EVOLUTIONARY STUDIES BASED ON THE 70-kDa HEAT SHOCK FAMILY OF PROTEIN SEQUENCES

By

MIZIED FALAH, B.Sc., M.Sc.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Doctor of Philosophy

McMaster University

EVOLUTIONARY STUDIES BASED ON THE 70-kDa HEAT SHOCK FAMILY OF PROTEIN SEQUENCES

Doctor of Philosophy (1997)
(Biochemistry)

McMaster University
Hamilton, Ontario, Canada

TITLE: Evolutionary Studies Based on the 70-kDa Heat Shock Family of Protein Sequences

AUTHOR: Mizied Falah, B.Sc., M.Sc. (The Hebrew University of Jerusalem)

SUPERVISOR: Dr. Radhey S. Gupta

NUMBER OF PAGES: [xii], [95]

ABSTRACT

The 70-kDa heat shock protein (HSP70) is one of the most conserved proteins known to date. The HSP70 genes from *Mycoplasma capricolum*, *Giardia lamblia*, *Rhizobium meliloti*, and *Pseudomonas cepacia* were cloned and sequenced. Sequence comparison of HSP70 shows that homologs from gram-positive bacteria and archaebacteria share a number of sequence signatures that are not found in other prokaryotic or eukaryotic species. Other sequence features were found to be specific to eukaryotes and gram-negative bacteria. Phylogenetic analyses of HSP70 sequence strongly support a specific evolutionary relationship between gram-positive bacteria and archaebacteria on the one hand, and gram-negative bacteria and eukaryotes on the other. The mycoplasma homologs of HSP70 contain all sequence signatures that are characteristic of gram-positive bacteria and archaebacteria. In the HSP70 tree, mycoplasma branch with the low-G+C-content gram-positive group of bacteria indicating their close evolutionary relationship to this group. The *G. lamblia* HSP70 homologs (from the cytosol and endoplasmic reticulum), similar to other eukaryotic homologs, were found to contain sequence signatures specific to gram-negative species. Phylogenetic analyses show that *G. lamblia* homologs are the deepest homologs in the eukaryotic tree. The cytosolic and endoplasmic reticulum homologs in eukaryotes, including those of *G. lamblia*, are similar and form two distinct subfamilies indicating that they evolved from a common eukaryotic ancestor by gene duplication early in the evolution of eukaryotic cells. The early divergence of *G. lamblia* and its evolutionary linkage to gram-

negative bacteria indicate that eukaryotes evolved from a gram-negative bacterium. The cloning of the endoplasmic reticulum-HSP70 gene from *G. lamblia* suggested the existence of endoplasmic reticulum in this ancient eukaryote. The existence of endoplasmic reticulum in *G. lamblia* has been investigated using the endoplasmic reticulum-HSP70 protein as a molecular marker. The endoplasmic reticulum-HSP70 gene of *G. lamblia* was expressed in *Escherichia coli* and a polyclonal antibody was raised to the purified protein. Using cryoelectron microscopy and antibody to endoplasmic reticulum-HSP70, we demonstrated the presence of endoplasmic reticulum in *G. lamblia*. The presence of endoplasmic reticulum in *G. lamblia*, which is one of the earliest diverging eukaryotic species lacking mitochondria, supports the view that both the nucleus and endoplasmic reticulum co-evolved in a common ancestor of eukaryotic cells. We have used the HSP70 sequences from *Rhizobium meliloti*, and *Pseudomonas cepacia* to study the evolution of mitochondria. *Rhizobium meliloti*, and *Pseudomonas cepacia* belong to the α- and β-subdivisions of gram-negative proteobacteria, respectively. On the basis of sequence alignment, several sequence signatures have been identified that are distinctive to mitochondrial homologs and gram-negative proteobacteria. Phylogenetic analyses of HSP70 sequences suggest the endosymbiotic origin of mitochondria from a member of the α-proteobacteria and that all mitochondrial homologs are monophyletic in origin.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Radhey S. Gupta for his advice and guidance throughout the course of this work and for his countless efforts to provide broad training in both educational and professional aspects. Without his continuous encouragement and criticism, the completion of this work would not have been possible. Similarly, I would like to extend my gratitude to the members of my Supervisory Committee, Professors K. B. Freeman, G. B. Golding, and G. Singh for their critical discussion and important suggestions on this work.

Special thanks are also extended to all my teachers and co-workers who shared in my scientific endeavors and helped me greatly. Dr. B. Singh was very helpful throughout, providing technical assistance and always being available for scientific discussion. Dr. B. J. Soltys helped in conducting immunofluorescence experiments and always presented me with an invaluable perspective. Mrs. Rajni Gupta provided help in the protein purification. J. Cechetto helped in printing the pictