251 CTTTACTTTA TGCATATAAC TGTATTTAGT AGGGAAACCT TGGCCTAGCG
3301 ATCCCAATTA CATTCATTAT GTATTGGGTG ATTGAAAAGT TACCACAGGG
3351 ATAACTGGCT TGTAGCAGGC AAGCGATCAT AGCGACTCTG CTTTTTGATT
3401 CTTCGATGTC GTCTCTTCTG ATCATCGTAG TGTATATGTT ACGAAGTGTT
3451 GGATTGTTCA CCCGGTAATG GGGAACGTGA GATGGGTTTA GACCGTCGTG
3501 AGACAGGTTA GTTTTACCCT ACTGTAAGTA TTTTTGAGTG AACTTTGATA
3551 GTACGAGAGG AACTCTAAGT GATGACCACT GGTAGTGCGA TTAACTATTA
3601 AGTTATGTTG CTTAGCTACG TCGTTTCGAT TAAGGCTGAA AGCCTCTTAA
3651 GCCTGAAGCG TAGCTCAAAG ATACATTTAG GAGAATACTA CTCTATTAAG
3701 CAGAACTGTA AGAATGAAGG TGTATGCCGA CATTTTGAGG TTACTGCTTT
3751 TTATGTGTTG GAGTAGGTTA ATTTTAATAT TGTAAGTTTG ATTTATTATT
3801 ATTTATTAGT GTAATTTATA GTTTGTTTAT TAGTTAGTGT GGTTATAATT
3851 AGCGTGGTTT ATAGTTAGTG TGGCTATTAC TTTAATAAGT TTGATTATAA
```

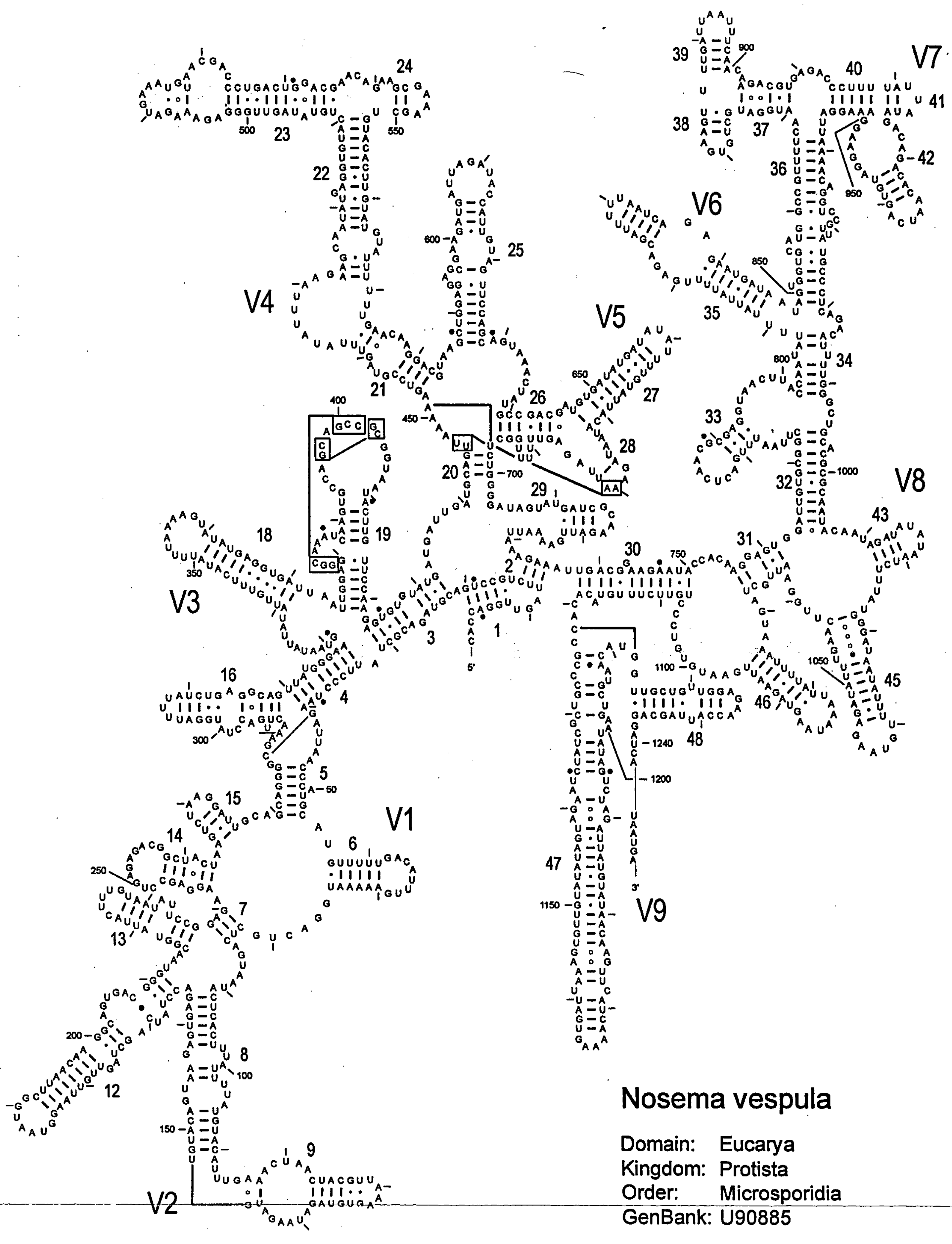
```
3901 TAATACTTTA TATTATTAAC TAACCTAAGC TAAGTTAAAT TAAATATCAT
3951 GTTTACAGTA ATGATATACT CTATTTAAGT GTGCAAATAA ACCTTTGCTT
4001 TATTGCTTAA TGAGTTATTC CTTTAAATGG TGATTAATTA AATATTTTTT
4051 TACGTACATT TACTAATGTT ACTTAAGATA TAGAAAAGTA ACTCATACAA
4101 TTACGGCCAT ATCTACTGTA AAGCACCGGA ACTCGTCAGC TCTCCGAAGT
4151 TAAGCCAGTA AGAGCCTAGT TAGTACTAAG ATGGGAGACC GCTTGGAAAA
4201 GCTGGGTGCT GTAATTTTTT AATCAAACTT ATCTATGTTT TTACTTGTAT
4251 AAAAAGTTTA TTTATTATTA AACACATGTA TTTTTAATTA AACTTTTCTT
4301 TGTTATTACT TGTATTATAA CTTAATTAAA AAAGTTCTAA ACACATGTAT
4351 TTTTTTTTAA ACTTATCTAT GTTTTTACTT GTATAGATAG TTAATTAAAA
4401 AGTGTTAAAC ACATGTGTAA TTATTTAATT AAACTAATTA TGTTTTAACT
4451 TGTATAGATA GTTAATTAAA AAGTTCTAAT AAATGTATAA TTATTTAATT
4501 AAATGTATCT ATGTTATTAC TTGTAAAAAA ATAGATTTGA TCAAAAAGTT
4551 ACCGGTTGGT TGGGTTAGGG GGTGTTCAGA TTGGTGAATG GGGTGTTGAT
4601 CCTTCTAATG GGTCGTGGGT TGGGGGTTTG TGGTGATTGA ACCAAGTTGG
4651 GTGAGTGACC TGTAACTGAA TGTAAAATGA TTCCTATTGA ATAATACCTA
4701 TAGAGAGAGA TTATATTTTT AAATTCCTAA ATCTGGTATT ACTAGAAAAG
4751 ACACATTATA GTTAAATAGC TCTTGTAATT TTCTTACTTT AAATGTTTGT
4801 TATGTAAGAT ATTCTCATGT TATTAGTCTG ACAATTAAAT GTTTATAAGG
4851 GCTATTCAAA TAAAATTAAT AAAAAAGTGA TAGATTTCTT GTTAGTTAGT
4901 TAAGTGATGT TTATAATATT AGTAGCACTA TACTAGTATA TATCGTAGAC
4951 CTTATTTCAT AGGTTAGTTG TTTAAATATG ATTAATAAAA TTATATTAGA
5001 TTTCTCGTTA GTTAGTTAAG TGATGTTTAT AATATTAATA ACATTATGCT
5051 AGTGTATATC TTAGACCTTA TTTTATAGGT TAGGTTAGTT GTTTAAATAT
5101 GATTAGTAAA AAAGTGCTAA ATCTGTCTTT AGTTAAGTGA TGTTTATAAT
5151 ATTAGTAGCA TTATAATGGT ATATTATGTA TATCATATTT CATAGGTTAG
5201 TTGTTTAAAT ATGATTAATA AAATTATAAT AGATTCCTCG CTAGTTAGTT
5251 AAGTGATGTT TATAATATTA GTAGCACTAT ACTAGTATAA TATGTATATC
5301 ATATTTAATA GGTTAGGTAA GTTGTTTAAA TATGATTATA AAAAAGTGGT
5351 AGATTTCTCG TTAGTTAGTT AAGTGATGTT AATTATTTCA ATAACATTAT
5401 AATGGTATAT TATGTATAAT CACATTTTAT AGGTTAGTTG TTTAAATATG
5451 ATTATTATAG TTTAGATATG CTAATCAAAA TAGTTTATAT TCATTTAACT
5501 ATGCACTAAA TGTTAAGTAA AGATAATAAC ATTTAAGTCA TTTAACTAAA
5551 AAAGTTAGTA GTTAATTAAA ATAAATACAT TTGGTATGGT TTTTTACATA
5601 GTAATTTAAA TTACAATGCT ACAATATATA TTTTAAGTAA CGTTCCATAC
5651 ACTTTATAAA AATTTAGCCC TTAAAAGTAA ATTATTCATC TGATTAGTAA
5701 AATTAATAAC GTTCATTATT ATTTTAAGAA AAGATGACTA TAAGATATAT
5751 AACTATTTTA TTATCATTAA GTTTGTGGTG AAAACTTAAG TTTAATTAAA
5801 TTTTAAAGTG TGATTTTAAA AAGACCGGGA AAAAATATTT AACTTGTTAA
5851 TTTTTAAATT TTAAGACTAA GTACTTTAGT TGTTAACTAA TTATTACTCA
5901 TAATAATATG TTGTTATGAT GTACTTTGTA TGATAAAAAA TATAATTAAA
```

```
5951 ATCTTTAGTT AAATATTTTA ACACCATATG TAAATTAACT TAAAATTAGT
6001 TGTCAATAAT GTTGATTAAA GTTTATTCTT ATTAGAGTCT TTTTGATCAA
6051 AGTCATTTAT ATTATTATTA ATCTATGACT TTTAAGTAAT TTATAGTTAA
6101 TGTATTAAAG CCTTTATACA ATTATGGATA TAAAAGATAT AAATCAGAAC
6151 TTTTTATATG AAAATAAAAT TACTATATAG TGTAATAACT ACAAGTTCTT
6201 TTTGATGTTT AAGTTATTTT TTATAGATTC TTTATTACTT TTCTATATAA
6251 TTAAATTAAT CTTACTATTA TGAGTCATAT TTTAAATAAT AATCACTCAT
6301 TACTATAATT ATTTAAATCT TTTCCCATTT GTAAATATAA ATAAGTTTTC
6351 TTTTATATAC ATATACATGC ATGTCTATCA ATTCATACCA AACCTTCCCA
6401 GTAGGTACTT GTTAAAGCTT AATTTTAAAC GTCATGAAAG GATCAAGGGT
6451 TATTTAGCAA AAGTGAGGGG TGGCATGTTT AAAGGCTTTA ATAAGTGTAG
6501 TTAGTGTACT TAGTAGGGTT AGTAGTTAAA TTAAATCTAG GTTAGTTAGG
6551 TAGTGTAGTG TGGTTTCTAC AATAATATGC AAAGCTGGGT ATACGGACAA
6601 CTTTGATAGA ACCTCTAATC TGAACCGTAG GCATTAAGTT GCCTTTTTGT
6651 TATATTAATT TTTATGCATT AATTAAAATA TATATCTTAT AAAATACATC
6701 TAATTATAAA GGACATTTTT ATAAAACTTT AATGTTATTT CTTCATATAC
6751 TTCTAATAAC ATCTTCTCCT ATTCATTTTT TATATCTTTT TTTACTGTGC
6801 CCGTAAAGGT GCAGTTAATT AATCTTATAA AATTAATTAT TTATGCATTT
6851 ATAACAATAT GCATCTTATT ATAAGAACAT CATTCTATTC TATTCATCAT
6901 TTTATTCTAT TACCTCAATC TAATTACATC ATTCTATTCA ATTACATCAC
6951 TCCATTCATC TATTACCCTA TTTACATCAT TATATTCATC TTTTTATTCA
7001 CAGGTAATCT TAAACTGGTC AAACATATCA TACATTTACT AGACTGTAAA
7051 TAAGTAATTG GTCAATCAAA TGCTTATTTA TTATTTTAAT TTTTTACCCT
7101 CTTTTTTTAA AGATTAATGC TTCTTGCAGG TTTAACTTTA AATTCAAACA
7151 TGATATTATA TCTATTTTCT TGTATCTTTC TTTTATTATA GTTTTCATAT
7201 ATTATATTTC CATATAAATA TGTATTTAGA AATTATAAAT TTCTATTATT
7251 CTATACAATA CTGATTATAA ATCTTATATT AGTATTTAAA TA
```

Figure 6.3

Putative secondary structure model for the SSU rRNA of *N. vespula*. The model is drawn in the format of Gutell (1994a). Structural components are numbered according to Neefs *et al.* (1991). Every 10th nucleotide is marked with a tick and every 50th position is numbered. Nucleotides marked with an adjacent dot are positions that are highly conserved within the three primary lineages and the character state shown for *N. vespula* is prokaryotic. Regions marked with a 'V' and a digit are hypervariable regions. Abbreviations are as in figure 4.3 and further details are described in the text.

# Secondary Structure: small subunit ribosomal RNA



## Nosema vespula

Domain: Eucarya
Kingdom: Protista
Order: Microsporidia
GenBank: U90885

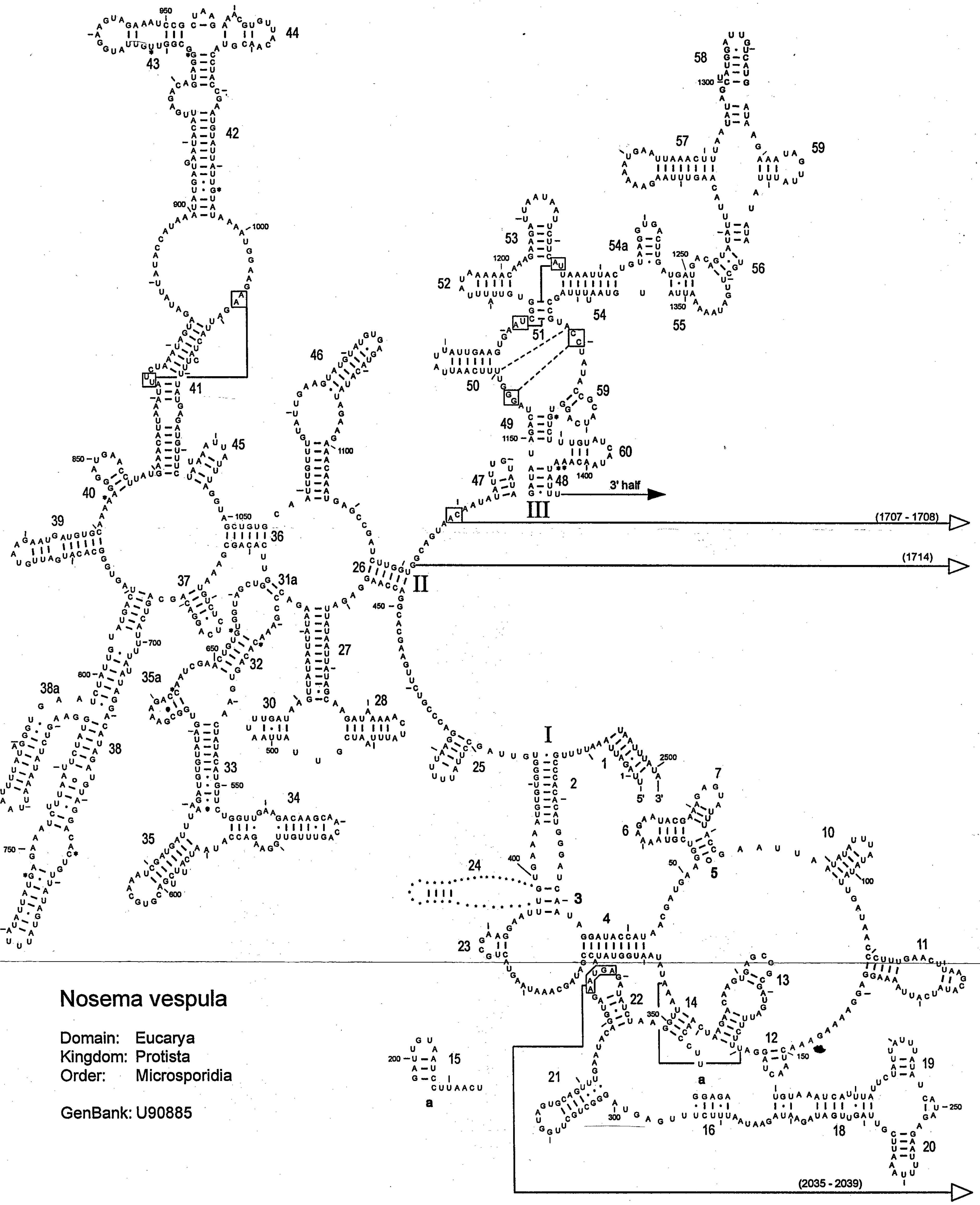
Figure 6.4a and b

Putative secondary structure model for the LSU rRNA of *N. vespula*. The model is drawn in the format of Noller *et al.*, (1981) and modified by Gutell *et al.,* (1993). Structural components are numbered according to Noller *et al.*, (1981). Every 10th nucleotide is marked with a tick and every 50th position is numbered. Nucleotides marked with an adjacent dot are positions that are highly conserved within the three primary lineages and the character state shown for *N. vespula* is prokaryotic. Abbreviations are as in figure 4.3 and further details are described in the text.

a. The 5' region of the putative LSU rRNA model.

b. The 3' region of the putative LSU rRNA model.

## Secondary Structure: large subunit ribosomal RNA - 5' half

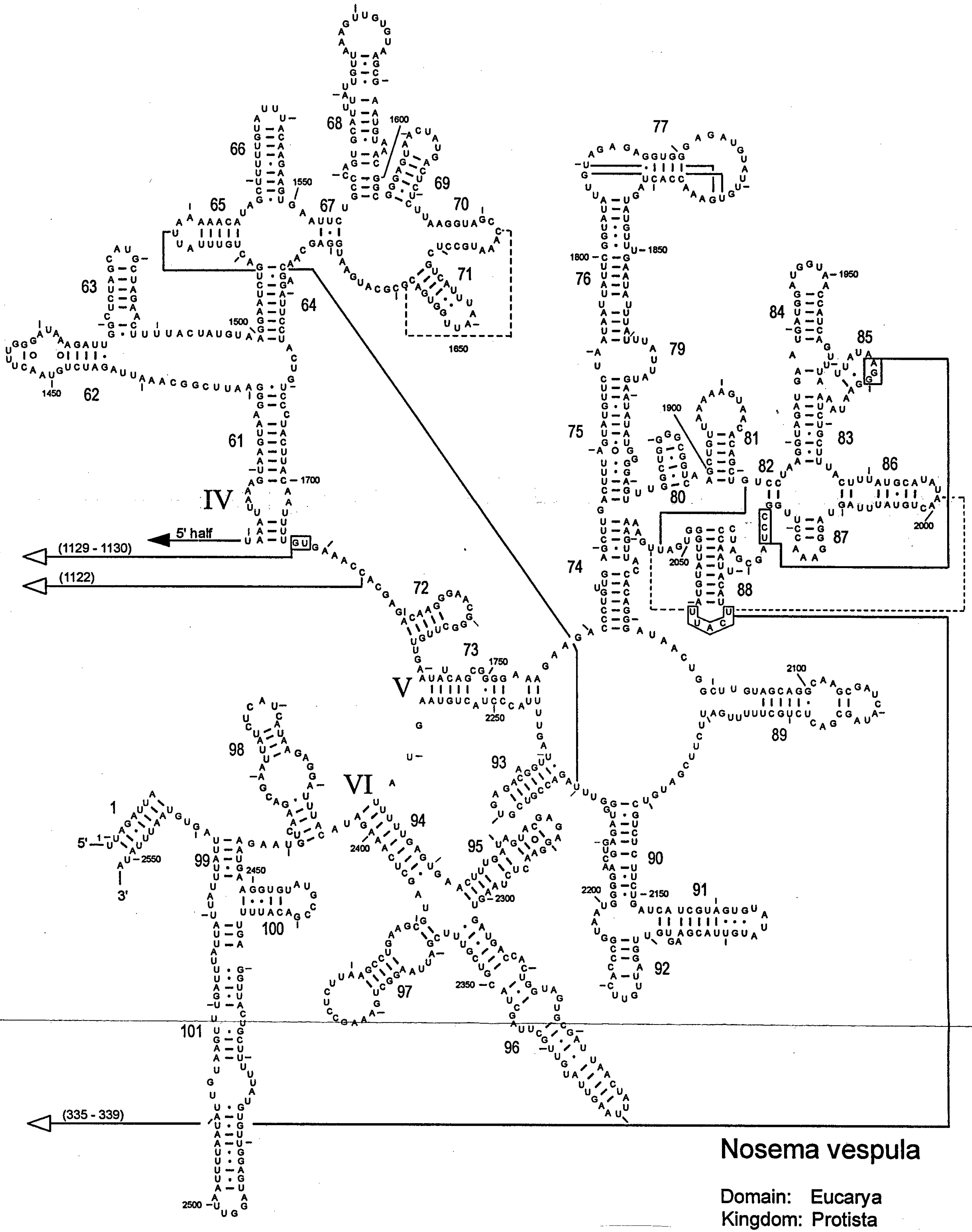


Nosema vespula

Domain: Eucarya
Kingdom: Protista
Order: Microsporidia

GenBank: U90885

# Secondary Structure: large subunit ribosomal RNA - 3' half



## Nosema vespula

| | |
|---|---|
| Domain: | Eucarya |
| Kingdom: | Protista |
| Order: | Microsporidia |

GenBank: U90885

Secondary Structure: 5S subunit ribosomal RNA

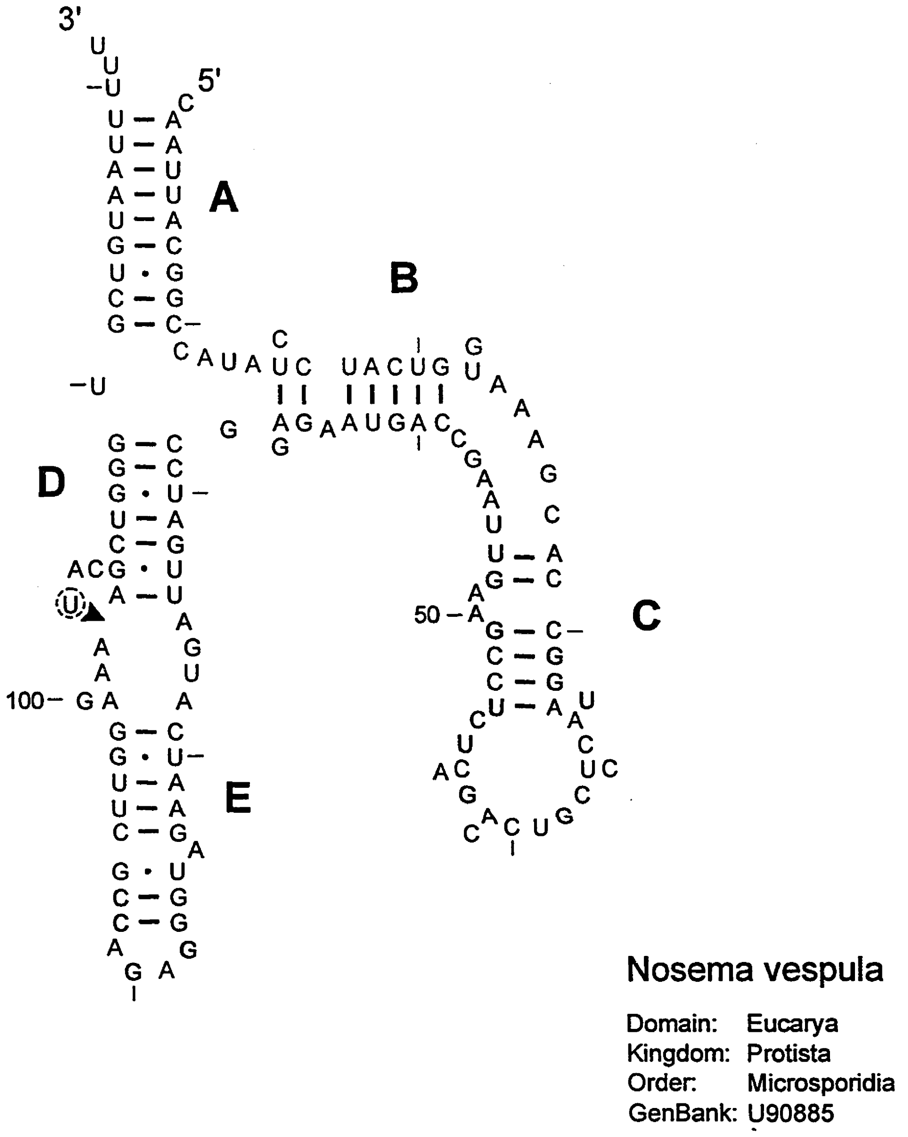


Figure 6.5

Putative secondary structure model for the 5S rRNA subunit of *N. vespula.* The model is drawn in the format of Wolters and Erdmann (1986) and Barciszewska and colleagues (1996). Every 10th nucleotide is marked with a tick and every 50th position is numbered. Nucleotides positions with two or more character states display the *N. vespula* character state within the structure and the eukaryotic character state adjacent to the structure.

Table 6.1

Nucleotide composition by region for the entire rDNA repeat unit of *N. vespula.* The various regions were determined from the putative secondary structure models of the SSU, LSU and 5S rRNAs. The table indicates for each region: the total number of nucleotides; the number of nucleotides possessing each alternative character state; the percent composition for each nucleotide character state; and the overall A + T and G + C composition expressed as a percentage of the total.

| Region | Nucleotide | Total | Percentage Composition | A + T% | G + C% |
|---|---|---|---|---|---|
| SSU gene | A | 403 | 32.37 | | |
| | C | 172 | 13.82 | | |
| | G | 286 | 22.97 | | |
| | T | 384 | 30.84 | | |
| Subtotal | | 1245 | | 63.21 | 36.76 |
| ITS rDNA | A | 12 | 50.00 | | |
| | C | 1 | 4.17 | | |
| | G | 3 | 12.50 | | |
| | T | 8 | 33.33 | | |
| Subtotal | | 24 | | 62.50 | 37.50 |
| LSU gene | A | 845 | 33.15 | | |
| | C | 319 | 12.52 | | |
| | G | 541 | 21.22 | | |
| | T | 844 | 33.11 | | |
| Subtotal | | 2549 | | 66.26 | 33.74 |
| NTS1 rDNA | A | 94 | 33.69 | | |
| | C | 25 | 8.96 | | |
| | G | 38 | 13.62 | | |
| | T | 122 | 43.73 | | |
| Subtotal | | 279 | | 77.42 | 22.58 |
| 5S gene | A | 35 | 28.45 | | |
| | C | 26 | 21.14 | | |
| | G | 30 | 24.40 | | |
| | T | 32 | 26.07 | | |
| Subtotal | | 123 | | 54.47 | 45.53 |
| NTS2 rDNA | A | 1088 | 35.42 | | |
| | C | 296 | 9.63 | | |
| | G | 378 | 12.30 | | |
| | T | 1310 | 42.64 | | |
| Subtotal | | 3072 | | 78.06 | 21.94 |
| Over all | A | 2477 | 33.97 | | |
| | C | 839 | 11.50 | | |
| | G | 1276 | 17.50 | | |
| | T | 2700 | 37.03 | | |
| Grand Total | | 7292 | | 71.0 | 29.0 |

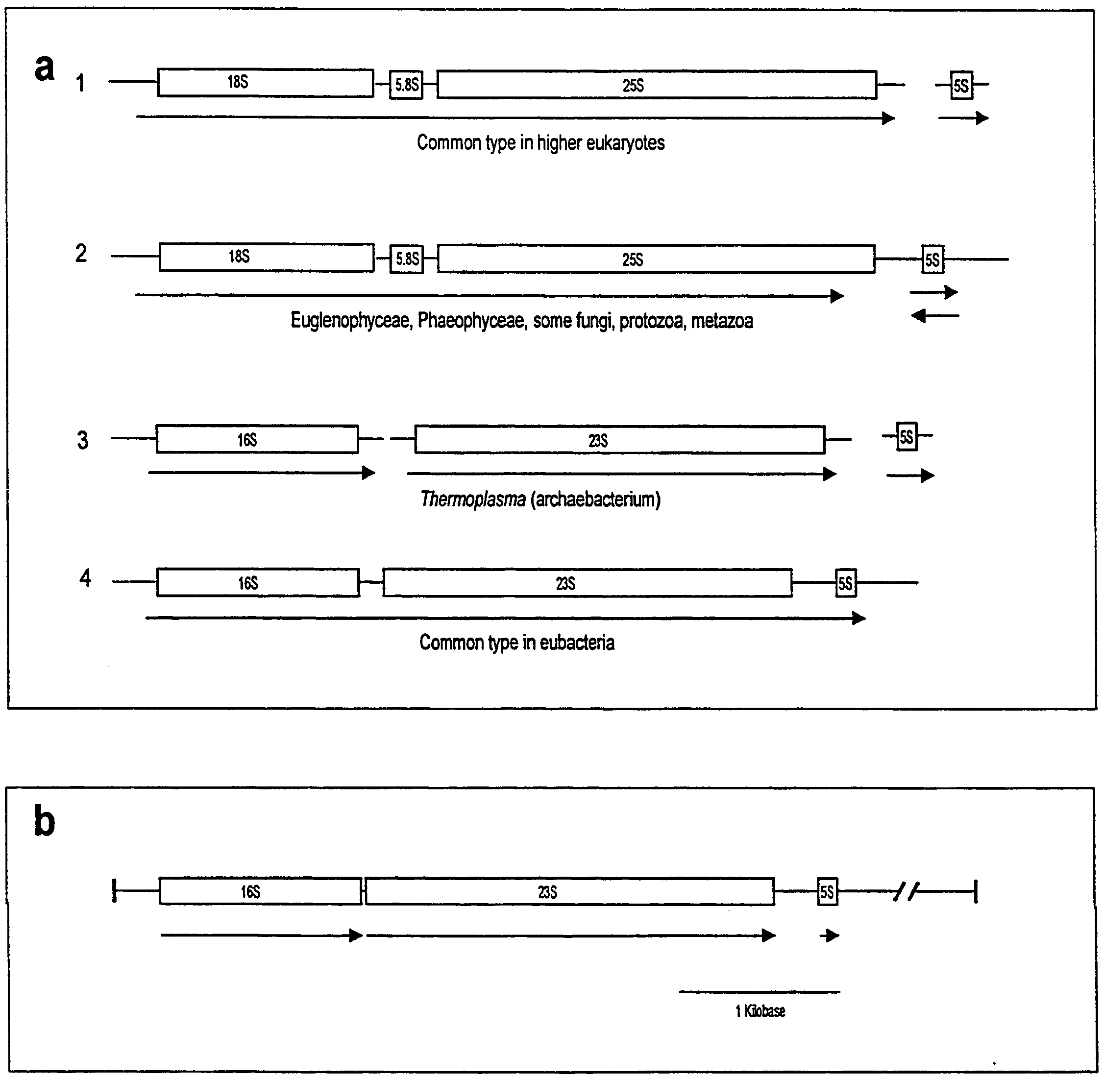

a
1
18S
5.8S
25S
5S
Common type in higher eukaryotes
2
18S
5.8S
25S
5S
Euglenophyceae, Phaeophyceae, some fungi, protozoa, metazoa
3
16S
23S
5S
*Thermoplasma* (archaebacterium)
4
16S
23S
5S
Common type in eubacteria
b
16S
23S
5S
1 Kilobase

Figure 6.6

Schematic representation depicting the organisation of rRNA genes within their respective operon types (after Kawai *et al.*, 1995). Arrows indicate direction and regions of transcription by either RNA polymerase I and III.

a. Known arrangements of rRNA genes within differing operon types.

b Arrangement of rRNA genes within the rDNA operon of *N. vespula*.

```
                                                                  50
Repeat 1  -.........T.....................................C.
Repeat 2  T.........................................A..A....
Repeat 3  ---.C.AAA..TG.CT..................................
Repeat 4  .......C....C...................................C.
Repeat 5  -...............................A..T..T.C.A..A....
Consensus ATAGATTTCTCGTTAGTTAGTTAAGTGATGTTTATAATATTAGTAGCATT

                                                                 100
Repeat 1  ............ATC...G.C-.T.....C-----...............
Repeat 2  ..G.....G...ATCT..G.C-.T...........T..............
Repeat 3  ...A.G...............-.......C-----...............
Repeat 4  ...........A.........-.......A.....T.....A........
Repeat 5  ...A.G..............A...C.....-----...............
Consensus ATACTAGTATATTATGTATATTCATATTTTATAGGATAGGTTAGTTGTTT

                                  122
Repeat 1  .................T....
Repeat 2  ...........G......AG.-
Repeat 3  .................T....
Repeat 4  ...........-......AG.G
Repeat 5  ...........-.T.T.G.T..
Consensus AAATATGATTAATAAAAATATA
```

Figure 6.7

Sequence alignment of the five repeat sequences located within the intergenic spacer of *N. vespula*. The sequences form a contiguous block in numerical order from repeat one to repeat five.

| REPEATS | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 115 | 83% | 82% | 80% | 79% |
|  | 0 | 5% | 1% | 5% | 1% |
| 2 | 101 | 120 | 70% | 83% | 74% |
|  | 7 | 0 | 7% | 1% | 7% |
| 3 | 95 | 85 | 113 | 76% | 78% |
|  | 2 | 9 | 0 | 7% | 3% |
| 4 | 97 | 101 | 93 | 120 | 75% |
|  | 7 | 2 | 9 | 0 | 5% |
| 5 | 92 | 91 | 91 | 91 | 115 |
|  | 2 | 9 | 4 | 7 | 0 |

Table 6.2

Similarity matrix for the five repeated sequences located within the intergenetic spacer. Each sequence is compared to every other sequence and two numbers are generated.

1. The number of residues that match exactly (identical residues) between the two sequences.

2. The number of residues lined up with a gap character.

Each number is expressed as a count and as percentage on different sides of the matrix diagonal. The diagonal shows how many locations have at least one residue for the single sequence.

Figures 6.8a - f

A comparison of SSU rRNA secondary structure elements. The secondary structure elements are drawn in the format of Gutell (1994a). Structural components are numbered according to Neefs *et al.* (1991). The secondary structures for eukaryotes (based on *C. neoformans*), eubacteria (*E. coli*), Microsporidia (*N. vespula* (NV)), and Diplomonads (*G. muris* (GM)) are displayed for comparison. The eukaryotic (Euk) and eubacterial (Eub) models indicate positions that are highly conserved (90-95% lower case; > 95% upper case) in species for which sequence data is available. Regions closed by a line are hypervariable in that lineage. Regions of interest are identified with a bracket while nucleotides are identified by a pointer.

a. Helix nine

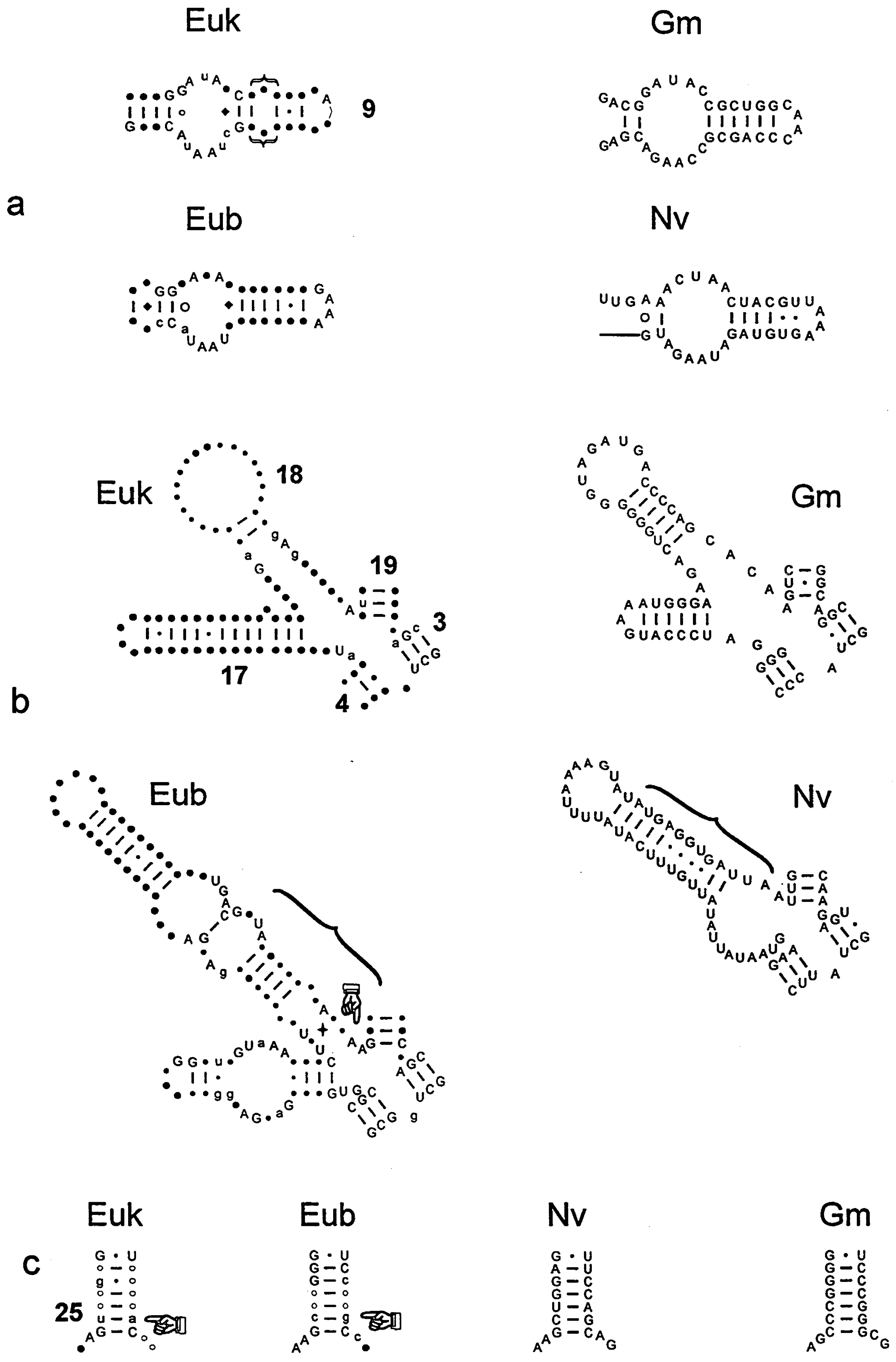
b. Helix eighteen

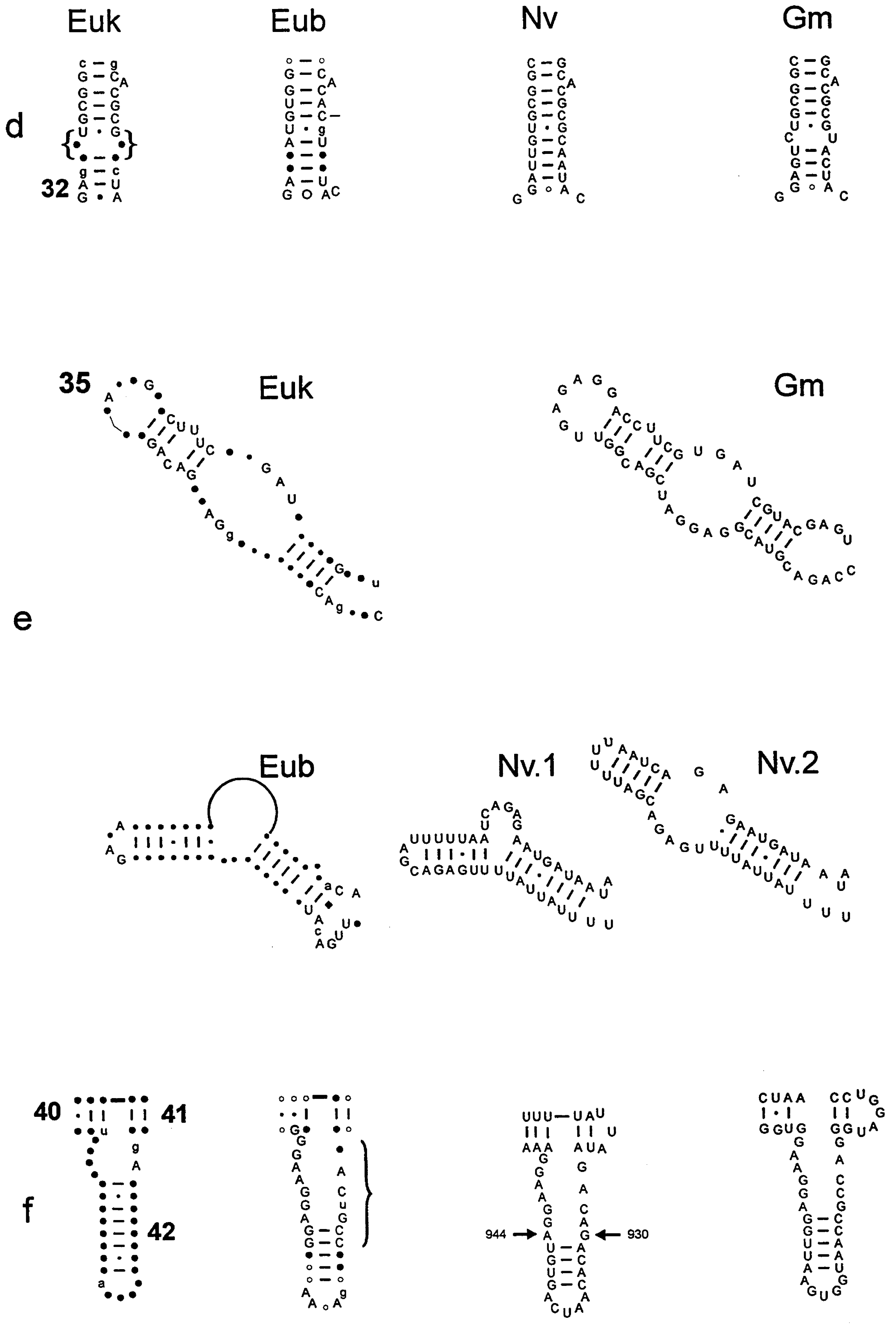
c. Helix twenty-five

d. Helix thirty-two

e. Helix thirty-five

f. Helix forty-two


a
Euk
9
Gm
Eub
Nv
b
Euk
18
19
3
17
4
Gm
Eub
Nv
c
Euk
25
Eub
Nv
Gm


Euk
Eub
Nv
Gm
d
32
e
35
Euk
Gm
Eub
Nv.1
Nv.2
f
40
41
42
944
930

Figures 6.9a - h

A comparison of LSU rRNA secondary structure elements. The secondary structure elements are drawn in the format of Noller *et al.*, (1981) and modified by Gutell *et al.,* (1993). The secondary structures for eukaryotes (based on *C. neoformans*), eubacteria (*E. coli*), Microsporidia (*N. vespula* (NV)), and Diplomonadida (*G. muris* (GM)) are displayed for comparison. The eukaryotic (Euk) and eubacterial (Eub) models indicate positions that are highly conserved (90-95% lower case; > 95% upper case) in species for which sequence data is available

a. Helix one

b. Helix five

c. Helix ten

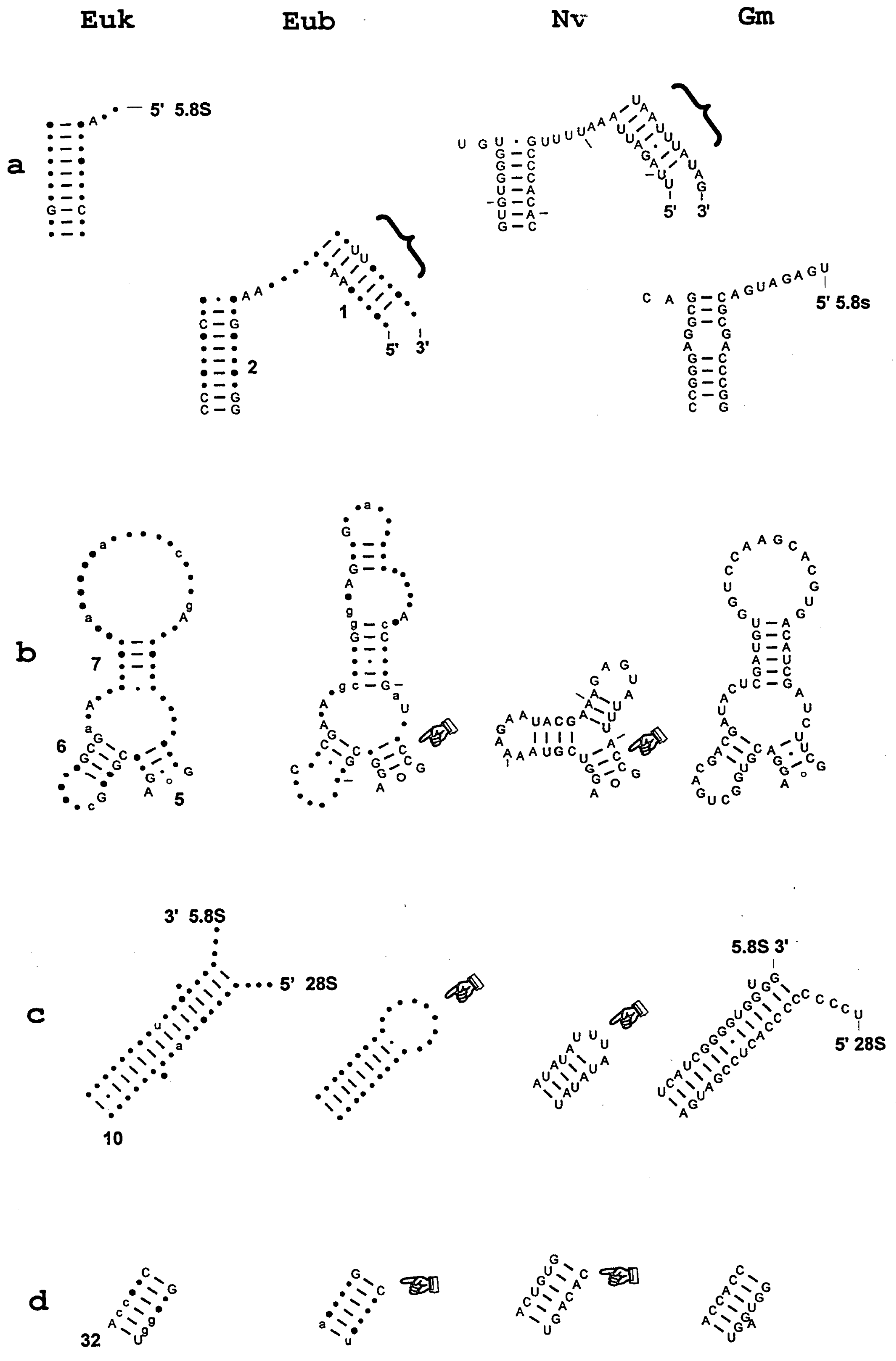
d. Helix thirty-two

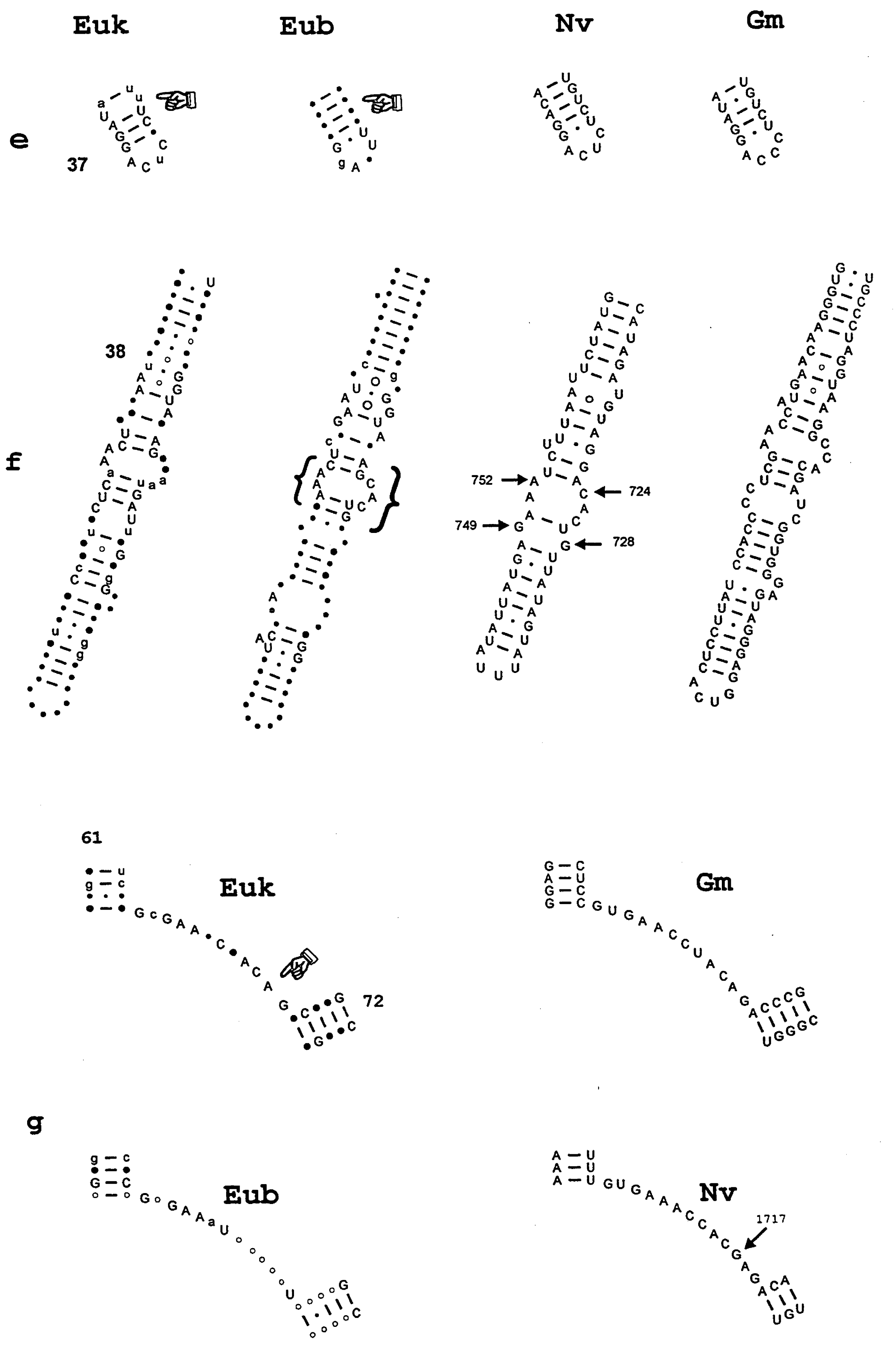
e. Helix thirty-seven

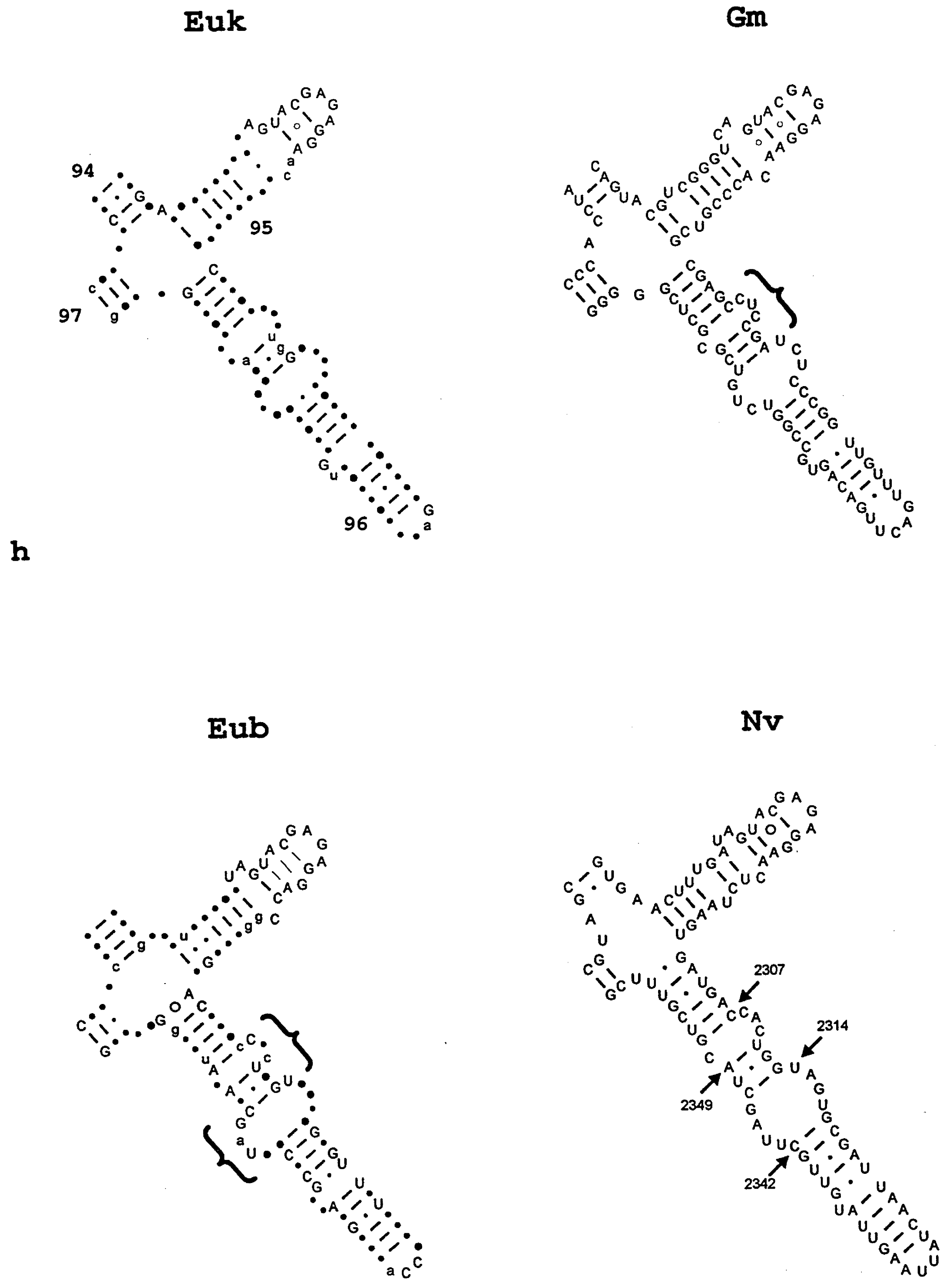
f. Helix thirty-eight, lower region

g. Helices sixty-one and seventy-two junction loop

h. Helix ninety-six


Euk
Eub
Nv
Gm
a
5' 5.8S
1
2
5' 3'
5' 5.8s
b
7
6
5
c
3' 5.8S
5' 28S
10
5.8S 3'
5' 28S
d
32


Euk
Eub
Nv
Gm
e
37
f
38
752
724
749
728
61
Euk
72
Gm
g
Eub
Nv
1717


h
Euk
Gm
Eub
Nv
94
95
96
97
2307
2314
2349
2342

# CHAPTER 7

# *Nosema vespula*: ANALYSIS OF DIVERGENCE AND PHYLOGENETIC RECONSTRUCTION

## 7.1 Introduction

During the late eighties and early nineties, published phylogenies of rRNA sequences developed into a consensus tree of eukaryotic evolution. According to the consensus tree, three amitochondrial protist groups (Microsporidia, Diplomonadida, and Trichomonadida) emerge at the base of the tree. The amitochondrial protists are followed by the Euglenozoa, then about eight protist phyla (most of which possess a lifecycle that includes an amoeboid stage). At the top of the tree is a vast terminal crown comprising the multicellular kingdoms and other protist phyla (Philippe and Adoutte, 1998). The Microsporidia, Diplomonadida, Trichomonoadida, and the pelobionts, oxymonads, retortmonads and hypermastigids, comprise the amitochondriate taxa known as the archaezoa (Cavalier-Smith, T., 1987; Sogin and Silberman, 1998). The apparent natural grouping of the taxa included in the archaezoa is premised on the absence of mitochondria, peroxisomes and typical stacked golgi dictyosomes (Cavalier-Smith, 1987; Keeling, 1998). Absence of these organelles implies that members of this group diverged at the base of the eukaryotic tree. If this is correct, the root of the eukaryotic tree lies within the protist assemblage. Detection of the eukaryotic should be possible using either phenotypic and or genotypic characters. In reality though, the true evolutionary relationship amongst protist taxa is often blurred, disguised by evolutionary forces that have lead to convergence, parallelism, and reversion of phylogenetic characters.

Recently, several lines of research have begun to question the validity of the archaezoan hypothesis including the identity of the first divergent eukaryote. First, mitochondrial genetic residues have been detected in the genome of some archaezoa. Clark and Rogers (1995) demonstrated the presence of mitochondrial homologues for pyridine nucleotide transhydrogenase and a 60-kilodalton chaperonin (cpn60) in *E. histolytica.* Mitochondrial homologues for the chaperionin gene (cpn10, cpn60, and cpn70) have been identified in *T. vaginalis* (Horner *et al.,* 1996; Bui *et al.,* 1996; Germot *et al.,* 1996; Roger *et al.,* 1996). A mitochondrial homologue for cpn70 has been identified in *N. locustae* (Germot *et al.,* 1997) and *V. necatrix* (Hirt *et al.,* 1997). A mitochondrial homologue of cpn60 has been identified in *G. lamblia* (Roger *et al.,* 1998). These results indicate that mitochondrial endosymbiosis is likely to have occurred before the evolution of the known eukaryotic lineages. Second, recent findings are critical of the phylogenetic placement of the Microsporidia (supported by SSU rRNA, elongation factors one-alpha and two, and isoleucyl-tRNA synthetases) at or near the base of the eukaryotic lineage. It has been suggested that the Microsporidia may be highly divergent fungi. Evidence in support of this hypothesis includes: the position of an insertion in the elongation factor 1alpha protein unique to Microspora, animals and fungi (Kamaishi *et al.,* 1996); the presence of dihydrofolate reductase and thymidylate synthase as two separate enzymes in Microspora, animals and fungi but fused in plants and other protists (*Vivarès et al.,* 1996); parallels between the unusual meiotic cycle of Microspora and certain fungi (Flegel and Pasharawipas, 1995); and more specifically, the phylogenetic placement of microsporidial alpha- and beta-tubulins within the fungal radiation (Keeling and Doolittle, 1996; Edlin *et al.,* 1996; Keeling, 1998; Philippe and Adoutte, 1998).

Rather than resolve the divergence order of early eukaryotes, the aforementioned research has raised even more questions about the evolution of eukaryotes and the genes therein. Clearly, these incongruent trees demonstrate that existing models do not adequately explain evolution, particularly when distantly related sequences are considered. More specifically, the models do not cope well with the level of sequence variation observed within the genes studied (Sogin and Silberman, 1998) nor the apparent level of mutational saturation found in some genes of various lineages (Philippe and Adoutte, 1998).

Sequence variation and its affects of phylogenetic reconstruction have been seen for the alpha-tublin, beta tubulin and rRNA genes. For example, Keeling and Doolittle (1996) acknowledge the alpha- and beta-tubulin genes of the Microsporidia and fungi are highly divergent. Therefore, the placement of the Microsporidia within the fungal clade may be the consequence of long branch attraction. Also, the nucleotide sequence of the SSU rRNA of the protists contains more sequence variation than observed within the archaebacteria or eubacteria (Sogin and Silberman, 1998). More specifically, the SSU rRNA of the Microsporidia exhibits a high degree of divergence, a bias in nucleotide composition, an accelerated nucleotide substitution rate, and heterogeneity of the nucleotide substitution rate (Vossbrinck *et, al.,* 1987; Kamaishi *et al.,* 1996b; Keeling and Doolittle, 1996; Edlin *et al.,* 1996; Germot *et al.,* 1997; Hirt *et al.,* 1997). Hence, it is not surprising that recent findings now question the validity of using the SSU rRNAs for inter-lineage phylogenetic reconstructions where eukaryotes are included.

Tourasse and Gouy (1997) have demonstrated that the SSU rRNA accumulates more substitutions in eukaryotes than the LSU rRNA. Computed distances indicate a substitution rate of greater than 1.5 substitutions per site for the SSU rRNA whereas the substitution rate of the LSU rRNA has been computed at about 1 substitution per site. Therefore, in a typical analysis maximising the number of retained positions, the SSU rRNA is saturated with substitutions for the eukaryote/prokaryote comparison. The LSU rRNA however, because of its lower substitution rate, still remains a useful molecule for the inter-lineage comparison.

In this chapter I present: an analysis of the degree of divergence for the SSU and LSU rRNAs of *N. vespula*, reconstruction of phylogenetic trees from aligned SSU and LSU rRNA sequences, and phylogenetic inferences from these analyses.

## 7.2 Materials and methods

### 7.2.1 Sequence alignment

Sequence data obtained for the SSU and LSU rRNA genes of *N. vespula* (Figure 6.2) were used. These sequences were aligned to the pre-aligned sequences of the SSU and LSU rRNAs for two eubacteria, two archaebacteria and ten eukaryotes as described in section 2.8.

### 7.2.2 Analysis of sequence divergence

The aligned sequences were analysed for the presence of nucleotide divergence at positions normally conserved in eukaryotes. The aligned sequences in DCSE format were manually converted from a horizontal to a vertical alignment matrix and transferred to a spreadsheet (Microsoft Excel®, Microsoft Corporation). This alignment included the character-state of those positions in eukaryotes that are conserved in either 90-95% or >95% of species (Section 2.8). Also included in the alignment was the character-state of those positions in the eubacteria and archaebacteria that are conserved in >95% of species. Sequence comparisons were made to detect the presence of eubacterial or archaebacterial character-states at positions also conserved in eukaryotes but where the conserved character-state differs. Outcomes from comparisons were categorised and summed over the entire data set. The significance of nucleotide divergences at conserved positions for each aligned species was determined using the chi square test of significance with three degrees of freedom. The null hypothesis (non-significant divergence) was assumed. The chi square test is a function available within the

Microsoft Excel® software package.

### 7.2.3 Phylogenetic Analysis

Sequence alignments in DCSE format were manually converted to Seaview format and the unambiguous alignment positions marked as previously described (Section 2.9). Unambiguous alignment positions were chosen using secondary structure models of the eukaryotic SSU and LSU rRNAs (Section 2.9). Phylogenetic analysis of selected aligned positions was achieved using the program PHYLO_WIN (Section 2.9).

Evolutionary relationships among 15, eukaryotic, archaebacterial, and eubacterial SSU and LSU rRNA sequences were reconstructed using maximum parsimony and the neighbor-joining method (Saitou and Nei, 1987) from distance estimates (Galtier and Gouy, 1995). These methods were chosen for consistency and comparative purposes with existing publications. Also, while it is desirable to include maximum likelihood analysis, available computer resources preclude the use of this method because of its computational complexity.

Maximum parsimony theorem maintains the simpler hypothesis is preferable to the more complex, and that *ad hoc* hypotheses should be avoided whenever possible. Tree estimation by whatever method under the principles of parsimony seeks to explain shared attributes among taxa as being inherited from a common ancestor. Conflicts in shared characters require the invocation of an *ad hoc* hypothesis to explain the observed data including assumptions of homoplasy (convergence, parallelism, or reversal). Application of parsimonious methods to nucleotide sequences is an attempt to construct evolutionary trees that require the least number of evolutionary steps (transformation from one character state to another) to explain the observed data (Farris 1983; Felsenstein 1982, 1988, Swafford *et al.,* 1996).

Distance methods measure pairwise differences between species. This data is used to compute trees of possible evolutionary relationships. Several distance methods recover the true evolutionary tree when the distance used is tree-like; the distance between any two sequence pair equals the shortest path on the tree connecting the pair (Saitou and Nei, 1987). For nucleotide data, the substitution rate per site between two DNA sequences is such a tree-like distance. The correct estimation of this value is sufficient for reconstructing the true evolutionary tree. Estimation of substitution rates per site between two DNA sequences requires assumptions to be made about the evolutionary process in each lineage. These assumptions about the evolutionary process usually include those of sequence homogeneity and stationarity. Homogeneity is the constancy of the evolutionary process with time and among the lineages, while stationarity is the constancy of base composition within each lineage. Violation of these assumptions in the actual data leads to inaccurate distance estimates. Both assumptions imply equality of nucleotide frequencies in all extant sequences (Galtier and Gouy, 1995). This however is not consistent with the observed data, for example, the bias nucleotide composition of some eukaryotes such as the diplomonads and Microsporidia. Comparisons of sequences with unequal nucleotide composition may lead to distance estimates that themselves may become biased. It has been demonstrated that violations of the assumptions of homogeneity and stationarity can substantially reduce the ability of tree-making methods to recover the true phylogenetic tree (Galtier and Gouy, 1995). The Galtier and Gouy distance measure estimates the evolutionary distance between sequence pairs without the assumptions of homogeneity and stationarity of the evolutionary process.

Statistical significance of the reconstructed trees was calculated using bootstrap analysis (Felsenstein, 1985). Three hundred replicates were used for maximum parsimony analysis and 1,000 bootstraps for distance analysis. The level of bootstrap analysis was subject to the constraints of computer resources but deemed to be sufficient for these analyses.

## 7.3 RESULTS AND DISCUSSION

### 7.3.1 Sequence alignments

The mature rRNA subunits are a mosaic of conserved domains interspersed by more rapidly evolving regions (Michot *et al.*, 1990). Secondary structure models of homologous subunits from different taxa reveal structural homology not apparent in the sequence alignments; evolutionary conservation has constrained the secondary structure more so then the gene sequence. A common core is maintained through compensatory changes in helical elements (Michot *et al.*, 1984; Raué *et al.*, 1988; Raué *et al.*, 1990) which can be detected by comparative sequence analysis (Fox and Woese, 1975). In total, approximately one-third of the total *E. coli* rRNA secondary structure is globally conserved much of which can be attributed to paired nucleotide replacement (Gray *et al.*, 1984; Cedergren *et al.*, 1988).

Accurate alignment of sequences is the basis of phylogenetic inference from nucleotide data. The underlying conservation of the rRNA secondary structure becomes a powerful tool when used to aid in the accurate alignment of sequences for phylogenetic inference. The importance of secondary-structure-based alignments cannot be underestimated particularly when the alignment process involves sequences from highly divergent genes such as those of the Microsporidia. The following example demonstrates how the use of secondary models in the alignment procedure highlighted the absence of a highly conserved secondary structural element in *N. vespula*.

The Microsporidia have undergone degenerate evolution of their SSU and LSU rRNAs. This is evident by the loss of conserved structural elements (Chapters 5 and 6). Figure 7.1 demonstrates one of the most striking example of degenerate evolution in the Microsporidia when the eubacterial, archaebacterial, and eukaryotic lineages are considered. Helix 24 (Figures 5.3a and 6.4a) of the LSU is conserved in the three primary lineages, exhibiting little variation in nucleotide number among the lineages (eubacteria - mean nucleotide number 36.2 standard deviation 0.8; archaebacteria 35.9, 0.9; eukaryotes 35.8, 0.6; see Table 5.2). Visual inspection of the alignment (Figure 7.1) demonstrates the presence of nucleotide conservation in both the prokaryotes (*S. acidcaldarius, H. maritima, T. maritima, E. coli)* and eukaryotes (excluding *N. vespula*). The conserved eukaryotic nucleotide positions are indicated in the 'conserved' alignment as either an 'M' (>95% conservation) or 'Q' (90-95%). The two nucleotides shown in the *N. vespula* alignment are associated with the secondary structure elements adjoining helix 24. Helix 24 in *N. vespula* is clearly absent.

Helix 24 is highly conserved in the three primary lineages. Alignment of this region would have been problematic if the process of sequence alignment had not been based on secondary structure models. Without the secondary models to guide the alignment process, the most likely assumption is that the region is highly conserved (based on the presence of conserved nucleotides) and so must be present. Therefore, in attempting to align this region, nucleotides would be drawn from adjacent regions disrupting their true alignment. Overall, this would lead to a decrease in alignment accuracy. Because of such inaccuracies, sequence comparisons of misaligned conserved regions might lead to overstated conclusions about the degree of sequence divergence, incorrect assumptions about substitution rates, misleading phylogenetic reconstructions, and increased branch lengths.

### 7.3.2 Phylogenetic inheritance and sequence divergence

The most obvious and direct evidence of phylogenetic inheritance is the possession of ancestral nucleotides in alignments of homologous genes. Divergence from the ancestral state arises through nucleotide substitutions. These substitutions themselves may become fixed such that they can be traced in the descendants or can in turn become substituted.

The SSU and LSU rRNA alignment data bases now contain sufficient sequences from a range of species to calculate the conservation level of each nucleotide at each position within the structural core of these subunits. These alignments reveal that all three primary domains share invariant nucleotides. In addition to these invariant nucleotides are conserved nucleotides located at specific alignment positions. When the three primary lineages are compared, the character-state of these conserved nucleotide positions varies among the lineages: character-state is conserved in all lineages; character-state is conserved in all lineages but differs in one lineage from the remaining two; and character-state is conserved in all lineages but differs in each lineage. Additionally, at some alignment positions nucleotide conservation occurs in only two lineages. At these positions the character-state may be shared by both lineages or be unique to each lineage. Finally, each lineage possesses unique conserved positions. All of these types of positional conservation occur interspersed throughout the rRNA genes of the three primary lineages.

Those positions that are invariant in the rRNA subunits of the three domains confirm the pre-existence of a common ancestor to today's extant cellular life-forms. Some highly conserved positions confirm phylogenetic relationships such as the possible sisterhood of the archaebacteria and the eukaryotes, while others confirm membership to a specific lineage. The more variable regions are useful in determining close taxonomic relationships such as membership of a genus. The hypervariable regions may be useful in determining strains or isolates of a single species. These discernible taxonomic relationships have arisen over time and because of nucleotide substitution events.

The probability of nucleotide substitution varies across the rDNA operon. While many of these events are informative, many are uninformative or potentially misleading. Positions that are misleading (homoplasies) are a prime source of phylogenetic noise and not always themselves easily identifiable. A further complication occurs at positions subject to multiple substitutions. In the extreme, multiple substitutions may become widespread within the gene causing mutational saturation. High substitution rates, and more specifically mutational saturation, causes the underestimation of divergence which separates pairs of sequences. Therefore, corresponding species on the tree will appear closer to each other than they are in reality (Philippe and Adoutte,1998). To correctly interpret phylogenetic trees, the alignment of gene sequences must be closely examined to identify informative or misleading positions.

Two categories of eukaryotic character-state conservation are of interest with respect to this thesis; those positions in this lineage where the character-state is conserved in either >95% or 90-95% of species. Also of interest are alignment positions of the eubacteria and archaebacteria that are conserved in 95% of species. The eukaryotic SSU rRNA model (Gutell, 1994a; Appendix 1) identifies 553 positions in the >95% category and 129 positions in the 90-95% category, while in the LSU rRNA model (Gutell *et al.,* 1993; Appendix 2a, b) there are respectively 1003 and 244 positions. Tables 7.1a,b and 7.2a,b summarise a series of comparisons made between the conserved eukaryotic, eubacterial, and archaebacterial positions and the corresponding positions of two eubacteria, two archaebacteria, and eleven eukaryotes. Each table is divided into sections which compared different character-states in a range of organisms.

*Section 1*

This section of each table shows the number positions in each species where the conserved eukaryotic character-state is absent, retained, or of alternate character-state'. 'Nucleotides absent' is the total number of positions in the aligned species where an alignment gap occurs. 'Positions retained' is the total number of positions at which the aligned species and the conserved eukaryotic model share the same character-state. Positions 'alternate character-state' is the total number of positions at which the aligned species and the conserved eukaryotic model differ in character-state.

This comparison reveals that the character-state for many of the conserved eukaryotic alignment positions have remained fixed since before the divergence of the eukaryotic domain. This is evident by the large number of retained eukaryotic character-states present in the prokaryotic taxa. This observation is not unexpected and concurs with previous findings relating to the universal conservation of secondary structure and the phylogenetic relationship of cellular life-forms (Gray *et al.,* 1984; Michot *et al.,* 1984; Cedergren *et al.,* 1988; Raué *et al.,* 1988; Raué *et al.,* 1990).

A further observation is that a large number of normally conserved eukaryotic nucleotides are absent in the SSU (19) and LSU (50) rRNAs of *N. vespula*. Furthermore, excluding the SSU rRNA >95% category, the number of absent nucleotides in *N. vespula* is larger than that observed for the prokaryotic taxa. This indicates that even normally invariant nucleotides have been lost from the SSU and LSU rRNAs of *N. vespula*, further supporting other observations of degenerate evolution in the Microsporidia.

Along with a number of absent nucleotides, *N. vespula* also has the greatest number of positions (>95% category ) in both subunits of the eukaryotes with an 'alternate character-state'. The normally high level of conservation at these positions in the eukaryotic taxa indicates that substitutions occur rarely at these positions. Two possible explanations as to why *N. vespula* has such a large number of 'alternate character-states' are: 1) *N. vespula* diverged before these positions were substituted and became fixed in other eukaryotes; and 2) the substitution rate at these positions in *N. vespula* is extremely high relative to other eukaryotic taxa.

Other species also have positions with 'alternate character-states'. Ranked after *N. vespula* for the SSU rRNA (>95% category) are *P. polycephalum* and *E. histolytica.* Likewise for the LSU rRNA, *N. vespula* is followed in ranking by *E. gracilis, E. histolytica,* and *P. polycephalum.* For both subunits, *G. muris* is ranked behind these species. Assuming that there is a link between the number of 'alternate character-states' and the period since divergence, this result appears to contradict the usual rRNA phylogenetic placement of the Diplomonadida at the base of the eukaryotic lineage. As yet to be discussed, a number of these 'alternate character-states' are shared by distantly related taxa and the prokaryotes suggesting that at least some of these are the ancestral character-state. The remainder are likely to be substitution events.

In contrast, in the 90-95% category for both subunits, *N. vespula* (SSU 52, LSU 80) and *G. muris* (SSU rRNA 52, LSU rRNA 82) share about equal numbers of positions with 'alternate character-states'. In this category, both organisms exhibit more 'alternate character-state' positions than the other eukaryotic species and almost as many 'alternate character-state' positions as the prokaryotic species. Once again, if you assume that an early divergence is characterised by the number of positions with 'alternate character-states' this observation supports published rRNA phylogenetic trees in which the Microsporidia and diplomonads diverge at the base of the tree.

These results raise the question: why is it that *G. muris* differs in relative number of 'alternate character-state' nucleotides between the two degrees of conservation? To explore this observation it is necessary to study the distribution of character-states of conserved eukaryotic positions.

### *Section Two*

The second section of each table examines in more detail the distribution of character-states at conserved eukaryotic positions excluding those positions defined as 'nucleotides absent'. Here, three aspects of the data are considered: 1) the total number of positions that retained the conserved character-state; 2) the total number of each character-state at those positions defined as 'alternate character-state'; and 3) the overall character-state preference.

The first notable observation from this analysis is the apparent A + G bias of the conserved positions in the eukaryotic SSU and LSU rRNA models. The nucleotide character-states A and G combined represent 59% of the SSU rRNA and 62% of the LSU rRNA conserved (≥90%) positions in the eukaryotic models. However, the distribution of character-states in some species does not fit this pattern and is biased toward other character-states.

Each species retains the majority of the conserved eukaryotic character-states. The bracketed value accompanying the conserved character-state value indicates the number of times that the character-state occurs at other conserved positions. The bracket values at these positions are 'alternate character-states'. The sum of these two figures for each character-state when compared to the expected number for each character-state (column 1) indicates if character-state bias is present.

Diplomonads and Microsporidia are well known for their G + C bias across the rRNA subunits. This observed pattern of bias is based on all positions within these subunits. By limiting the choice of positions assessed for nucleotide bias to only highly conserved positions (>95% and 90 - 95%) a different picture emerges. Within the higher category of nucleotide conservation, only *N. vespula* exhibits a low G + C bias. For example, in table 7.2a there are 208 C and 323 G positions within the eukaryotic LSU rRNA that are highly conserved (>95%). *N. vespula* has 47 C (expected - observed) and 58 G positions that have 'alternate character-states', these being predominantly either A or T. Consequently, the total number of A and T nucleotides at highly conserved LSU rRNA positions in *N. vespula* exceeds the total number of conserved As (289) and Ts (183) in the eukaryotic LSU rRNA model. *N. vespula* has an additional 43 As and 55 Ts. In comparison, of the 289 A and 183 T nucleotides in the eukaryotic LSU rRNA model, only 12 A and 4 T positions have an 'alternate character-state' in *G. muris*. The total number of C nucleotides in *G. muris* remains constant at 208, while the total number of G nucleotides increases by 1 to 324. Clearly, in the >95% category, *N. vespula* is character-state biased while *G. muris* is character-state neutral. Some of the other protist species in this analysis (eg., *P. polycephalum, E. histolytica,* and *E. gracilis*) exhibit results similar to that of *G. muris*. These species possess a number of positions with 'alternate character-states'. However, when the total number of each character-state is considered, the degree of bias in these species could best be described as slight and not necessarily as G + C but other combinations such as C + G + T (eg.

table 7.2a for *E. gracilis*).

The effects of high G + C bias in the *G. muris* LSU rRNA is only apparent as the level of nucleotide conservation declines from >95% to 90-95%. At the lower level of conservation, the G + C bias of *G. muris* becomes almost as pronounced as the A + T bias observed for *N. vespula*. For example, character-state bias for these organisms is apparent in table 7.2b. At conserved (90 - 95%) position in the LSU rRNA there are 77 A, 38 T, 70 G, and 59 T nucleotides. At these positions, each character-state is under-represented in both *N. vespula* and *G. muris* (eg, *N. vespula* observed A nucleotides 77, retained A nucleotides 54; *G. muris* observed G nucleotides 70, retained G nucleotidess 53). However, overall, the total number of A and T nucleotides for *N. vespula* (161) and G and C nucleotides (137) exceeds the total number of conserved A and T nucleotides (136) and G and C nucleotides (108). Of the other eukaryotic species analysed, only *E. histolytica* exhibits a degree of low G + C bias in both subunits.

Evidently, the G + C bias to some extent has influenced the character-state at many conserved positions in the SSU and LSU of *N. vespula* and *G. muris*. This influence of character-state bias has contributed to the number of alignment positions with an 'alternate character-state'.

To detect potential ancestral links between the prokaryotes and eukaryotic taxa it is necessary to undertake detailed comparative studies of aligned sequences. The highly conserved core of the SSU and LSU makes it possible to align both prokaryotic and eukaryotic rRNA sequences. Analysis of such alignments, and in particular those positions defined as 'alternate character-state', potentially reveal the presence of ancestral character-states supporting direct ancestral links.

*Section Three*

In terms of phylogenetic inference, positions with 'alternate character-states' are of particular interest. It is these positions that link the prokaryotes and eukaryotes by possession of ancestral character-state other than those shared by all three lineages. Section three of each table examines the character-state relationships between conserved prokaryotic positions and those positions in the aligned species deemed to have 'alternate character-states'. There are seven possible outcomes. 'Other' are alignment positions that are conserved in only the eukaryotic model where the character-state differs in the organism compared. 'Unique' are alignment positions at which the character-state is conserved across all three lineages but differs in the compared organism. 'Prokaryote domain' are alignment positions conserved in the three lineages where the character-state is shared in the eubacteria and archaebacteria models but differs from that of the eukaryotic model. 'Eubacteria domain' and 'archaebacteria domain' are alignment positions conserved in all lineages but with a different character-state in each lineage. The organism compared at these alignment positions respectively possess the eubacterial or archaebacterial character-state. 'Eubacteria only' and 'archaebacteria only' are alignment positions that are conserved in the eukaryotic model and either the eubacterial or archaebacteria models only. The character-state of the eubacteria or archaebacteria nucleotide differs from the eukaryotic model. The data was categorised in this way to identify nucleotides with an ancestral character-state. Also, to potentially reveal any obvious patterns of inheritance, for example, a greater portion of inherited archaebacterial than eubacterial nucleotides in the eukaryotic species.

The majority of 'alternate character-state' positions fall into the category of 'other' and therefore are uninformative. These nucleotides in the protists may represent character-states present before their divergence that subsequently were substituted and have since become fixed in the higher eukaryotes. Alternatively, they may simply be divergent nucleotides that have been subjected to one or more substitution events. The remaining 'alternate character-state' nucleotides are positions with either a 'unique' or an ancestral character-state.

Within the >95% category, *N. vespula* has the greatest number of ancestral nucleotides. There are five positions in the SSU rRNA and ten positions in the LSU rRNA. Ranked behind *N. vespula* is P. *polycephalum* (3 and 4), *E. histolytica* (1 and 2), and *T. thermophila* (1 and 2). Neither *G. muris* nor any of the other eukaryotic species possess nucleotides that have the ancestral character-state at the >95% level of conservation.

For the 90-95% category, the situation is reversed. *G. muris* has more alignment positions with the ancestral character-state than any other of the included eukaryotic species. *G. muris* has 16 positions in the SSU and 9 positions in the LSU with the ancestral character-state, while *N. vespula* has 8 and 6, P. *polycephalum* has 4 and 4, *E. histolytica* has 3 and 0, and *T. thermophila* 1 and 0. The shift in ranking observed for *G. muris* from the least to the most number of prokaryotic character-states parallels the shift in number of 'alternate character-state' nucleotides between the two conservation categories, >95% 45 and 90-95% 134. Certainly in the 90-95% category *G. muris* rivals *N. vespula* which has 132 positions with 'alternate character-states'. These alignments also show that in the >95% category for both subunits, *T. thermophila* has fewer 'alternate character-state' positions than *G. muris* (SSU 6 verses 17, LSU 13 Vs. 27) yet more with an ancestral character-state (SSU 1 Vs. 0, LSU 2 Vs.

0).

The most obvious explanation for these protists having more ancestral nucleotides than *G. muris* in the >95% category is that these are reversions to the ancestral character-state. Alternatively, this may represent phylogenetic inheritance at these positions that were lost in *G. muris*. Or simply, that the divergence of these species pre-dates *G. muris*. Whatever selective pressure caused this phenomenon, it demonstrates that at some alignment positions there are selective differences for character-states between the lower and higher eukaryotes. Also, this analysis demonstrates that many of the highly conserved positions in eukaryotes are specific only to the higher eukaryotes. Sogin and Silberman (1998) observed that the degree of sequence variation (nucleotide divergence) in the protist SSU is greater than that of the eubacteria or archaebacteria. The results presented here support their findings.

In terms of nucleotide character-state inheritance by the eukaryotic species, there appears to be a weak preference towards 'prokaryote domain' inheritance. Several species have ancestral character-states that occur in both the eubacteria and archaebacteria but not in the 'eubacteria only' or 'archaebacteria only'. The pattern (albeit weak) of 'prokaryotic domain' character-state inheritance was not unexpected. It was anticipated that the order of character-state preference inheritance would be 'prokaryotic domain' then 'archaebacterial domain', 'archaebacteria only', 'eubacteria only', and finally 'eubacteria domain'. This pattern of nucleotide inheritance would agree with the sister-lineage relationship between the archaebacteria and the eukaryotes. However, there was a surprising number of 'eubacterial only' nucleotides. Although, it should be noted that while two categories of eukaryotic conservation are considered, only positions in the >95% category are considered for the eubacteria and archaebacteria. Hence, a position scored as 'eubacteria domain' or 'eubacteria only' may also be conserved in the archaebacteria but at a level below 95%. This method of scoring only highlights specific character-state preferences at highly conserved positions.

The rRNA trees predict that the diplomonads are the earliest divergent eukaryote. Therefore, it would be expected that if the divergence of the Diplomonadida predate all other eukaryotic taxa, then in any eukaryotic alignment position, if the ancestral character-state is not also shared by *G. muris* that the ancestral character-state must be due to homoplasies. However, this statement ignores two possibilities: 1) that in *G. muris* either these positions are variable; or 2) they have become fixed under the influence of G + C bias while remaining conserved in the other species. Two scenarios support the inheritance of ancestral nucleotides that do not occur in *G. muris:*1) the putative ancestral nucleotides required substitution events to occur against the direction of G + C bias; or 2) two unrelated species share the same ancestral character-state. In the second case, the most parsimonious explanation is that the loss of the prokaryotic character-state in one species (here *G. muris*) is more likely than the reversion to the ancestral character-state in two unrelated species.

Those alignment positions in the eukaryotic species identified as ancestral are displayed in tables 7.3 and 7.4. These positions are displayed as an alignment of character-states respectively for the eubacteria, archaebacteria, eukaryotes, and eukaryotic species. Included in these tables is data about the type of alignment (eg. 'prokaryotic domain'), any other eukaryotic species that shares the same prokaryotic character-state, and the position relative to the *N. vespula* SSU and LSU models (Figures 6.3, 6.4a,b).

Within the two categories of eukaryotic nucleotide conservation for all eukaryotic species assessed, there are 46 SSU and 46 LSU rRNA positions that possess ancestral character-states. The 46 SSU rRNA positions consist of 9 >95% and 37 90-95% category alignment positions, while the 46 LSU rRNA positions consist of 19 >95% and 27 90-95% category alignment positions. Twenty-nine of the 46 SSU rRNA positions occur in either the *N. vespula* or *G. muris* alignments. Thirteen of the 29 positions occur in *N. vespula*. Similarly, there are 46 LSU rRNA positions of which 25 occur in either *N. vespula* (16) or *G. muris* (9).

For the >95% category of the *N. vespula* SSU rRNA, 5 of the 6 ancestral nucleotides may have arisen from biased substitutions such as G to A. However, for the LSU rRNA of *N. vespula* only 3 of 10 ancestral nucleotides could have arisen from biased substitutions, the remaining eight positions would be substitutions against the direction of bias, such as A to G substitutions.

In the 90-95% category both *N. vespula* and *G. muris* have ancestral nucleotides. Within this category for *N. vespula*, 5 of the 8 SSU rRNA and 5 of the 6 LSU rRNA ancestral nucleotides would have to had arisen as substitution against the direction of bias. In comparison, 4 of 9 SSU rRNA and 1 of 16 LSU rRNA positions in *G. muris* would have arisen from substitution against the direction of bias. Because of the overall A + T substitution bias in *N. vespula*, these results present a strong argument for the *N. vespula* nucleotides being ancestral. As the archaebacteria and the diplomonads are both G + C rich, there is no way to tell from these results how many of the G or C nucleotides in *G. muris* are ancestral or reversions. In comparison to *N. vespula*, *G. muris* possess few nucleotides that could be substitutions against the direction of nucleotide bias.

In addition to nucleotides requiring substitutions against the direction of bias to have occurred, several positions in the alignment indicate ancestral nucleotides shared by unrelated taxa. There are 14 ancestral nucleotides in *N. vespula* and 9 in *G. muris* that also occur in the other aligned species. Of these 23 nucleotides, 15 occur in the SSU rRNA. *Nosema vespula* shares ancestral nucleotides with *G. muris, E. histolytica, P. polycephalum, E. gracilis*, and *T. thermophila* (SSU rRNA only). *G. muris* shares ancestral nucleotides with *E. histolytica, P. polycephalum,* and *E. gracilis*. Of the remaining eukaryotic species, *P. polycephalum* and *E. histolytica* share one archaebacterial nucleotide in the SSU rRNA, while *E. gracilis* and *P. polycephalum,* and *E. gracilis* and *T. brucei* share one prokaryotic nucleotide each in the LSU rRNA. Overall, *N. vespula* has the largest number of shared ancestral nucleotides followed by *G. muris.* This result supports a divergence of Microsporidia that pre-dates the diplomonads as shown in some of the distance based trees such as elongation 1 alpha.

Three other aspects of the data are considered in tables 7.3 and 7.4. First, the structural element of the rRNA subunit in which the ancestral nucleotide occurs. Second, when the location of the nucleotide with the ancestral character-state occurs in a helix, whether the paired nucleotide is conserved, and if so, is the character-state of the nucleotide eukaryotic or prokaryotic. Third, the position of the ancestral nucleotides with respect to the *N. vespula* rRNA secondary models. The third aspect of the data is intended to reveal any positional preference for the inheritance of ancestral nucleotides.

An analysis of secondary structure motifs of the included species reveals that within the SSU, 32 helix, 8 loop, 5 hairpin and 1 bulged nucleotide positions have ancestral nucleotides. Similarly, for the LSU there are 21 helix, 16 loop, 6 hairpin, and 3 bulge positions that have ancestral nucleotides (tables 7.3 and 7.4). Twenty-nine of these nucleotides occur in *N. vespula* while there are 25 such nucleotides in *G. muris*. According to the eukaryotic SSU and LSU models (Gutell, 1994a; Gutell *et al.,* 1993), 24 of the SSU (Table 7.3 column 4) and 15 LSU (Table 7.3b columns 4) helical positions are also conserved. Of these 24 SSU nucleotide pairs, three remain structurally unchanged such that a conserved C:G pair becomes U:G (*N. vespula*) and 2 A:U pair become G:U pairs (*P. polycephalum, T. brucei*). Similarly, of the 15 LSU nucleotide pairs, 5 remain structurally unchanged such that 3 G:U pairs are replaced by more stable G:C pairs and 2 G:C pairs are replaced by less stable G:U pairs. These occur in *N. vespula, E. gracilis*, and *P. polycephalum.* Also, in *E. gracilis* there is one substitution that results in a G:C bonded being replaced by a C<>C pair. This pair would disrupt the helix at this point. For the majority of these paired nucleotides to become a 'eukaryotic' nucleotide pair, both nucleotides have to be substituted and for this reason they are more likely to be ancestral than reversions.

Visual inspection of the involved SSU positions reveals no apparent distribution pattern. However, within the LSU the distributions seems to favour some specific regions, the most obvious of which is the distal half of helix 42, helix 43, and helix 44 (Figure 6.4a). This region is known as the "GTPase center", is the binding site of the *E. coli* EL11 and *S cerevisiae* L15 ribosomal-proteins, and is an example of structural conservation across the prokaryotic/eukaryotic evolutionary barrier (Raué *et al.,* 1990). There are three alignment positions in the GTPase centre of *P. polycephalum* that have the 'prokaryotic domain' character-state. In contrast, there are two positions in *N. vespula,* one in *G. muris,* and one in *E. histolytica* that are present and of the 'eubacterial only' character-state. This observation agrees with the beta-tublin trees (Philippe and Adoutte, 1998) that suggest a divergence of *P. polycephalum* that predates both the diplomonads and the Microsporidians.

*Section four*

In the final section of tables 7.1a,b and 7.2a,b, the degree of divergence from the conserved positions of the eukaryotic SSU and LSU rRNA models is statistically tested for each species. To measure the significance of divergence for each species the chi square statistic (three degrees of freedom) is computed. The statistic was computed using the total number of each conserved eukaryotic character-state and the actual total number of character-states retained for each species.

Tables 7.1a and 7.2a (Section 4) demonstrate that at the highly conserved (>95%) eukaryotic positions within the SSU and LSU rRNA, *N. vespula* is as significantly (chi square $p < 0.05$) divergent as the included prokaryotic species. In the lower category of conservation, the conserved positions of the SSU rRNA of the prokaryotic species, *N. vespula* and *G. muris* are significantly divergent. For the LSU rRNA only *N. vespula* is significantly divergent in the higher category. In the lower category, in addition to *N. vespula, G. muris, P. polycephalum, E. histolytica, T. brucei, and E gracilis* are also significantly divergent. The increase in the number of species between the two categories with significant chi square values most likely parallels an increase in substitutions rates between the two categories. Also, it appears that the substitution rate of the LSU rRNA is greater than the SSU rRNA. Tourasse and Gouy (1997) suggest that in general the substitution rate of the SSU rRNA is greater than the LSU rRNA. Their conclusion was based on the analyses 102 species for which both SSU and LSU rRNA alignments were available. In this analysis only a restricted subset of mostly early divergent species were included and hence a result apparently in conflict with Tourasse's and Gouy's.

### 7.3.3 Phylogenetic reconstruction

A set of 1,100 SSU and 2078 LSU aligned and gap-free rRNA sequence positions were used for each of the 15 taxa examined. These positions were determined from sequence alignments based on secondary structure models. Only those positions considered to be hypervariable were excluded (Gutell, 1994a; Gutell *et al.,* 1993). Statistical significance was determined using 300 bootstraps (bs) for maximum parsimony (MP) and 1000 bs for neighbor-joining/Galtier and Gouy distance method (NJ). Trees were rooted by the inclusion of two eubacterial sequences (*E. coli, T. maritima*).

With respect to the divergence of the Microsporidia and the diplomonads, the MP and NJ trees (Figures 7.2a,b, 7.3b) agree with earlier published trees. The SSU rRNA MP tree places *G. muris* as the earliest divergent eukaryote followed by *N. vespula* (also: Sogin et. al., 1989; Cavalier-Smith, 1993; Van De Peer *et al.,* 1993; Van Keulen *et al.*, 1993) while in the SSU rRNA and LSU rRNA NJ trees, this branching order is reversed (also: Galtier and Gouy, 1995; Philippe and Adoutte, 1998). In all three instances, the branching order is well supported by their respective bootstrap values: MP for SSU rRNA, 100% for the divergence of eukaryotes with *G. muris* as the earliest branch then the divergence of *N. vespula* 86%; NJ for SSU rRNA and LSU rRNA, 100% for the divergence of eukaryotes with *N. vespula* as the earliest branch then the divergence of *G. muris* with respectively 98 and 90% support. These results are incongruent regarding the divergence of *N. vespula* and *G. muris* and so at least one of these trees must be incorrect.

An unexpected result was obtained for the LSU rRNA tree using MP (Figure 7.3a). *N. vespula* appears as a sister species to *E. histolytica* with both these species forming a monophyletic clade with *P. polycephalum*. However, by removing the *E. histolytica* sequence from the analysis (data not shown) *N. vespula* returns to a basal position branching immediately after *G. muris*. A possible explanation for this phenomenon is that these sequences are very A + T rich [*N. vespula* (66.7%), *E. histolytica* (66.7%), and *P. polycephalum* (62.6%)] and as such MP analysis artificially group these sequences together. By comparison, for the SSU rrNA alignment, only *N. vespula* (62.3%) and *E. histolytica* (67%) are A + T rich while *P. polycephalum* is more balanced with respect to nucleotide composition (47.9%). Equally, as *N. vespula* is attracted to *E. histolytica* and *P. polycephalum,* the basal position of *G. muris* may be due to G + C rich attraction between *G. muris* (SSU 59.1%, LSU 57.3%) and the archaebacteria

[*H. marismortui* (53.2%, 55.7%) and *S. acidocaldarius* (62.3% and 60.3%)].

The order of species divergence for the NJ trees of the SSU rRNA and LSU rRNA is more congruent than the MP trees. The apparent differences between the NJ trees relates to the divergence of *P. falciparum*, *T. thermophila* and *P. micans*. First, in the SSU rRNA tree these species diverge as independent, sister-lineages, whereas for the LSU rRNA tree they diverge as a monophyletic clade. Second, the divergence order of these species differs between the two trees.

The branching pattern of the archaebacteria, and *E. gracilis* and *T. brucei* in the LSU rRNA MP tree is also atypical compared to other similar trees. Rather than appearing as an archaebacterial and eukaryotic monophyletic clades, these species diverge as sister lineages; divergences well supported by their bs values. Such variations in tree topology, particular those which associate apparently distantly related species as sister-groups, have been found to be associated with strong composition bias, grouping sequences of similar base composition. As seen above, these topologies are often highly supported by high bootstrap values (Galtier and Gouy, 1995). The other phenomenon of long branch attraction causes species with higher rates of evolution to group with each other instead of associating with their true sister-taxa.

Of the published phylogenetic trees, including both the Microsporidia and the diplomonads, only the elongation factor 1 alpha tree is congruent with the SSU rRNA tree. In these trees both the diplomonads and the Microsporidia diverge at the base of the eukaryotic tree with the Microsporidia as the first branch (Kamaishi *et al.*, 1996a,b). Trees reconstructed from beta tubulin sequences place the divergence of the Microsporidia within the fungal taxa. These trees also suggest that the divergence of the Diplomonadida was preceded by the divergence of *E. histolytica*, some Mycetozoa (eg. *P. polycephalum*) and Trichomonadida (Edlin *et al.*, 1996; Keeling and Doolittle, 1996; Philippe and Adoutte, 1998).

Similarly, the actin tree demonstrates incongruent branching patterns with the rRNA and beta tublin trees. Microsporidian actin sequences are yet to be used in phylogenetic reconstructions; however the diplomonads diverge at the base of the tree followed by the Ciliophora and the trichomonads (Philippe and Adoutte, 1998). To account for such incongruent trees, Philippe and Adoutte (1998) propose a link between the frequency of use for the molecule explored and the position of divergence for the species on the tree. In species in which a molecule is little used, the species is seen to diverge further down the tree, and conversely when the molecule is used frequently, the species is seen to diverge further up the tree. Furthermore, they suggest that a decline in usage of the molecule allows the rate of evolution of the molecule to increase in some lineages. Because of accelerated evolution, the molecule becomes saturated with substitutions and may, in the extreme, result in two species branching as an unresolved multifurcation. Philippe and Adoutte further suggest that the high rate of saturation of the rRNAs indicates that the rRNA molecules are evolving at rates that are lineage dependent. Certainly, the data presented above ('Unique': tables 7.1 - 7.4) suggest the presence of differing rates of evolution in the lineages represented. Also, this data indicates that in some lineages, even the normally highly conserved nucleotide position can be subjected to elevated substitution rates. From the data presented above the rRNA genes of *G. muris, N. vespula, P. polycephalum,* and *E. histolytica* have undergone lineage-dependant accelerated evolution.

## 7.4 Summary

Phylogenies based on SSU rRNA, elongation factors one-alpha and two, place the amitochondrial protists, and more specifically the microsporidia, at the base of the eukaryotic evolutionary tree. In contradiction, beta-tublin and actin trees suggest that the amitochondrial protist diverged further up the eukaryotic tree and that the microsporidian diverged within the fungi. The cause of these incongruent branching patterns lies within the nucleotide and protein sequence alignments. It has been clearly demonstrated that the evolutionary process is lineage dependant causing some lineages to evolve more quickly than others. Central to the evolutionary process is nucleotide deletion, insertion, and substitution. As neither insertions nor deletions are usually considered in tree reconstruction from alignments, the incongruent branching patterns observed must therefore be attributable to substitution events. It has been demonstrated that the substitution process itself is highly variable. In some species only some regions of genes are affected by substitutions, while in extreme cases in other species there are many substitutions at many positions leading to mutational saturation. One aspect of the substitution process observed in the Microsporidia and the diplomonads are substitutions that result in G + C bias. Numerous evolutionary models have attempted to describe the process of nucleotide substitution, including substitution, bias and hence the controversy that now exits.

The analysis in this chapter further supports the work of others revealing the significant level of divergence through substitution that has occurred in the lower protists. However, as shown, underlying the divergence of the lower protists, there exists physical links through the possession of ancestral nucleotides at conserved eukaryotic positions. It is possible and likely that some of these ancestral nucleotides are homoplasies, but data presented herein suggest that the majority of positions identified are likely to be inherited. *N. vespula* and *G. muris* possess by far the greatest number of these ancestral nucleotides. Of these two species, *N. vespula* has the greatest number of ancestral nucleotides when both categories of conservation and both subunits are considered. In most cases for these character-states to have arisen through reversion, it would require substitutions against the direction of bias to have taken place. Also at many positions in helices, the paired nucleotide is conserved, thus requiring substitution at two positions.

| Organism & Lineage | Sequence of Helix 24 | Nucleotide Number |
|---|---|---|
| AraTha Eu | AT-GAAA-AGGACTTTGA-AAAGAGA-GTCAAAGAGTGCT | 36 |
| CryNeo Eu | AT-GAAA-AGCACTTTGG-AAAGAGA-GTTAAACAGTACG | 36 |
| PlaFal Eu | AT-GAAATAGTACTCAGG-AATGAGCAATTAAATAGTACC | 38 |
| ProMic Eu | GT-GAAA-AGGACTTTGA-AAAGAGA-GTT-AAAAGTGCC | 35 |
| TetTer Eu | AT-GAAA-AGAACTTTGA-AAAGAGG-GCT-AAAAG-ACT | 34 |
| EugGra Eu | AT-GCAA-AGAACTCCGC-GAAGAGG-GTT-AAAAGTCCC | 35 |
| TryBru Eu | TTTGAAA-AGTACTTTGG-AAAGAGA-GTGACATAGAACC | 37 |
| PhyPol Eu | CT-GAAA-AGCACCTCGT--TGAGGA-GTT-AAAAGAGCA | 34 |
| EntHis Eu | CT-TAAA-AGAACTTTGG-AAAAAGA-GTG-AAAAGAGCT | 35 |
| GiaMur Eu | GT-GAGA-AGGATGCCGACCCAGGCA-CGTCAAAAGACCC | 37 |
| NosVes Eu | T--------------------------------------G | 2 |
| SulAci Ab | CT-GAAA-AGAACCCCGGAAGGGGGA-GTGCCAAAGAGCC | 37 |
| HalMor Ab | CT-GCAA-AGTACCCTCAGAAGGGAG-GCGAAATAGAGCA | 37 |
| TheMar Eb | GT-GAAA-AGCACCCCG-GAAGGGGA-GTGAAAGAGGACC | 36 |
| EscCol Eb | GC-GAAA-AGAACCCCGGCGAGGGGA-GTGAAAAAGAACC | 37 |
| Conserved | -M-QQMM-MM-QM---Q-----M-Q--QQ--MM-MM—M- | 19 |

Figure 7.1

A comparison of aligned sequence for helix 24 of the LSU rRNA (Figure 6.4a) showing an example of degenerate evolution in *N. vespula*. Nucleotides position in the 'Conserved' sequence represented as either an 'M' or 'Q' have respectively, a level of conservation of either 95%+ or 90-95% in the eukaryotic model (Gutell *et al.*, 1993). The first and last nucleotides of the *N. vespula* sequence are respectively positions 398 and 399 in the *N. vespula* LSU model. The number of nucleotides within this helix for each organism is also displayed. Abbreviation are: Eu - Eukaryote; Ab - Archaebacteria; Eb - Eubacteria; AraTha - *Arabidopsis thaliana;* CryNeo - *Cryptococcus neoformans;* - PlaFal - *Plasmodium falciparum;* ProMic - *Prorocentrum micans;* TetTer - *Tetrahymena thermophila;* EugGra - *Euglena gracilis;* TryBru - *Trypanosoma brucei;* PhyPol - *Physarum polycephalum;* EntHis - Entamoeba histolytica; GiaMur - *Giardia muris;* NosVes - *Nosema vespula*; SulAci - *Sulfolobus acidcaldarius;* HalMor - *Halobacterium maritima;* TheMar - *Thermotoga maritima;* EscCol - *Escherichia coli.*

Tables 7.1a, b and 7.2a, b

An analysis of character-state comparisons between the conserved eukaryotic SSU and LSU rRNA nucleotides and the equivalent aligned nucleotide position of two eubacteria, two archaebacteria, and 11 eukaryotes. Tables 7.1 and 7.2 respectively compared conserved SSU rRNA eukaryotic nucleotides that are conserved in either >95% or 90-95% of all known eukaryotic SSU rRNA sequences. Abbreviation are:

Heading

Euk - Eukaryotes; Ec - *Escherichia coli;* Tm - *Thermotoga maritima;* Hm - *Halobacterium maritima*; Sa - *Sulfolobus acidcaldarius;* Nv - *Nosema vespula*; Gm - *Giardia muris;* Pp - *Physarum polycephalum;* Eh - Entamoeba histolytica; Tb - *Trypanosoma brucei;* Eg - *Euglena gracilis;* Tt - *Tetrahymena thermophila;* Pf - *Plasmodium falciparum;* Pm - *Prorocentrum micans;* Cn - *Cryptococcus neoformans;* At - *Arabidopsis thaliana.*

Section 1

Nucleotides Absent - the total number of positions in the aligned species where an alignment gaps occur adjacent to a conserved eukaryotic nucleotide; Positions Retained - the total number of positions at which the aligned species and the conserved eukaryotic model share the same character-state; 'alternate character-state' - the total number of positions at which the aligned species and the conserved eukaryotic model differ in character-state.

Section 2

'A', 'C', 'G', 'U' - the character-states of the conserved rRNA subunit eukaryotic nucleotide.

Section 3

Other - alignment positions that are conserved in only the eukaryotic model where the character-state differs in the organism compared; Unique - alignment positions at which the character-state is conserved across all three lineages but differs in the compared organism; Prokaryote - alignment positions conserved in the three lineages where the character-state is shared in the eubacteria and archaebacteria models but differs from that of the eukaryotic model; Eubacteria and archaebacteria - alignment positions conserved in all lineages but with a different character-state in each lineage; Eubacteria only and archaebacteria only - alignment positions that are conserved in the eukaryotic model and either the eu- or archaebacteria only.

Section 4

Chi square - the chi square statistic from data in section 2 using the expected values for each conserved eukaryotic character-state and the comparative species values.

Table 7.1a

| Eukaryotic 16S<br>Conserved >95% | *Escherichia coli* | *Thermotoga maritima* | *Halobacterium marismortui* | *Sulfolobus acidocaldarius* | *Nosema vespula* | *Giardia muris* | *Physarum polycephalum* | *Entamoeba histolytica* | *Trypanosoma brucei* | *Euglena gracilis* | *Tetrahymena thermophila* | *Plasmodium falciparum* | *Prorocentrum micans* | *Cryptococcus neoformans* | *Arabidopsis thaliana* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Section 1** | | | | | | | | | | | | | | | |
| Nucleotides Absent | 13 | 10 | 13 | 10 | 10 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Positions Retained | 416 | 407 | 429 | 449 | 468 | 532 | 525 | 534 | 553 | 541 | 547 | 546 | 553 | 553 | 553 |
| Alternate Character-state | 124 | 136 | 111 | 94 | 75 | 17 | 28 | 19 | 0 | 12 | 6 | 7 | 0 | 0 | 0 |
| **Section 2** | | | | | | | | | | | | | | | |
| Eukaryotic Nucleotides | | | | | | | | | | | | | | | |
| A 160 | 132<br>(24) | 126<br>(18) | 129<br>(27) | 130<br>(15) | 143<br>(29) | 152<br>(3) | 156<br>(10) | 158<br>(10) | 160 | 157<br>(2) | 160<br>(3) | 160<br>(2) | 160 | 160 | 160 |
| C 115 | 79<br>(34) | 78<br>(44) | 90<br>(31) | 95<br>(33) | 94<br>(6) | 111<br>(5) | 106<br>(4) | 110<br>(2) | 115 | 110 | 112<br>(1) | 111<br>(1) | 115 | 115 | 115 |
| G 182 | 138<br>(39) | 137<br>(53) | 143<br>(26) | 153<br>(34) | 147<br>(5) | 177<br>(5) | 173<br>(1) | 173<br>(2) | 182 | 179<br>(4) | 179<br>(1) | 179<br>(1) | 182 | 182 | 182 |
| U 96 | 67<br>(27) | 66<br>(21) | 67<br>(27) | 71<br>(12) | 84<br>(35) | 92<br>(4) | 90<br>(13) | 93<br>(5) | 96 | 95<br>(6) | 96<br>(1) | 96<br>(3) | 96 | 96 | 96 |
| Total 553 | | | | | | | | | | | | | | | |
| **Section 3** | | | | | | | | | | | | | | | |
| Unique | 0 | 5 | 1 | 0 | 25 | 8 | 6 | 6 | 0 | 10 | 1 | 2 | 0 | 0 | 0 |
| Prokaryote | 16 | 16 | 16 | 16 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria | 25 | 23 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria | 0 | 2 | 7 | 8 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Eubacteria Only | 19 | 20 | 4 | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria Only | 2 | 6 | 11 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Other | 62 | 64 | 72 | 54 | 45 | 9 | 19 | 12 | 0 | 2 | 4 | 5 | 0 | 0 | 0 |
| **Section 4** | | | | | | | | | | | | | | | |
| Chi Square | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.84 | 0.65 | 0.85 | 1.00 | 0.95 | 0.99 | 0.98 | 1.00 | 1.00 | 1.00 |

Table 7.1b

| Eukaryotic 16S<br>Conserved 90-95% | *Escherichia coli* | *Thermotoga maritima* | *Halobacterium marismortui* | *Sulfolobus acidocaldarius* | *Nosema vespula* | *Giardia muris* | *Physarum polycephalum* | *Entamoeba histolytica* | *Trypanosoma brucei* | *Euglena gracilis* | *Tetrahymena thermophila* | *Plasmodium falciparum* | *Prorocentrum micans* | *Cryptococcus neoformans* | *Arabidopsis thaliana* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Section 1** | | | | | | | | | | | | | | | |
| Nucleotides Absent | 4 | 3 | 2 | 3 | 9 | 3 | 1 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Positions Retained | 61 | 72 | 61 | 71 | 68 | 74 | 108 | 107 | 106 | 101 | 111 | 119 | 129 | 129 | 129 |
| Alternate Character-state | 64 | 54 | 66 | 55 | 52 | 52 | 20 | 19 | 23 | 28 | 17 | 10 | 0 | 0 | 0 |
| **Section 2** | | | | | | | | | | | | | | | |
| Eukaryotic Nucleotides | | | | | | | | | | | | | | | |
| A 32 | 14 | 14 | 16 | 12 | 17 | 14 | 25 | 31 | 26 | 20 | 29 | 31 | 32 | 32 | 32 |
| | (15) | (9) | (11) | (8) | (19) | (10) | (5) | (9) | (9) | (6) | (5) | (7) | | | |
| C 24 | 13 | 17 | 14 | 16 | 10 | 17 | 20 | 16 | 18 | 21 | 18 | 23 | 24 | 24 | 24 |
| | (22) | (22) | (19) | (19) | (6) | (18) | (3) | (1) | (2) | (6) | (5) | | | | |
| G 44 | 23 | 29 | 20 | 31 | 20 | 29 | 37 | 32 | 37 | 36 | 36 | 37 | 44 | 44 | 44 |
| | (15) | (17) | (25) | (18) | (10) | (20) | (5) | | (5) | (10) | (4) | (1) | | | |
| U 29 | 11 | 12 | 11 | 12 | 21 | 14 | 26 | 28 | 25 | 24 | 28 | 28 | 29 | 29 | 29 |
| | (12) | (6) | (11) | (10) | (17) | (4) | (7) | (9) | (7) | (6) | (3) | (2) | | | |
| Total 129 | | | | | | | | | | | | | | | |
| **Section 3** | | | | | | | | | | | | | | | |
| Unique | 0 | 1 | 0 | 0 | 7 | 10 | 5 | 7 | 7 | 13 | 3 | 5 | 0 | 0 | 0 |
| Prokaryote | 12 | 12 | 12 | 12 | 2 | 4 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 |
| Eubacteria | 4 | 3 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria | 0 | 0 | 3 | 3 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria Only | 10 | 10 | 6 | 7 | 0 | 4 | 3 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria Only | 3 | 7 | 19 | 9 | 3 | 5 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Other | 35 | 21 | 26 | 24 | 37 | 26 | 11 | 9 | 14 | 13 | 13 | 4 | 0 | 0 | 0 |
| **Section 4** | | | | | | | | | | | | | | | |
| Chi Square | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.11 | 0.23 | 0.07 | 0.35 | 0.75 | 1.00 | 1.00 | 1.00 |

Table 7.2a

| Eukaryote 23S<br>Conserved >95% | *Escherichia coli* | *Thermotoga maritima* | *Halobacterium marismortui* | *Sulfolobus acidocaldarius* | *Nosema vespula* | *Giardia muris* | *Physarum polycephalum* | *Entamoeba histolytica* | *Trypanosoma brucei* | *Euglena gracilis* | *Tetrahymena thermophila* | *Plasmodium falciparum* | *Prorocentrum micans* | *Cryptococcus neoformans* | *Arabidopsis thaliana* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Section 1** | | | | | | | | | | | | | | | |
| Nucleotides Absent | 6 | 6 | 6 | 4 | 25 | 1 | 3 | 1 | 3 | 0 | 1 | 0 | 1 | 0 | 0 |
| Positions Retained | 748 | 781 | 772 | 822 | 852 | 975 | 960 | 955 | 992 | 946 | 989 | 998 | 995 | 1003 | 1003 |
| Alternate Character-state | 249 | 216 | 225 | 177 | 126 | 27 | 40 | 47 | 8 | 57 | 13 | 5 | 7 | 0 | 1 |
| **Section 2** | | | | | | | | | | | | | | | |
| Eukaryotic Nucleotides | | | | | | | | | | | | | | | |
| A 289 | 245 | 242 | 240 | 243 | 263 | 278 | 279 | 285 | 284 | 264 | 287 | 286 | 288 | 289 | 289 |
| | (63) | (42) | (52) | (35) | (43) | (3) | (11) | (25) | | (9) | (7) | (1) | (4) | | (1) |
| C 208 | 141 | 154 | 153 | 175 | 161 | 201 | 197 | 192 | 207 | 198 | 203 | 208 | 207 | 208 | 208 |
| | (60) | (62) | (66) | (66) | (11) | (7) | (12) | (4) | (5) | (17) | (2) | | (2) | | |
| G 323 | 241 | 259 | 249 | 278 | 265 | 317 | 307 | 298 | 321 | 310 | 318 | 322 | 317 | 323 | 322 |
| | (63) | (68) | (57) | (55) | (17) | (6) | (9) | (3) | (2) | (17) | | (2) | | | |
| U 183 | 121 | 126 | 130 | 126 | 163 | 179 | 177 | 180 | 180 | 174 | 181 | 182 | 183 | 183 | 183 |
| | (63) | (44) | (50) | (21) | (55) | (11) | (8) | (15) | (1) | (14) | (4) | (2) | (1) | | |
| Total 1003 | | | | | | | | | | | | | | | |
| **Section 3** | | | | | | | | | | | | | | | |
| Unique | 0 | 3 | 1 | 1 | 47 | 10 | 11 | 16 | 4 | 26 | 4 | 2 | 5 | 0 | 1 |
| Prokaryote | 27 | 30 | 27 | 27 | 4 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria | 16 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria | 0 | 1 | 8 | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria Only | 48 | 40 | 20 | 10 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Archaebacteria Only | 14 | 18 | 25 | 25 | 3 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| Other | 144 | 114 | 144 | 106 | 69 | 17 | 25 | 29 | 4 | 31 | 7 | 3 | 2 | 0 | 0 |
| **Section 4** | | | | | | | | | | | | | | | |
| Chi Square | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.82 | 0.57 | 0.35 | 0.98 | 0.31 | 0.97 | 0.99 | 0.99 | 1.00 | 1.00 |

Table 7.2b

| Eukaryote 23S<br>Conserved 90-95% | *Escherichia coli* | *Thermotoga maritima* | *Halobacterium marismortui* | *Sulfolobus acidocaldarius* | *Nosema vespula* | *Giardia muris* | *Physarum polycephalum* | *Entamoeba histolytica* | *Trypanosoma brucei* | *Euglena gracilis* | *Tetrahymena thermophila* | *Plasmodium falciparum* | *Prorocentrum micans* | *Cryptococcus neoformans* | *Arabidopsis thaliana* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Section 1** | | | | | | | | | | | | | | | |
| Nucleotides Absent | 3 | 7 | 6 | 4 | 25 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Positions Retained | 154 | 147 | 149 | 149 | 139 | 161 | 195 | 198 | 200 | 175 | 241 | 224 | 239 | 244 | 242 |
| Alternate Character-state | 87 | 90 | 89 | 91 | 80 | 82 | 47 | 45 | 43 | 68 | 3 | 20 | 5 | 0 | 2 |
| **Section 2** | | | | | | | | | | | | | | | |
| Eukaryotic Nucleotides | | | | | | | | | | | | | | | |
| A 77 | 48<br>(15) | 50<br>(9) | 53<br>(24) | 53<br>(13) | 54<br>(33) | 49<br>(14) | 65<br>(14) | 72<br>(15) | 65<br>(13) | 56<br>(10) | 76 | 71<br>(7) | 76<br>(2) | 77 | 76 |
| C 38 | 20<br>(29) | 23<br>(39) | 23<br>(32) | 22<br>(40) | 14<br>(5) | 30<br>(34) | 28<br>(11) | 22<br>(7) | 31<br>(12) | 26<br>(27) | 38<br>(2) | 33<br>(3) | 37<br>(3) | 38 | 38 |
| G 70 | 57<br>(25) | 54<br>(32) | 45<br>(22) | 50<br>(24) | 31<br>(8) | 53<br>(20) | 55<br>(6) | 51<br>(5) | 50<br>(8) | 51<br>(18) | 69 | 54<br>(2) | 68 | 70 | 69<br>(1) |
| U 59 | 29<br>(18) | 20<br>(10) | 28<br>(11) | 24<br>(14) | 40<br>(34) | 29<br>(14) | 47<br>(16) | 53<br>(18) | 54<br>(10) | 42<br>(13) | 58<br>(1) | 66<br>(8) | 58 | 59 | 59<br>(1) |
| Total 244 | | | | | | | | | | | | | | | |
| **Section 3** | | | | | | | | | | | | | | | |
| Unique | 0 | 0 | 2 | 1 | 13 | 19 | 7 | 12 | 12 | 13 | 0 | 5 | 1 | 0 | 1 |
| Prokaryote | 8 | 8 | 8 | 8 | 3 | 4 | 2 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Archaebacteria | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Eubacteria Only | 12 | 11 | 3 | 3 | 2 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| Archaebacteria Only | 3 | 2 | 4 | 4 | 1 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Other | 61 | 68 | 72 | 75 | 61 | 54 | 36 | 33 | 30 | 49 | 3 | 15 | 2 | 0 | 1 |
| **Section 4** | | | | | | | | | | | | | | | |
| Chi Square | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.01 | 0.03 | 0.00 | 0.99 | 0.13 | 0.99 | 1.00 | 0.99 |

Tables 7.3a and b

An analysis of character-state comparisons between the conserved eukaryotic SSU and LSU rRNA nucleotides and the equivalent aligned nucleotide position of 11 eukaryotes. At these positions in the aligned eukaryotic species the character-state was identified as ancestral. Tables 7.3 compares conserved SSU rRNA eukaryotic nucleotides that are conserved in either >95% or 90-95% of all known eukaryotic SSU rRNA sequences. Tables 7.4 compares conserved LSU rRNA eukaryotic nucleotides that are conserved in either >95% or 90-95% of all known eukaryotic LSU rRNA sequences.

Abbreviation are:

**Taxa** - Species and the range of eukaryotic nucleotide conservation compared; **Sec. Struc.** - the type of rRNA subunit secondary structure in which a particular nucleotide position is located; **Align** - the character-state of each aligned nucleotide where the nucleotides respectively are the conserved (>95%) eubacterial nucleotide, the conserved (>95%) archaebacterial nucleotide, the nucleotide of the species in question, and the conserved (as shown being either >95% or 90-95%) eukaryotic nucleotide; **Paired Euk.** - the character-state of the bonded nucleotide from a structural helix where both the nucleotide in question and its bonded partner are conserved in the eukaryotic model; **Paired Sp.** - the character-state of the bonded nucleotide in the structural helix of the species compared where both the nucleotide in question and its bonded partner are conserved in the eukaryotic model; **Type** - indicates whether the character-state of the species at the conserved eukaryotic position is either of the prokaryotic domain (PD), eubacterial only (EO), or archaebacterial only (AO); **Nv. Pos.** - indicates the position of the nucleotide in question on the *N. vespula* secondary models (Figures 6.3 and 6.4a,b); **Bonds** - indicates the position on the *N. vespula* secondary model of the bonded nucleotide from Paired Euk.; **Shared** - indicates other species within the analysis that also share the same 'Type' of nucleotide at the alignment position in question.

Table 7.3a

| Taxa | Sec. Struc. | Align | Paired Euk. | Paired Sp. | Type | Nv Pos. | Bonds | Shared |
|---|---|---|---|---|---|---|---|---|
| Nv | | | | | | | | |
| 95%+ | loop | AAUA | | | PD | 411 | | |
| | loop | AGGA | | | ED | 386 | | |
| | helix | CAGA | C | U | AD | 747 | | Tt |
| | helix | GGAG | U | C | PD | 1195 | | |
| | loop | ZUAU | | | AO | 38 | | |
| 90-95% | helix | ACUA | A | U | ED | 5 | 19 | Gm |
| | helix | CUCU | G | G | AD | 425 | | Eh |
| | helix | GGAG | | | PD | 329 | | Eg |
| | hairpin | GGAG | | | PD | 789 | | Gm, Eg |
| | helix | UGAU | U | A | ED | 19 | 5 | Gm |
| | helix | ZCUC | A | G | AO | 588 | 629 | Gm |
| | helix | ZGAG | | | AO | 531 | | Tt |
| | helix | ZGAG | U | C | AO | 629 | 588 | Gm |
| Gm | | | | | | | | |
| 90-95% | helix | ACUC | A | G | AD | 5 | 19 | Nv |
| | helix | AGGA | U | U | ED | 793 | | |
| | loop | CCUC | | | PD | 871 | | |
| | helix | CCUC | A | G | PD | 1112 | | |
| | helix | CZUC | A | G | EO | 260 | | |
| | helix | CZUC | | | EO | 432 | | |
| | helix | GGAG | U | C | PD | 745 | | |
| | hairpin | GGAG | | | PD | 789 | | Nv, Eg |
| | helix | GZAG | U | C | EO | 245 | | Pp |
| | loop | GZAG | | | EO | 434 | | |
| | helix | UGAG | U | C | AD | 19 | 5 | Nv |
| | helix | ZCUC | A | G | AO | 162 | 93 | |
| | helix | ZCUC | A | G | AO | 588 | 629 | Nv |
| | helix | ZGAG | | | AO | 507 | | |
| | helix | ZGAG | U | C | AO | 629 | 588 | Nv |
| | helix | ZGAG | U | C | AO | 93 | 162 | |
| Pp | | | | | | | | |
| 95%+ | helix | AUUA | A | U | ED | 493 | | |
| | helix | CCUC | | | PD | 233 | | |
| 90-95% | bulge | CZGC | | | EO | 972 | | |
| | helix | GZAG | U | U | EO | 245 | | |
| | helix | UZGU | | | EO | 1074 | | Gm |
| | loop | ZUGU | | | AO | ~ 820 | | Eh |
| Tt | | | | | | | | |
| 95%+ | helix | CAGA | C | U | AD | 747 | | Nv |
| 90-95% | helix | ZGAG | | | AO | 531 | | Nv |
| Tb | | | | | | | | |
| 90-95% | loop | AAGA | | | PD | 982 | | |
| | helix (T) | GGAG | U | U | PD | 451 | | |
| Eg | | | | | | | | |
| 90-95% | loop | GGAG | | | PD | 329 | | Nv |
| | hairpin | GGAG | | | PD | 789 | | Nv, Gm |
| Eh | | | | | | | | |
| >95% | hairpin | AGAG | | | AD | 224 | | |
| 90-95% | helix | CUCU | G | A | AD | 425 | | Nv |
| | helix | CZAC | U | G | EO | 549 | | |
| | helix | UZGU | | | EO | 1074 | | Pp |
| Pf | | | | | | | | |
| 90-95% | hairpin | AAGA | | | PD | 227 | | |

Table 7.3b

| Taxa | Secondary | Alignment | Paired | Paired | Type | *N. vespula* | Bonds with | Shared |
|---|---|---|---|---|---|---|---|---|
| Nv | | | | | | | | |
| 95%+ | loop | AAUA | | | PD | 1404 | | |
| | helix | CCGC | C | G | PD | 533 | 655 | |
| | helix | CGCG | G | U | AD | 2100 | | |
| | loop | GGAG | | | PD | 2201 | | |
| | helix | GGCG | G | C | PD | 655 | 533 | |
| | bulge | GZAG | | | EO | 927 | | |
| | loop | UZGU | | | EO | 932 | | |
| | loop | ZGCG | | | AO | 747 | | Eh |
| | helix | ZGCG | G | C | AO | 1385 | | |
| | loop | ZACA | | | AO | 1403 | | |
| 90-95% | loop | AAGA | | | PD | 844 | | Gm |
| | loop | CCAC | | | PD | 726 | | Eg |
| | helix | GGAG | | | PD | 2010 | | Gm |
| | helix | GZCG | G | U | EO | 993 | | |
| | bulge | GZUG | | | EO | 1672 | | |
| | helix | ZCUC | G | G | AO | 643 | | Pp, Eg |
| Gm | | | | | | | | |
| 90-95% | loop | AAGA | | | PD | 844 | | Nv |
| | hairpin | AAGA | | | PD | 938 | | |
| | helix | AZCA | G | U | EO | 928 | | |
| | helix | CCAC | | | PD | ~ 1536 | | |
| | bulge | CZAC | | | EO | 1753 | | |
| | helix | GGAG | | | PD | 2010 | | Nv |
| | helix | GZUG | | | EO | 616 | | Eg |
| | loop | UAAU | | | ED | ~ 1205 | | |
| | loop | UZCU | | | EO | 2248 | | |
| Tt | | | | | | | | |
| 95%+ | helix | ZACA | G | U | AO | 2105 | 2113 | |
| | helix | ZUGU | C | A | AO | 2113 | 2105 | |
| Pm | | | | | | | | |
| 90-95% | loop | CAAC | | | ED | 2293 | | |
| | hairpin | CZUC | | | AO | 309 | | |
| Eg | | | | | | | | |
| 90-95% | hairpin | AACA | | | PD | 2375 | | Pp |
| | loop | CCAC | | | PD | 726 | | Nv |
| | loop | GGAG | | | PD | 2241 | | Tb |
| | helix | GZUG | | | EO | 616 | | Gm |
| | helix | ZCGC | C | C | AO | 1941 | | |
| | helix | ZCUC | G | G | AO | 643 | | Nv, Pp |
| Pp | | | | | | | | |
| 95%+ | hairpin | CCUC | | | PD | 943 | | |
| | helix | CCUC | A | G | PD | 963 | 970 | |
| | helix | GGAG | U | C | PD | 970 | 963 | |
| | helix(T) | GGCG | | | PD | 1008 | | |
| 90-95% | hairpin | AACA | | | PD | 2375 | | Eg |
| | loop | AAGA | | | PD | 844 | | Nv, Gm |
| | hairpin | ZCGC | | | AO | 851 | | |
| | helix | ZCUC | G | G | AO | 643 | | Nv, Eg |
| Eh | | | | | | | | |
| 95%+ | helix | UZCU | G | A | EO | 979 | | |
| | loop | ZGCG | | | AO | 747 | | Nv |
| Tb | | | | | | | | |
| 95%+ | loop | GGAG | | | PD | 2241 | | Eg |

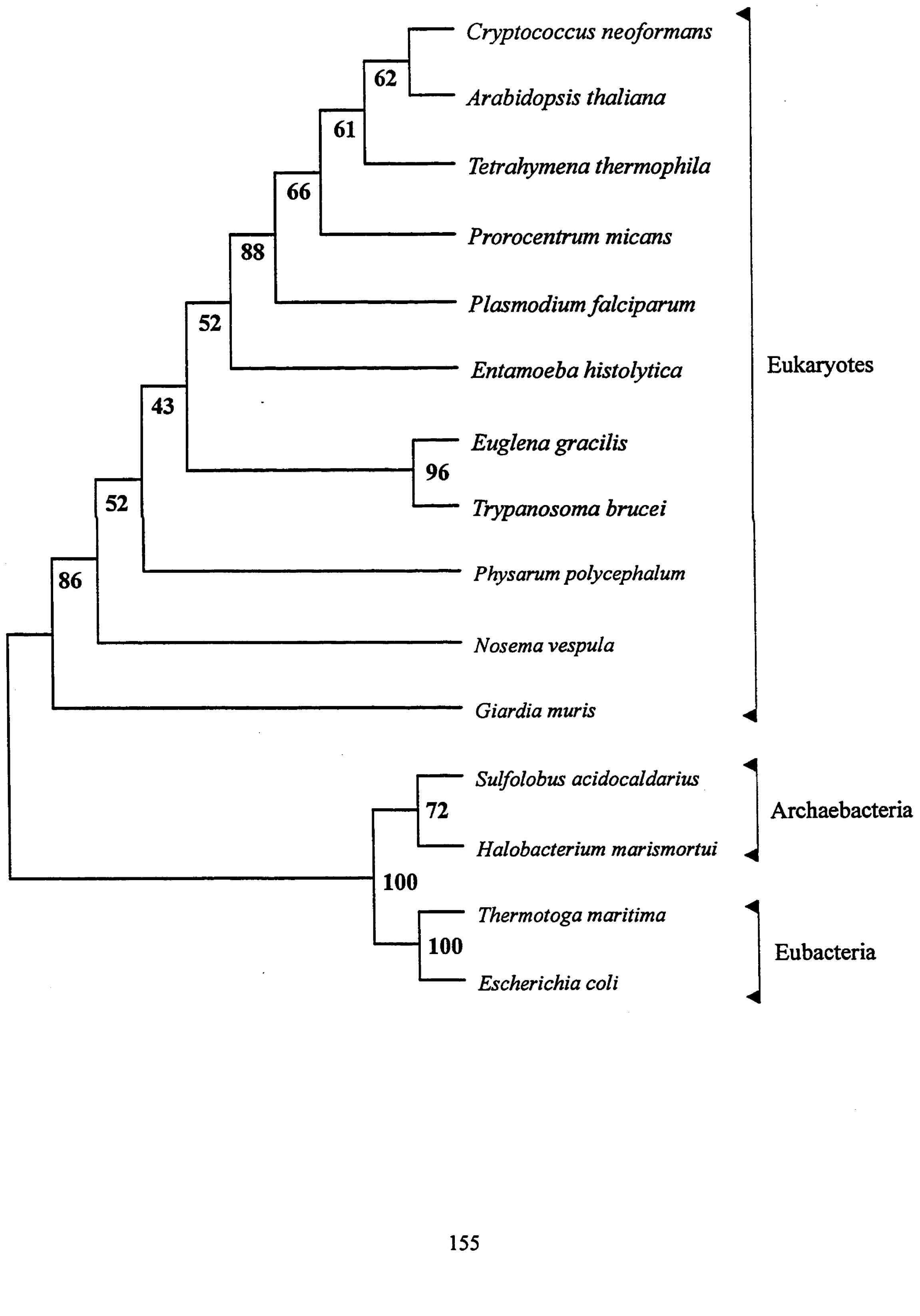
Figure 7.2a and b

SSU rRNA phylogenetic trees. The trees were constructed using maximum parsimony analysis (a) and the neighbor-joining method using Galtier-Gouy distance estimates (b). Bootstrap proportions are indicated at each node for both maximum parsimony analysis and the neighbor-joining method. Horizontal branches are drawn proportional to the inferred evolutionary distance. Substitutions per site are indicated for the neighbor-joining method (see scale). Evolutionary lineage, eubacteria, arcahebacteria or eukaryote, is indicated adjacent to species name in figure 7.2a.

a. Maximum Parismony Tree
1,100 positions (647 informative), 2,642 steps, consensus tree from 300 bootstraps

b. Neighbor-joining Tree (Galtier and Gouy distance)
1,100 positions, 1000 bootstraps

Figure 7.2a



Cryptococcus neoformans
Arabidopsis thaliana
Tetrahymena thermophila
Prorocentrum micans
Plasmodium falciparum
Entamoeba histolytica
Euglena gracilis
Trypanosoma brucei
Physarum polycephalum
Nosema vespula
Giardia muris
Sulfolobus acidocaldarius
Halobacterium marismortui
Thermotoga maritima
Escherichia coli
62
61
66
88
52
43
96
52
86
72
100
100
Eukaryotes
Archaebacteria
Eubacteria

Figure 7.2b

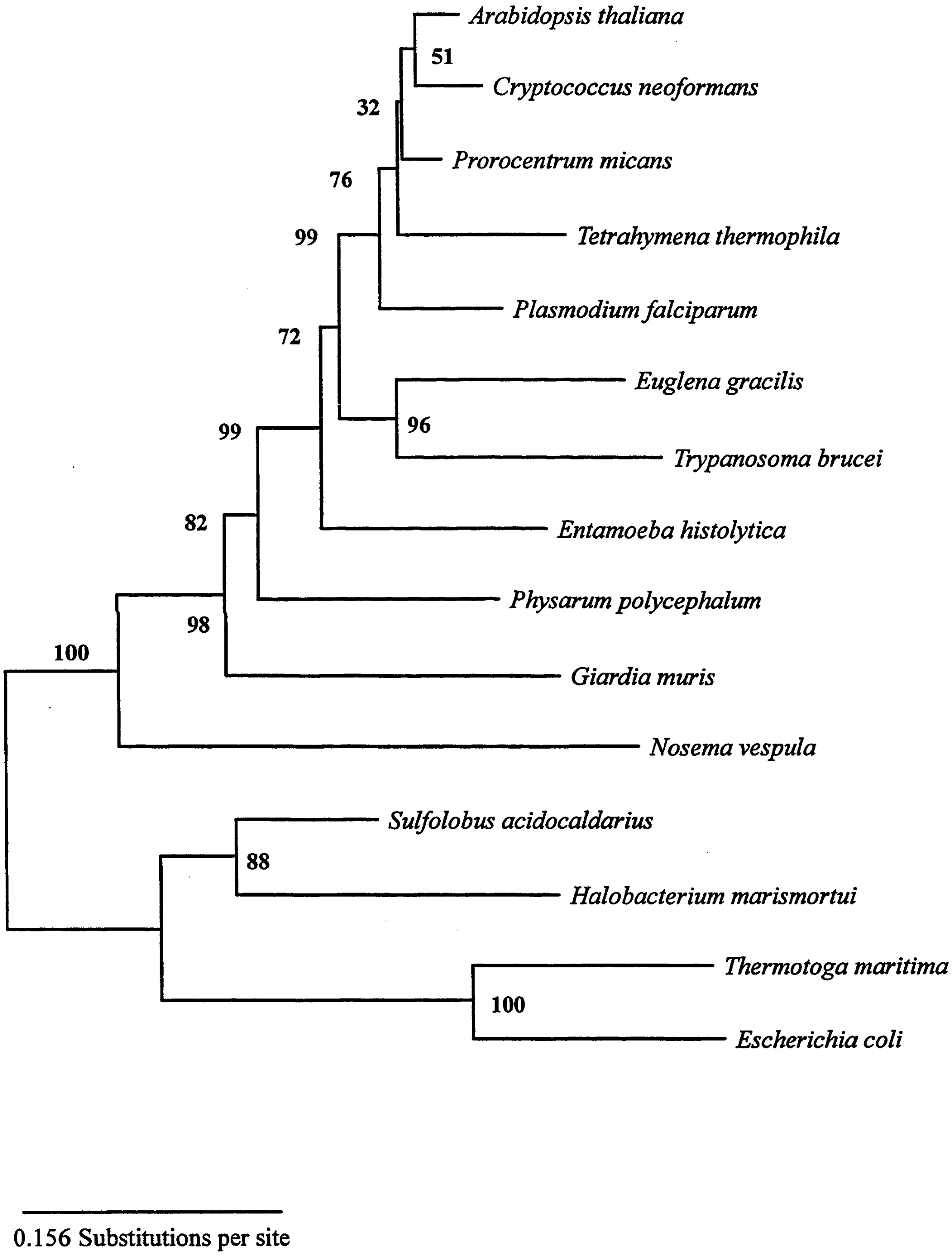

Arabidopsis thaliana
51
Cryptococcus neoformans
32
Prorocentrum micans
76
Tetrahymena thermophila
99
Plasmodium falciparum
72
Euglena gracilis
99
96
Trypanosoma brucei
82
Entamoeba histolytica
Physarum polycephalum
98
100
Giardia muris
Nosema vespula
Sulfolobus acidocaldarius
88
Halobacterium marismortui
Thermotoga maritima
100
Escherichia coli
0.156 Substitutions per site

Figure 7.3a and b

LSU rRNA phylogenetic trees. The trees were constructed through maximum parsimony analysis (a) and the neighbor-joining method using Galtier-Gouy distance estimates (b). Bootstrap proportions are indicated at each node for both maximum parsimony analysis and the neighbor-joining method. Horizontal branches are drawn proportional to the inferred evolutionary distance. Substitutions per site are indicated for the neighbor-joining method (see scale). Evolutionary lineage, eubacteria, arcahebacteria or eukaryote, is indicated adjacent to species name in figure 7.2a.

c. Maximum Parismony Tree
2,008 positions (1179 informative), 5,297 steps, consensus tree from 300 bootstraps

d. Neighbor-joining Tree (Galtier and Gouy distance)
1,100 positions, 1000 bootstraps

Figure 7.3a

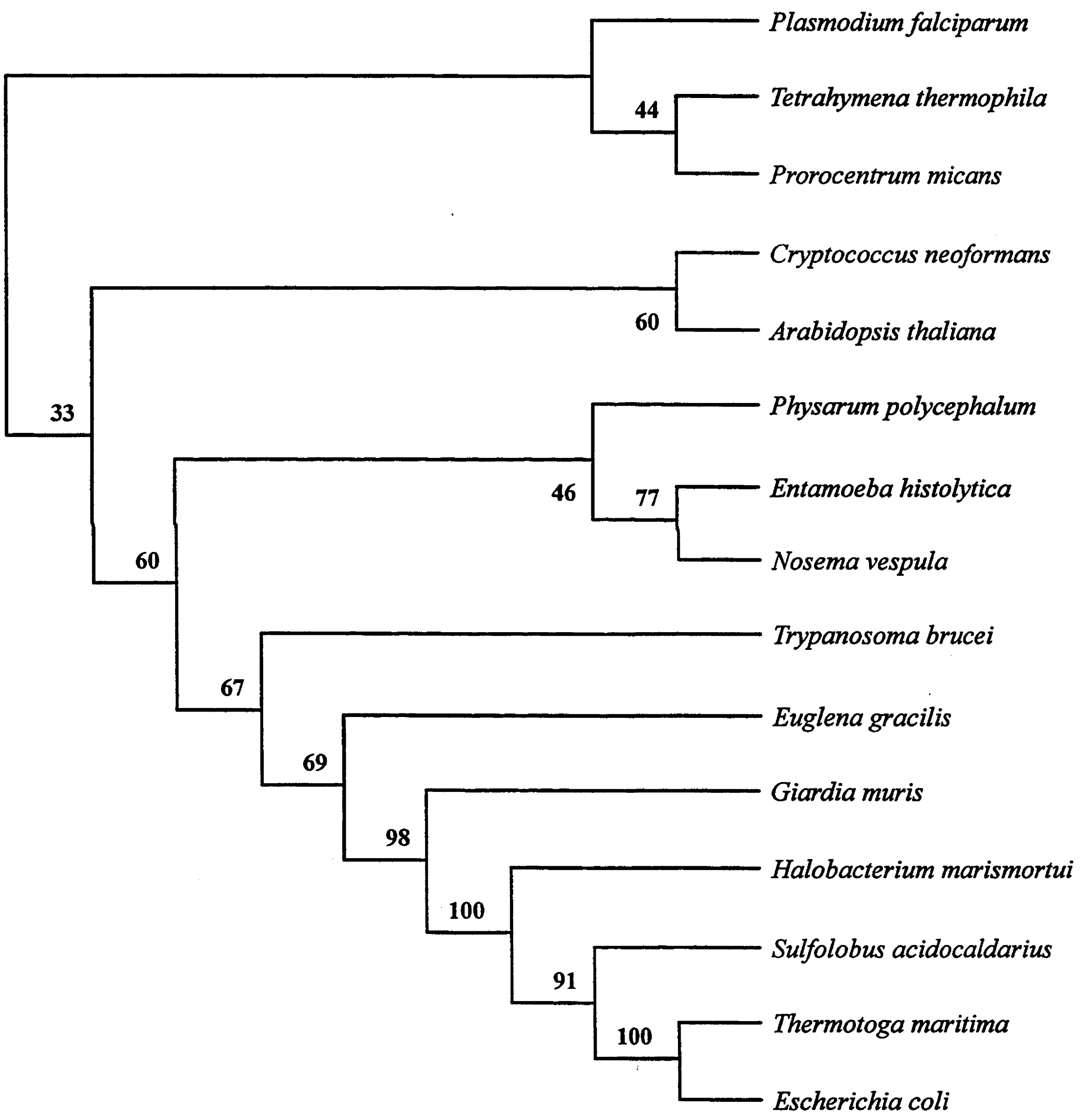
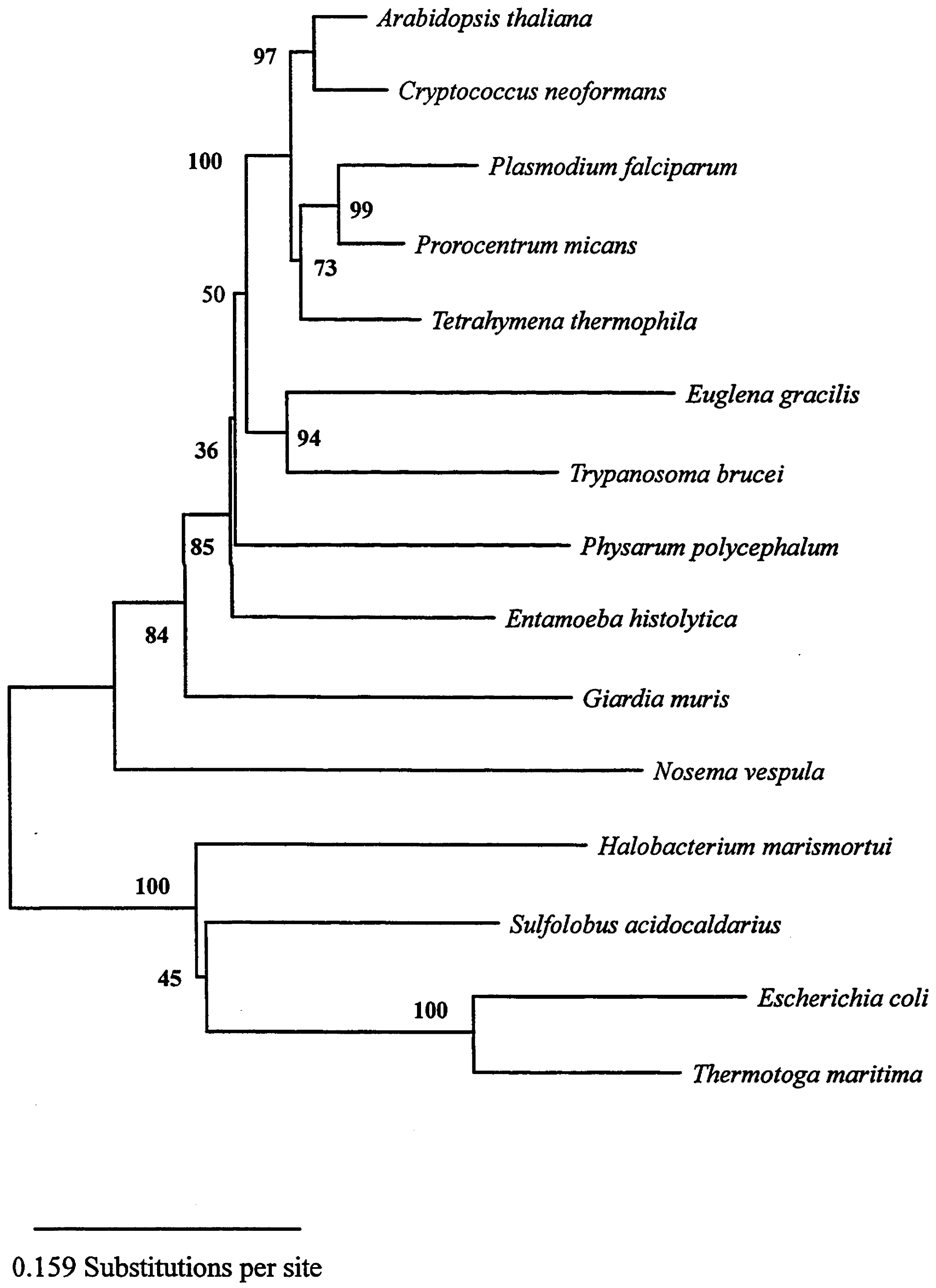

Plasmodium falciparum
44
Tetrahymena thermophila
Prorocentrum micans
Cryptococcus neoformans
60
Arabidopsis thaliana
33
Physarum polycephalum
46
77
Entamoeba histolytica
Nosema vespula
60
Trypanosoma brucei
67
Euglena gracilis
69
Giardia muris
98
Halobacterium marismortui
100
Sulfolobus acidocaldarius
91
Thermotoga maritima
100
Escherichia coli

Figure 7.3b


Arabidopsis thaliana
97
Cryptococcus neoformans
100
Plasmodium falciparum
99
Prorocentrum micans
73
50
Tetrahymena thermophila
Euglena gracilis
94
36
Trypanosoma brucei
85
Physarum polycephalum
Entamoeba histolytica
84
Giardia muris
Nosema vespula
Halobacterium marismortui
100
Sulfolobus acidocaldarius
45
100
Escherichia coli
Thermotoga maritima
0.159 Substitutions per site

# CHAPTER 8

# GENERAL CONCLUSION

## 8.1 CONCLUSION

The cellular and molecular characteristics of the Microsporidia, Diplomonadida and the other archaezoans has lead to them being placed at or near the base of the eukaryotic lineage. This hypothesis was initially supported by phylogenetic evidence based on SSU rRNA sequence data. However, subsequent research revealed cellular and molecular data that conflicted with these placement, in particular the placement of the Microsporidia.

The research presented in this thesis has sought to provide further evidence as to the phylogenetic status of the Microsporidia by examining molecular aspects of the microsporidial rRNA genes not previously studied. Essentially, four aspects of the microsporidian rRNA genes were examined: 1) arrangement of the rRNA genes within the operon; 2) the secondary structure and tertiary interactions therein, including the presences of prokaryotic-like motifs; 3) sequence variability and the retention of plesiomorphic nucleotides; and 4) phylogenetic inferences using the SSU and LSU rRNA sequences.

The genes of the rDNA operon of *N. vespula* are arranged in a prokaryotic-like manner—the 5.8S and LSU rRNA genes are covalently linked and the 5S rRNA gene is located downstream from the LSU rRNA gene. However, this location for the 5S rRNA gene is not consistent in all Microsporidia. It also appears that the 5S rRNA gene is transcribed separately by RNA polymerase III, as indicated by the presences of the RNA polymerase III termination signal.

The secondary structure comparisons are more supportive of an early divergence of the Microsporidia. Within the secondary structure of the rRNA subunits are prokaryotic-like structure motifs. Furthermore, within some of these motifs are nucleotides that share the highly conserved prokaryotic character-state at positions that in eukaryotes are also highly conserved but differ in character-state. In addition to these prokaryotic nucleotides are other instances of prokaryotic nucleotides at positions whose character-state in eukaryotes differs but is also highly conserved.

It was demonstrated that the sequences of the SSU and LSU rRNA genes are significantly divergent from other eukaryotes. It was demonstrated that many of the highly conserved nucleotides either are of a different character-state or are simply absent. This observation confirms the Microsporidia have undergone a process of degenerate evolution. This high degree of sequence divergence has a dramatic impact on accuracy when trying to recover the true evolutionary tree using tree building algorithms/methods; particularly for the inter-lineage comparisons.

Finally, phylogenetic reconstructions that include highly divergent sequences or sequences that exhibit character-state bias suffer from a number of misleading phenomena (eg., long branch attraction). Therefore, in reconstructing phylogenies of distant related taxa, it is possible that the 'noise' of evolution, when comparing the results of various molecular data, may in fact disguise the remnants of the true evolutionary process. Under these circumstances the presence or absence of such remnant evidence, being either nucleotide or protein sequence, can only be revealed through detailed comparative alignment analysis as presented here.

More so than any other eukaryote studied, the rRNA genes of the Microsporidia are highly divergent and truncated, and yet are a collection of prokaryotic-like and eukaryotic features. The most parsimonious explanation for the presence of these prokaryotic-like features is by direct inheritance. Therefore, I conclude that the analysis presented supports the early divergence of the Microsporidia and the Diplomonadida and that the divergence of the Microsporidia is likely to have predated that of the diplomonads.

# APPENDECES

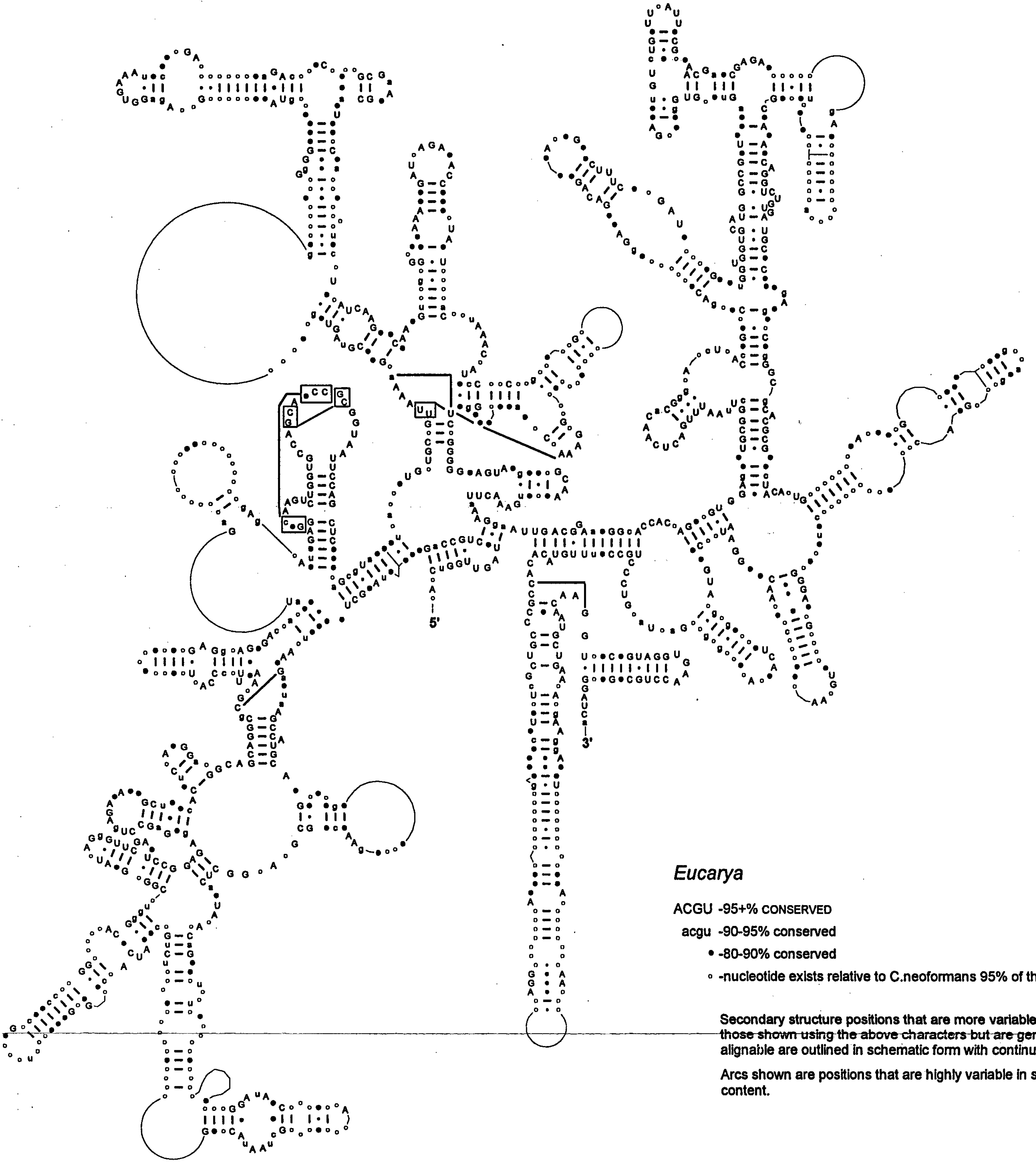
### Appendix 1 SSU rRNA model

The generic eukaryotic SSU rRNA model (Gutell, 1994a).

### Appendix 2a, b LSU rRNA model

The generic eukaryotic LSU rRNA model (Gutell *et al.,* 1993).

# Phylogenetic conservation superimposed onto the *Cryptococcus neoformans* SMALL SUBUNIT RIBOSOMAL RNA secondary structure



*Eucarya*

ACGU -95+% CONSERVED

acgu -90-95% conserved

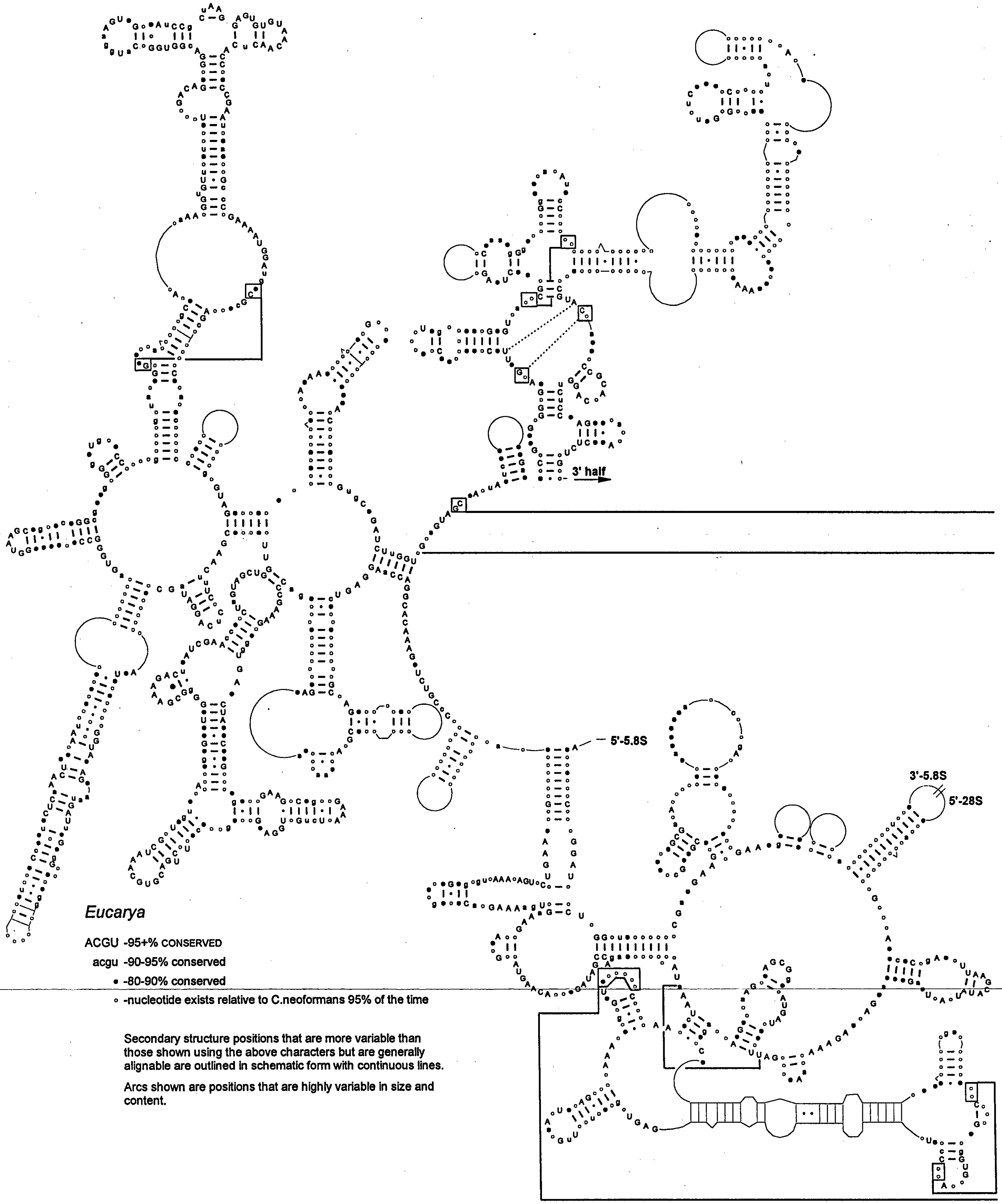
● -80-90% conserved

○ -nucleotide exists relative to C.neoformans 95% of the time

Secondary structure positions that are more variable than those shown using the above characters but are generally alignable are outlined in schematic form with continuous lines.

Arcs shown are positions that are highly variable in size and content.

# Phylogenetic conservation superimposed onto the *Cryptococcus neoformans* LARGE SUBUNIT RIBOSOMAL RNA secondary structure



*Eucarya*

ACGU -95+% CONSERVED

acgu -90-95% conserved

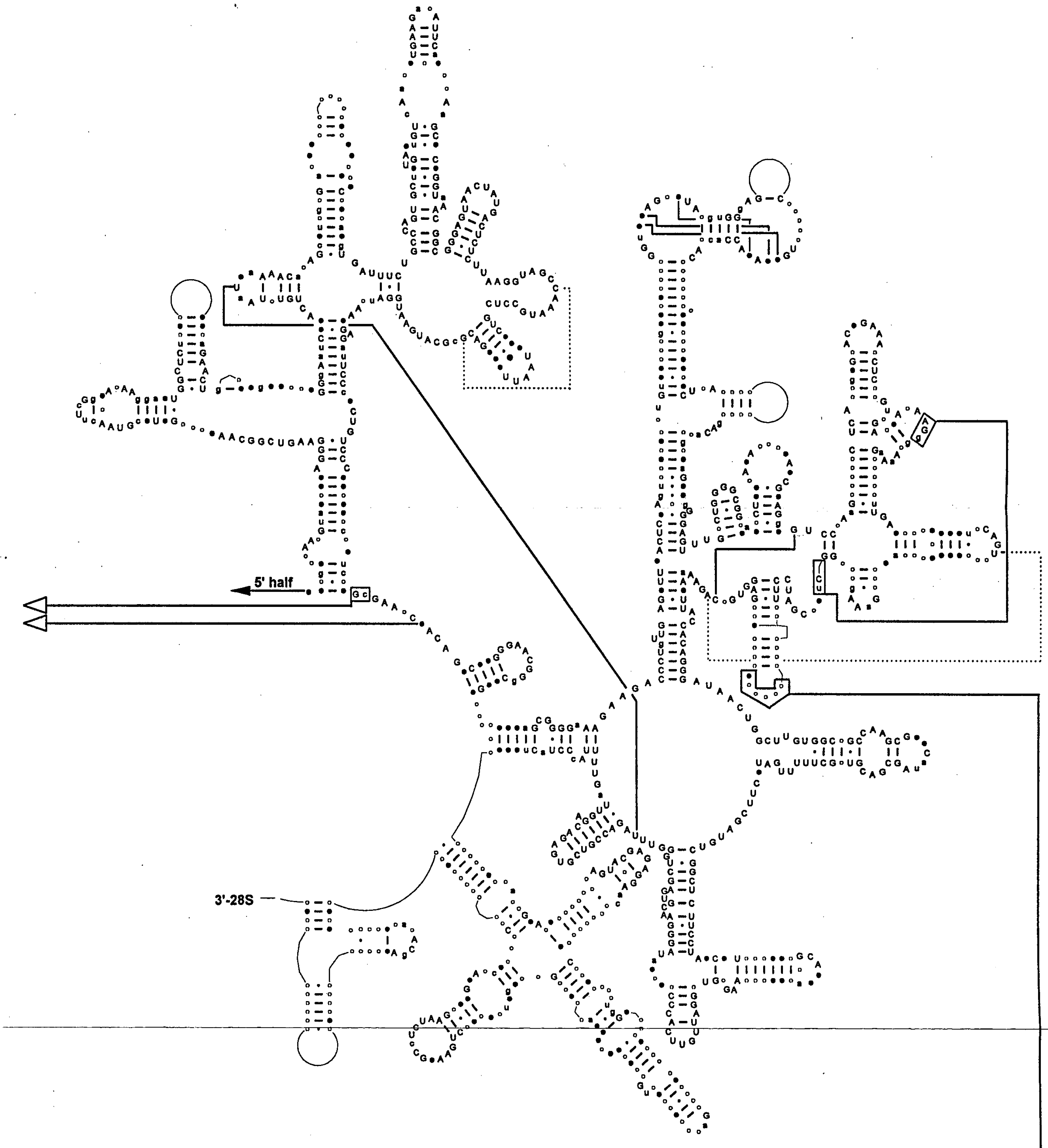
• -80-90% conserved

∘ -nucleotide exists relative to C.neoformans 95% of the time

Secondary structure positions that are more variable than those shown using the above characters but are generally alignable are outlined in schematic form with continuous lines.

Arcs shown are positions that are highly variable in size and content.

# Phylogenetic conservation superimposed onto the *Cryptococcus neoformans* LARGE SUBUNIT RIBOSOMAL RNA secondary structure



Secondary structure helices that are more variable than those shown using the above characters but are generally alignable are outlined in schematic form with continuous lines.

Arcs shown are positions that are highly variable in size and content.

*Eucarya*

ACGU -95+% CONSERVED

acgu -90-95% conserved

● -80-90% conserved

○ -nucleotide exists relative to C.neoformans 95% of the time

# REFERENCES

Aagaard, C. and Douthwaite, S. 1994. Requirements for a conserved, tertiary interaction in the core of 23S ribosomal RNA. *Proc. Natl. Acad. Sci. USA,* 91: 2989-2993.

Aksoy, S., Shay, G.L., Villanueva, M.S., Beard, C.B. and Richards, F.F. 1992. Spliced leader RNA sequence of *Trypanosoma rangeli* are organized within the 5S rRNA-encoding genes. *Gene,* 113: 239-243.

Andreadis, T.G. 1985. Experimental transmission of a microsporidian pathogen from mosquitoes to an alternate copepod host. *Proc. Natl. Acad. Sci. USA,* 82: 5574-5577.

Anderson, D.L. CSIRO, Division of Entomology, Black Mountain, Canberra, Australian Capital Territory, Australia.

Andreadis, T.G. and Hall, D.W. 1979. Development, ultrastructure, and mode of transmission of *Amblyospora sp.* (Microspora) in the mosquito. *J. Protozool.*, 26: 444-452.

Appels, R. and Honeycutt, R.L. 1987. rDNA: evolution over a billion years. In S.K. Dutta (ed.). *DNA Systematics,* pp. 81-135. CRC Press, Boca Raton, USA.

Bailey, L. and Ball, B.V. 1991. *Honey Bee Pathology* (2nd edition). pp 62 - 74. Academic Press, London.

Baker, M.D., Vossbrinck, C.R., Maddox, J.V. and Undeen, A.H. 1994. Phylogenetic relationships among *Vairimorpha* and *Nosema* species (Microspora) based on ribosomal RNA sequence data. *J. Invertebr. Pathol.*, 64: 100-106.

Baker, M.D., Vossbrinck, C.R., Dider, E.S., Maddox, J.V. and Shadduck, J.A. 1995. Small subunit ribosomal DNA phylogeny of various microsporidia with emphasis on AIDS related forms. *J. Euk. Microbiol.*, 42(5): 564-570.

Baranov, P.V., Sergiev, P.V., Dontsova, O.A., Bogdanov, A.A. and Brimacombe, R. 1998. The database of ribosomal cross links (DRC). *Nucl. Acids Res.*, 26: 187-189.

Barciszewska, M.Z., Erdmann, V.A. and Barciszewshi, J. 1996. Ribosomal 5S RNA: tertiary structure and interactions with proteins. *Biol. Rev.*, 71: 1-25.

Barnes, W.M. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from λ bacteriophage templates. *Proc. Natl. Acad. Sci. USA,* 91: 2216-2220.

Biderre C., Pagès, M., Méténier, G., David, D., Bata, J., Prensier, G. and Vivarès, C.P. 1994. On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates. *C. R. Acad. Sci. Paris.* 317: 399-404.

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

Branlant, C., Krol, A., Machatt, M.A., Pouyet, J., Ebel, J. P., Edwards, K. and Kössel, H. 1981. Primary and secondary structures of *Escherichia coli* MRE 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs. *Nucl. Acids Res.,* 9: 4303-4324.

Brimacombe, R. 1984. Conservation of structure in ribosomal RNA. *Trends Biochem. Sci.*, 9: 273-277.

Brown, J.R. and Doolittle, W.F. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Pro. Natl. Acad. Sci. USA,* 92: 2441-2445.

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

Bulla, Jr., L.A. and Cheng, T.C. 1997. Systematics of the Microsporidia. In L. A. Bulla Jr. and T. C. Cheng (eds.) and V. Sprague and Jírí Vávra Contributing (eds.). *Comparative Pathobiology.* Vol. 2. Plenum Press, New York.

Buys, B. 1972. Nosema in brood. *South African Bee J.*, 44: 2-4.

Buys, B. 1977. A nosema disease affecting honeybee brood. In S. Colibaba (ed.). *Symposium of Bee Biology and Pathology.* pp. 73-76. Apimondia Publishing House, Bucharest, Romania.

Cali, A. and Owen, R.L. 1988. Microsporidiosis. In A. Barlows, W. Hausler, Jr. and E.H. Lennette (eds.). *The Laboratory Diagnosis of Infectious Diseases: Principles and Practice.* Vol 1. pp. 928-949. Springer-Verlag, New York.

Canning, E.U. 1953. A new microsporidian, *Nosema locustae* n. sp., from the fat body of the African migratory locust, *Locusta migratoria migratorioides* R. and F. *Parasitology*, 43: 287-290. [ Cited from Bulla and Cheng (1977)]

Canning, E.U. 1989. Phylum Microspora. In L. Margulis, J.O. Corliss, M. Melkonian and D.J. Chapman (eds.). *Handbook of Protoctista.* Jones and Bartlett, Publishers, Boston.

Canning, E. U. and Hollister, W.S. 1987. Microsporidia of mammals-widespread pathogens or opportunistic curiosities? *Parasitol. Today,* 3: 267-273.

Canning, E. U. 1993. Microsporidia. In J. P. Kreier and J. R. Baker (ed.), *Parasitic protozoa.* 2$^{nd}$ ed., Vol. 6. pp. 299-385. Academic Press, Inc., New York.

Cantwell, G.E. 1970. Standard methods for counting nosema spores. *Amer. Bee. J.*, June: 22-23.

Cao, Y., Adachi, J., Yano, T. and Hasegawa, M. 1994a. Phylogenetic place of guinea pigs: No support of the rodent-poly-phyly hypothesis from maximum-likelihood analyses of multiple protein seqquences. *Mol. Biol. Evol.,* 11: 593-604.

Cao, Y., Adachi, J., Janke, A., Pääbo, S. and Hasegawa, M. 1994b. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J. Mol. Evol.,* 39: 519-527.

Cavalier-Smith, T. 1987. Eukaryotes with no mitochondria. *Nature,* 326: 332-333.

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

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

Cedergren, R., Gray, M.W., Abel, Y. and Sankoff, D. 1988. The evolutionary relationships among known life forms. *J. Mol. Evol.,* 28: 98-112.

Cheng, S., Fockler, C., Barnes, W.M. and Higuchi, R. 1994a. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci. USA,* 91: 5695-5699.

Cheng, S., Chang, S., Gravitt, P. and Respess, R. 1994b. Long PCR. *Nature,* 369: 684-685.

Christian, P., CSIRO, Division of Entomology, Black Mountain, Canberra, Australian Capital Territory, Australia.

Clark, T.B. 1980. A second microsporidian in the honeybee. *J. Invertbr. Pathol.,* 35: 290-294.

Clark, C.G., Tague, B.W., Ware, V.C. and Gerbi, S.A. 1984. *Xenopus laevis* 28S ribosomal RNA: a secondary structure model and its evolutionary and functional implications. *Nucl. Acids Res.,* 12: 6197-6220.

Clark, C.G., Martin, D.S. and Diamond, L.S. 1995. Phylogenetic relationships among trypanosomes as revealed by riboprinting. *J. Euk. Microbiol.,* 42: 92-96.

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

Clark, C.G. and Diamond, L.S. 1997. Intraspecific variation and phylogenetic relationships in the genus *Entamoeba* as revealed by riboprinting. *J. Euk. Microbiol.,* 44: 142-154.

Cordesse, F., Cooke, R., Tremousaygue, D., Grellet, F. and Delseny, M. 1993. Fine structure and evolution of the rDNA intergenic spacer in rice and other cereals., *J. Mol. Evol.* 36: 369-379.

Croizer, Y.C., Koulianos, S. and Croizer, R.H. 1991. Am improved test for africanized

Croizer, Y.C., Koulianos, S. and Croizer, R.H. 1991. Am improved test for africanized honeybee mitochondrial DNA. *Experientia*, 47: 968-969.

Cruces, J., Diaz-Guerra, M., Gil, I. and Renart, J. 1989. The 5S rRNA-histone repeat in the crustacean *Artemia*: structure, polymorphism and variation of the 5S rRNA segment in different populations. *Nucl. Acids Res.*, 17: 6283-6297.

Curgy J.J., Vávra, J. and Vivarès, C. 1980. Presence of ribosomal RNAs with prokaryotic properties in Microsporidia, eukaryotic organisms. *Biol. Cellulaire*, 38: 49-52.

Dalrymple, B.P., Dimmock, C.M., Parrodi, F. and Wright, I.G. 1992. *Babesi bovis, Babesi bigemina, Babesi canis, Babesia microti* and *Babesia* rodhaini: comparison of ribosomal RNA gene organization. *Int. J. Parasitol.,* 22: 851-855.

De Graaf, D.C., Masschelein, G., Vandergeynst, F., De Brabander, H.F. and Jacobs, F.J. 1993. *In vitro* germination of *Nosema apis*(Microspora: Nosematidae) spores and its effect on their alpha,alpha-trehalose/D-glucose ratio. *J. Invertbr. Pathol.,* 62: 220-225.

De Rijk, P. and De Wachter, R. 1993. DCSE v2.54, an interactive tool for sequence alignment and secondary structure research. *Comput. Applic. Biosci.,* 9: 735-740.

De Rijk, P., Van de Peer, Y., Van Den Broeck, I. and De Wachter, R. 1995. Evolution according to large ribosomal subunit RNA. *J. Mol. Evol.*, 41: 366-375.

De Rijk, P., Caers, A., Van de Peer, Y., De Wachter, R. 1998. Database on the structure of large ribosomal subunit RNA. *Nulc. Acids Res.,* 26: 183-186.

De Winter, R.F.J. and Moss, T. 1987. A complex array of sequences enhances ribosomal transcription in *Xenopus laevis*. *J. Mol. Biol.,* 196: 813-827.

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

Doelling, J. H., Gaudino, R.J. and Pikaard, C.S. 1993. Functional analysis of *Arabidopsis thaliana* rRNA gene and spacer promoters *in vivo* and by transient expression. *Proc. Natl.*

Doelling, J. H., Gaudino, R.J. and Pikaard, C.S. 1993. Functional analysis of *Arabidopsis thaliana* rRNA gene and spacer promoters *in vivo* and by transient expression. *Proc. Natl. Acad. Sci. USA*, 90: 7528-7532.

Drouin, G. and Moniz de Sá, M. 1995. The concerted evolution of 5S ribosomal genes linked to the repeat units of other multigene families. *Mol. Biol. Evol.,* 12: 481-493.

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

Ehresmann, B., Ehresmann, C., Romby, P., Mougel, M., Baudin, F., Westof, E. and Ebel, J. 1990. In W.E. Hill, A.E. Dahlberg, R.A. Garrett, P.B. Moore, D. Schlessinger and J.R. Warner (eds.). *The Ribosome, Structure, Function and Evolution.* pp. 148-159. Am. Soc. Microbiol., Washington.

Erdmann, V.A., Wolters, J., Huysmans, E. and DeWachter, R. 1985. Collection of published 5S, 5.8S and 4.5S ribosomal RNA sequences. *Nucleic Acids Res.* 13:r105-r153.

Erlandson, M.A., Mukerji, M.K., Ewen, A.B. and Gillot, C. 1985. Comparative pathogenicity of *Nosema acridophagus* Henry and *Nosema cuneatum* Henry (Microsporidia: Nosematidae) for *Melanoplus sanguinipes* (FAB) (Orthoptera: Acrididae). *Canadian Entomol.,* 117: 1167-1175.

Eschbach , S., Wolters, J. and Sitte, P. 1991. Primary and secondary structure of the nuclear small subunit ribosomal RNA of the cryptomonad *Pyrenomonas salina* as inferred from the gene sequence: evolutionary implications. *J. Mol. Evol.,* 32: 247-252.

Farris J.S. 1983. The logical basis of phylogenetic analysis. *Advan. Clad.,* 2: 7-36.

Faulkner, D.V. and Jurka, J. 1988. Multiple alignment sequence editor (MASE). *Trends In Biochem. Sci.,* 13: 321.

Felsenstein, J. 1972. Numerical methods for inferring evolutionary trees. *Quar. Rev. Bio.,* 57: 379-404.

Felsenstein, J. 1978. Cases in which parsimony and compatability methods will be positively misleading. *Syst. Zool.*, 27: 401-410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17: 368-376.

Felsenstein, J. 1982. Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.*, 57: 379-404.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* 39: 783-791.

Felsenstein, J. 1988. Phylogenies from molecular sequences: Inferences and reliability. *Ann. Rev. Genet,* 22: 521-565.

Felsenstein, J. 1993. *PHYLIP (Phylogeny Inference Package)*, version 3.5c. Department of Genetics, University of Washington, Seattle.

Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic trees. *Science,* 155: 279-284.

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.*, 20: 406-416.

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

Fox, G.E. and Woese, C.R. 1975. 5S RNA secondary structure. *Nature*, 256: 505-507.

Fries. I. 1993. *Nosema apis*—A parasite in the honey bee colony. *Bee World,* 74: 5-19.

Fries, I., Feng, F., da Silva, A., Slemenda, S.B. and Pieniazek, N.J. 1996. *Nosema ceranae* n. sp. (Microspora, Nosematidae), morphological and molecular characterisation of a microsporidian parasite of the asian honey bee *Apis cerana* (Hymenoptera, Apidae). *Europ. J. Protistol.*, 32: 356-365.

Galtier, N. and Gouy, M. 1995. Inferring phylogenies from DNA sequences of unequal base composition. *Proc. Natl. Acad. Sci. U.S.A,* 92: 11317-11321.

Galtier, N., Gouy, M. and Gautier, C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *CABIOS,* 12: 543-548.

Gatehouse, H.S. and Malone, L.A. 1998. The ribosomal RNA gene region of *Nosema apis* (Microspora): DNA sequence for small and large subunit rRNA genes and evidence of a large tandem repeat. *J. Invertbr. Pathol.*, 71: 97-105.

Georgiev, O.I., Nikolaev, N., Hadjiolov, A.A., Skryabin, K.G., Zakharyev, V.M. and Bayev, A.A. 1981. The structure of the yeast ribosomal RNA genes. Complete sequences of the 25S rRNA gene from *Saccharomyces cerevisiae*. *Nucl. Acids Res.*, 9: 6953-6958.

Gerbi, S. A. 1985. Evolution of ribosomal DNA. In R.J. MacIntyre (ed.). *Molecular Evolutionary Genetics*. pp. 419-517. Plenum, New York.

Germot, A., Philippe, H. and Le Guyader, H. 1996. Presence of a mitochondrial-type HSP70 in *Trichomonas* suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proc. Natl. Acad. Sci. USA,* 93: 14614-14617.

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

Glotz,, C., Zwieb, C., Brimacombe, R., Edwards, K. and Kössel, H. 1981. Secondary structure of the large subunit ribosomal RNA from *Escherichia coli, Zea mays* chloroplast, and human and mouse mitochondrial ribosomes. *Nucl. Acids Res.*, 9: 3287-3306.

Gray, M.W., Sankoff, D. and Cedergren, R.J. 1984. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucl. Acids Res.,* 12: 5837-5852.

Gray, M.W. and Schnare, M.N. 1990. Evolution of the modular structure of rRNA. In W.E. Hill, A.E. Dahlberg, R.A. Garrett, P.B. Moore, D. Schlessinger and J.R. Warner (eds.). *The Ribosome, Structure, Function and Evolution*. pp. 589-597. Am. Soc. Microbiol., Washington.

Grimaldi, G. and Dinocera, P.P. 1988. Multiple repeat units in *Drosophila melanogaster* ribosomal DNA spacer stimulates rRNA precursor transcription. *Proc. Natl. Acad. Sci. USA*, 85: 5502-5506.

Guay, J., Huot, A., Gagnon, S., Tremblay, A. and Levesque, R. 1992. Physical and genetic mapping of cloned ribosomal DNA from *Toxoplasma gondii*: primary and secondary structure of the 5S gene. *Gene*, 114: 165-171.

Gutell, R.R., Gray, M.W. and Schnare, M.N. 1993. A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucl. Acids Res.,* 21: 3055-3074.

Gutell, R.R. and Fox, G.E. 1988. A compilation of large subunit RNA sequences presented in a structural format. *Nucl. Acids Res.,* 16(Suppl.), r175-r269.

Gutell, R.R. and Woese, C.R. 1990. Higher order structural elements in ribosomal RNAs: pseudoknots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. USA,* 87: 663-667.

Gutell, R.R., Gray, M.W. and Schnare, M.N. 1992. A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucl. Acids Res.,* 20 : 2095-2109.

Gutell, R.R. 1994a. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucl. Acids Res.,* 22: 3502-3507.

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

Hadjiolov, A.A., Georgiev, O.I., Nosikov, V.V. and Yarachev, L.P. 1984. Primary and secondary structure of rat 28S ribosomal RNA. *Nucl. Acids Res.,* 12: 3677-3693.

Hartskeerl, R.A., Schuitema, A.R.J. and deWachter, R. 1993. Secondary structure of the small subunit ribosomal RNA sequence of the microsporidium *Encephalitozoon cuniculi*. *Nucl. Acids Res.*, 21: 1489.

Hasegawa, M. and Hashimoto, T. 1993. Ribosomal RNA trees misleading? *Nature,* 361: 23-23.

Haselman, T., Gutell, R.R., Jurka, J. and Fox, G.E. 1989. Additional Watson-Crick interactions suggest a structural core in large subunit ribosomal RNA. *J. Biomol. Struct. Dynam.,* 7: 181-186.

Hausmann, K. and Hülsmann, N. 1996. *Protozoology* (2nd Edition). Thieme Medical Publishers, Inc., New York, U.S.A.

Hendriks, L., De Baere, R., Van de Peer, Y., Neefs, J., Goris, A. and De Wachter, R. 1991. The evolutionary position of the rhodophyte *Porphyra umbilicalis* and the basidiomycete *Leucosporidium scotti* among other eukaryotes as deduced frm complete sequences of small ribosomal subunit RNA. *J. Mol. Evol.,* 32: 167-177.

Henry, J.E. 1967. *Nosema acridophagus* sp. n., a microsporidian isolated from grasshoppers. *J. Invertebr. Pathol.,* 9: 331-341. [ Cited from Bulla and Cheng (1977)]

Henry, J.E. 1971. *Nosema cuneatum* sp. n. (Microsporidia: Nosematidae) in grasshoppers (Orthoptera: Acridiadae). *J. Invertebr. Pathol.,* 17: 164-171. [ Cited from Bulla and Cheng (1977)]

Henze, K., Badr, A., Wettern, M., Cerff, R. and Martin, W. 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses in protist evolution. *Proc. Natl. Acad. Sci. USA,* 92: 9122-9126.

Hill, W.E., Dahlberg, A.E., Garrett, R.A., Moore, P.B., Schlessinger, D. and Warner, J.R., (eds.). 1990. *The Ribosome, Structure, Function and Evolution.* Am. Soc. Microbiol., Washington.

Hillis, D. M. and Moritz, C., Mable B.K. (Eds.), 1996. Molecular Systematics (2nd Edition). Sinauer Associates. Inc., Massachusetts, U.S.A.

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

Hogan, J.J., Gutell, R.R. and Noller, H.F. 1984. Probing the conformation of 26S rRNA in yeast 60S ribosomal subunits with kethoxal. *Biochemistry,* 23: 3330-3335.

Höpfl, P., Ludwig, W., Schleifer, H and Larsen, N. 1989. Higher order structure 23S rRNA of *Pseudomonas cepacia* and other prokaryotes. *Eur. J. Biochem.*, 185: 355-364.

Horner, D.S., Hirt, R.P., Kilvington, S., Lloyd, D. and Embley, T.M. 1996. Molecular data suggest an early acquisition of the mitochondrion endosybiont. *Proc. R. Soc. Lond. [Biol].,* 1263: 1053-1059.

Huang, M.M., Arnheim, N. and Goodman, M.F. 1992. Extension of base mispairs by Taq DNA polymerase: Implications for single nucleotide discrimination in PCR. *Nucl. Acids Res.,* 20: 4567-4573.

Ishihara, R. and Hayashi. 1968. Some properties of ribosomes from sporoplasm of *Nosema bombycis. J. Invertebr. Pathol.*, 11: 377-385.

Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. In H.H. Munro (ed.). *Manual of Protein Metabolism.* pp. 21-32. Academic Press, New York.

Kawai, H., Muto, H., Fuji, T. and Kato, A. 1995. A linked 5S rRNA gene in *Scytosiphon lomentaria* (Scytosiponales, Phaeophyceae). *J. Phycol.*, 31: 306-311.

Kawai, H., Nakayama, T., Inouye, I. and Kato, A. 1997. Linkage of 5S ribosomal DNA to other rDNAs in the chromophytic algae and related taxa. *J. phycol.*, 33: 505-511.

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

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

Katiyar, S.K., Govinda, S.V. and Edlind, T.D. 1995. Comparisons of ribosomal RNA sequences from amitochondrial protozoa: implications for processing, mRNA binding and paromomycin susceptibility. *Gene,* 152: 27-33.

Kawakami, Y., Inoue, T., Ito, K., Kitamizu, K., Hanawa, C., Ando, T., Iwano, H. and Ishihara, R. 1994. Identification of a chromosome harboring the small-subunit ribosomal-RNA gene of *Nosema bombycis*. *J. Invertbr. Pathol.,* 64: 147-148.

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

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

Keller, M., Tessier, L.H., Chan, R.L., Weil, J.H. and Imbault, P. 1992. In *Euglena*, spliced-leader RNA (SL-RNA) and 5S rRNA genes are tendemly repeated. *Nucl. Acids Res.*, 20: 1711-1715.

Kibe, M.K., Ole-MoiYoi, O.K., Nene, V., Khan, B., Allsopp, B.A., Collins, N.E., Morzaria, S.P., Gobright, E.I. and Bishop, R.P. 1994. Evidence for 2 single-copy units in *Theileria parva* ribosomal-RNA genes. *Mol. Biochem. Parasitol.*, 66: 249-259.

Kishino, H. and Hasegawa, M. 1990. Converting distance to time: Application to human evolution. *Meth. Enzymol.,* 183: 550-557.

Kooi, E.A., Rutgers, C.A., Mulder, A., Van't Riet, J., Venema, J. and Raué, H.A. 1993. The phylogenetically conserved doublet tertiary interaction in domain III of the large subunit rRNA is crucial for ribosomal protein binding. *Proc. Natl. Acad. Sci. USA,* 90: 213-216.

Kuhner, M.K. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.,* 11: 459-468.

Lake, J.A. 1987. Rate-independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Biol. Evol.,* 4: 167-191.

Larsen, N. 1992. Higher order interactions in 23S rRNA. *Proc. Natl. Acad. Sci. USA,* 89: 5044-5048.

Larsson, R. 1986. Ultrastructure, function and classification of microsporidia. *Progr. Protistol.,* 1: 325-390.

Leffers, H., Kjems, J., Østergaard, L., Larsen, N. and Garrett, R.A. 1987. Evolutionary relationships among archaebacteria. A comparative study of 23S rRNAs of a sulphur-dependent thermophile, an extreme halophile and a thermophilic methanogen. *J. Mol. Biol.,* 195: 43-61.

Leipe, D.D., Gunderson, J.H., Nerad, T.A. and Sogin, M.L. 1993. Small subunit ribosomal RNA$^{+}$ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. *Mol. Biochem. Parasitol.,* 59: 41-48.

Levaditi, C., Nicolau, S., and Schoen, R. 1923. L'agent étiologique de l'encéphalite épizootique du lapin (*Encephalitozoon cuniculi*). *C. R. Soc. Biol.* Paris 89: 984-986. [Cited from Weber *et al.,* 1994]

Linnaeus, C. (1758). *Systema Nature.* 10th edn. Holmiae Laur Salvii. [ Cited from Bulla and Cheng (1977).]

Lockwood, J.A. and Debrey, L.D. 1990. Direct and indirect effects of a large scale application of *Nosema locustae* (Microsporidia: Nosematidae) on rangeland grasshoppers (Orthoptera: Acrididae). *Entomol. Soc. Am.,* 83: 377-383.

Long, E.O. and Dawid, I.B. 1980. Repeated genes in eukaryotes. *Annu. Rev. Biochem.,* 49: 727-764.

Loomis, W.F. and Smith, D.W. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. USA,* 87: 9093-9097.

Malone, L.A. and McIvor C.A., 1993. Pulsed-field gel electrophoresis from four microsporidian isolates. *J. Invertebr. Pathol.*, 56: 243-248.

Malone, L.A., Broadwell, A.H., Lindridge, E.T., McIvor, C.A. and Ninham, J.A. 1994. Ribosomal RNA genes of two microsporidia, *Nosema apis* and *Vavraia oncoperae,* are very variable. *J. Invertbr. Pathol.,* 64: 151-152.

Malone, L.A. and McIvor C.A., 1996. Use of nucleotide sequence data to identify a microsporidian pathogen of *Pieris rapae* (Lepidoptera, Pieridae). *J. Invertebr. Pathol.*, 68: 231-238.

Margulis, L. 1974a. Five-kingdom classification and the origin and evolution of cells. *Evol. Biol.*, 7: 45-78.

Margulis, L. 1974b. The classification and evolution of prokaryotes and eukaryotes. In R.C. King (ed.). *Handbook of Genetics.* Vol. 1, pp1-4. New York. Plenum Press.

Margulis, L. 1980. Undulipodia, flagella and cilia. *Biosystems,* 12: 105-108.

Margulis, L. 1989. Introduction. In L. Margulis, J.O. Corliss, M. Melkonian and D.J. Chapman (eds.). *Handbook of Protoctista.* Jones and Bartlett Publishers, Boston, USA.

Marquardt, W.C. and Demaree, R.S., *Parasitology.* MacMillan, NewYork.

Maslov, D.A., Elgort, M.G., Wong, S., Peckova, H., Lom, J., Simpson, L. and Campbell, D.A. 1993. Organization of mini-exon and 5S rRNA genes in the kinetoplastid *Trypanoplasma borreli. Mol. Biochem. Parasitol.,* 61: 127-136.

Matheson, A. 1993a. World bee health report. *Bee World,* 74: 176-212.

Matheson, A. 1993b. Practical Beekeeping in New Zealand. GP Publications Ltd, Wellington, New Zealand.

Matsubayashi, H., Koike, T., Mikata, T. and Hagiwara, S. 1959. A case of *Encephalitozoon*-like body infection in man. *Arch. Pathol.*, 67: 181-187.

McCutchan, T.F., Li. J., McConkey, G.A., Rogers, M.J. and Waters, A.P. 1995. The cytoplasmic ribosomal RNAs of *Plasmodium spp. Parasitol. Today,* 11: 134-143.

Michot, B., Hassouna, N. and Bachellerie, J.-P. 1984. Secondary structure of mouse 28S rRNA and general model for the folding of the large rRNA in eukaryotes. *Nucl. Acids Res.,* 12: 4259-4279.

Michot, B. and Bachellerie, J-P. 1987. Comparisons of large subunit rRNAs reveal some eukaryote-specific elements of secondary structure. *Biochimie,* 69: 11-23.

Michot, B., Qu, L., Bachellerie, J.P. 1990. Evolution of large-subunt rRNA structure: The diversification of divergent D3 domain among major phylogenetic groups. *Eur. J. Biochem.,* 188: 219-229.

Mooers, A.O., Page, R.D.M., Purvis, A. and Harvey, P.H. 1995. Phylogenetic noise leads to unbalanced cladistic tree reconstructions. *Syst.Biol.,* 44: 332-342.

Morton, A., Tabrett, A.M., Carder, J.H. and Barbara, D.J. 1995. Sub-repeat sequences in the ribosomal RNA intergenic regions of *Verticillium alboatrum* and *V. dahliae*. *Mycol. Res.,* 99: 257-266.

Moritz, C. and Hillis, D.M. 1996. Molecular systematics: context and controversies. In David M. Hillis, Craig Moritz, Barbara K. Mable (eds.). *Molecular Systematics* (2$^{nd}$ Edition). pp. 1-13. Sinauer Associates, Inc. Sunderland, Massachusetts USA.

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

Mullis, K.B. and Faloona, F.A. 1987. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Meth. Enzymol.,* 155: 335-350.

National Centre for Biotechnology Information, National Library of Mdeicine, 8600 Rockville Pike, Bethesda, MD, USA. http://www.ncbi.nlm.nih.gov/

Neefs, J., Van de Peer, Y., De Rijk, P., Goris, A. and De Wachter, R. (1991). Compilation of small ribosomal subunit RNA sequences. *Nucl. Acids Res.,* 19(supplement): 1987-2015.

Nei, M. 1991. Relative efficiencies of different tree-making methods for molecular data. In M.M. Miyamota and J. Cracraft (eds.). *Phylogenetic analysis of DNA sequences.* pp. 90-128. Oxford University Press, Oxford, England.

Noller, H.F. 1991. Ribosomal RNA and translation. *Annu. Rev. Biochem.,* 60: 191-227.

Noller, H.F., Kop, J., Wheaton, V., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F., Herr, W., Stahl, D.A., Gupta, R. and Woese, C.R. 1981. Secondary structure model for 23S ribosomal RNA. *Nucl. Acids Res.,* 9: 6167-6189.

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

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

Olsen, G.J. 1988. Phylogenetic analysis and ribosomal RNA. *Meth. Enzymol.,* 164: 793-812.

Pakes, S.P., Shadduck, J.A. and Cali, A. 1975. Fine structure of *Encephalitozoon cuniculi* from rabbits, mice and hamsters. *J. Protozool.,* 22: 481-488.

Peyretaillade, E., Biderre, C., Peyret, P., Duffieux, F., Méténier, G., Gouy, M., Michot, B. and Vivarés, C.P. 1998. Microsporidia *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.

Philippe, H. and Adoutte, A. 1998. The molecular phylogeny of Eukaryota: solid facts and uncertainties. In G.H. Coombs, K. Vickerman, M.A. Sleigh and A. Warren (eds.). *Evolutionary Relationships Among Protozoa.* pp. 25-56. Academic Publishers, London.

Pilley, B.M. 1976. A new genus, *Vairimorpha* (Protozoa: Microsporidia), for *Nosema necatrix* Kramer, 1965: pathogenicity and life cycle in *Spodoptera exempta (Lepidoptera: Noctuidae). J. Invertebr. Pathol.,* 28: 177-183.

Pomport-Castillon, C., Romestand, B. and De Jonckheere, J. 1997. Identification and phylogenetic relationships of microsporidia by riboprinting. *J. Euk. Microbiol.,* 44: 540-544.

Raué, H.A., Klootwuk, J. and Musters, W. 1988. Evolutionary conservation of structure and function of high molecular weight ribosomal RNA. *Prog. Biophys. molec. Biol.,* 51: 77-129.

Raué, H.A., Musters, W., Rutgers, C.A., Van't Riet, J. and Planta, R.J. 1990. rRNA: from Structure to Function. In W.E. Hill, A.E. Dahlberg, R.A. Garrett, P.B. Moore, D. Schlessinger and J.R. Warner (eds.). *Ribosome, Structure, Function and Evolution.* pp. 217-235. Am. Soc. Microbiol., Washington.

Reeder, R.H. 1990. rRNA synthesis in the nucleolus. *Trends In Genetics,* 6: 390 - 395.

Reik, E.F. 1979. Hymenoptera. In: *The Insects of Australia.* Melbourne University Press.

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

Roger, A.J., Svärd, S.G., Tovar, J., Clark, G., Smith, M.W., Gillin, F.D. and Sogin, M. 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. USA,* 95: 229-234.

Rosendahl, G., Hansen, L.H. and Douthwaite, S. 1995. Pseudoknot in domain II of 23S rRNA is essential for ribosome function. *J. Mol. Biol.,* 249: 59-68.

Ruttner, F. and Maul, V. 1983. Experimental analysis of the reproductive interspecific isolation of *Apis mellifera* L. and *Apis cerana* Fabr. *Apidologie,* 14: 309-327.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N., (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science,* 230: 1350-1354.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science,* 239: 487-491.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.,* 4: 406-425.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular Cloning. A laboratory Manual.* Second edition. Cold Spring Harbour Press. Cold Spring Harbour, New York. USA.

Santor, M. 1963. Ribosomal RNA on the surface of the ribosome. *Science*, 141: 1049-1050.

Sato, R. and Watanabe, H. 1980. Purification of mature microsporidian spores by iso-density equilibrium centrifugation. *J. Seric. Sci. Japan,* (in Japanese with English summary). 49: 512-516.

Schaal, B.A. and Learn, G.H. 1988. Ribosomal DNA variation within and among populations. *Ann. Missouri. Bot. Gard.,* 75: 1207-1216.

Schnare, M.N. and Gray, M.W. 1990. Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. *J. Mol. Biol.,* 215: 85-91.

Schnare, M.N., Damberger, S.H., Gray, M.W. and Gutell, R.R. 1996. Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. *J. Mol. Biol.,* 256: 701-719.

Siddall, M.E., Hong, H. and Desser, S.S. 1992. Phylogenetic analysis of the Diplomonadida (Wenyon, 1926) Brugerolle, 1975: evidence for heterochrony in Protozoa and against *Giardia lamblia* as a "missing link." *J. Protozool.,* 39: 361-367.

Singh, Y. 1975. Nosema in indian honey bee (*Apis cerana* indica). *Amer. Bee J.,* 115: 59.

Smith, C. L., Econome, J. G., Schutt, A., Klco, S. and Cantor, C. R., 1987. A physical map of the *Escherichia coli* K12 genome. *Science*, 236: 1448-1453.

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

Sogin, M.L. and Silberman, J.D. 1998. Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Internat. J. Parasitol.*, 28: 11-20.

Soltys, B.J. and Gupta, R.S. 1994. Presence and cellular distribution of a 60-hDa protein related to mitochondrial HSP 60 in *Giardia lamblia. J. Parasitol.,* 80: 580-590.

Sprague, V., Becnel, J.J. and Hazard, E.I. 1992. Taxonomy of Phylum Microspora. *Crit. Rev. Microbiol.,* 18: 285-395.

Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA,* 85: 2653-2657.

Sweeney, A.W., Hazard, E.I. and Graham, M.F. 1985. Intermediate host for an *Amblyospora sp*. (Microspora) infecting the mosquito, *Culex annulirostris. J. Invertebr. Pathol.,* 46: 98-102.

Swofford, D.L. 1985. PAUP: *Phylogenetic Analysis Using Parsimony, version 2.4.* Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. 1996. Phylogenetic Inference. In David M. Hillis, Craig Moritz, Barbara K. Mable (eds.). *Molecular Systematics* (2$^{nd}$ Edition). pp. 407-514. Sinauer Associates, Inc. Sunderland, Massachusetts USA.

Tanada, Y. and Kaya, H.S. 1993. *Insect pathology*, Academic Press, Boston. U.S.A.

Teakle, R.E. and Jensen, J.M. 1985. *Heliothis punctigera.* In P. Singh and R.F. More (eds.). *Handbook of Insect Rearing.* Vol. 2 pp. 312-322., Elsevier, Amsterdam.

Tourasse, N.J. and Gouy, M. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. *Mol. Biol. Evol.,* 14: 287-298.

Uliana, S.R.B., Fischer, W., Stempliuk, V.A. and Floeter-Winter, L.M. 1997. Structural and functional characterization of the *Leishmania amazonensis* ribosomal RNA promoter. *Mol. Biochem. Parasitol.,* 76: 245-255.

Van De Peer, Y., J. -M. Neefs, P. De Rijk and R. De Wachter. 1993. Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. *J. Mol. Evol.,* 37: 221-232.

Van De Peer, Y., Caers, A., De Rijk, P., De Wachter, R. 1998. Database on the structure of small ribosomal subunit RNA. *Nulc. Acids Res.,* 26: 179-182.

Van Keulen, H., Gutell, R.R., Campbell, S.R., Erlandsen, S.L. and Jarroll, E.L. 1992. The nucleotide sequence of the entire ribosomal DNA operon and the structure of the large subunit rRNA of *Giardia muris. J. Mol. Evol.,* 35: 318 - 328.

Van Keulen, H., Campbell, S.R., Erlandsen, S.L. and Jarroll, E.L. 1991. Cloning and restriction enzyme mapping of ribosomal DNA of *Giardia duodenalis, Giardia ardeae,* and *Giardia muris. Mol. Biochem. Parasitol.,* 46: 275-284.

Veldmann, G.M., Klootwijk, J., de Regt, V.C.H.F., Planta, R.J., Branlant, C., Krol, A. and Ebel, J.-P. 1981. The primary and secondary structure of yeast 26S rRNA. *Nucl. Acids Res.* 9: 6935-6952.

Vivarès, C., Biderre, C., Duffieux, F., Peyretaillade, E., Peyret, P., Méténier, G. and Pagès, M. 1996. Chromosomal localization of five genes in *Encephalitozoon cuniculi* {Microsporidia). *J. Euk.. Microbiol.,* 43: 97S.

Vossbrinck, C.R. and Woese, C.R. 1986. Eukaryotic ribosomes that lack a 5.8S RNA. *Nature*, 320: 287-288.

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

Vossbrinck,C.R., Baker,M.D., Didier,E.S., DeBrunner-Vossbrinck,B.A. and Shadduck,J.A. 1993. Ribosomal DNA sequences of *Encephalitozoon hellem* and *Encephalitozoon cuniculi*: species identification and phylogenetic construction. *J. Euk.. Microbiol.,* 40, 354-362.

Warhurst, P and Goebel, R. 1995. The Bee Book: beekeeping in the warmer areas of Australia. Information Series Q194060, Department of Primary Industries, Queensland, Australia.

Weber, R., Bryan, R.T., Schwartz, D.A. and Owen, R.L. 1994. Human microsporidial infections. *Clin. Microbio. Rev.,* 7: 426-461.

Weiser, J. 1957. Mikrosporidien des Schwammspinners und Goldafters. *Z. Angew. Entomol.,* 40: 509-527. [ Cited from Bulla and Cheng (1977)]

Woese, C.R. and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA*, 74: 5088-5090.

Woese, C.R. 1980. Just so stories and Rube Goldberg machines: speculations on the origin of the protein synthetic machinery. In G. Chambliss, G.R. Craven, J. Davies, K. Davis, L. Kahan and M. Nomura (eds.). *Ribosomes. Structure, Function, and Genetics*. pp. 357-373. University Park Press, Baltimore.

Wolters J and Erdmann, V.A. 1986. Clasdistic analysis of 5S rRNA and 16S rRNA secondary and primary structures—The evolution of eukaryotes and their relation to archaebacteria. *J. Mol. Evol.,* 24: 152-166.

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

Wright, J.H. and Craighead, E.M. 1992. Infectious motor paralysis in young rabbits. *J. Exp. Med.,* 36: 135-140.

Yamamoto, A., Hashimoto, T., Asaga, E., Hasegawa, M. and Goto, N. 1997. Phylogenetic position of the mitochondrion-lacking protozoan *Trichomonas tenax*, based on amino acid sequences of elongation factors 1alpha and 2. *J. Mol. Evol.,* 44: 98-105.

Yakobson, B., Pothichot, S. and Wongsiri, S. 1992. Possible transfer of *Nosema apis* from *Apis mellifera* to *Apis cerana.* In: *Abstracts of papers of Int. Conf. Asian honey bees and bee mites,* Bangkok, 9-14 February, 1992, p. 97. Bee Biol. Res. Unit, Dept. Biology, Chulalong Univ., Bangkok, Thaliand.

Zander, E. (1909). Tierische Parasiten als Krankheitserreger bei der Bein. *Leipzig. Bienenztg.* 147-150, 164-166. [ Cited from Bulla and Cheng (1977)]

Zhu, X., Wittner, M., Tanowitz, H.B., Cali, A. and Weiss. L.M. 1993. Nucleotide sequence of the small ribosomal RNA of *Encephalitozoon cuniculi. Nucl. Acids Res.,* 21: 1315.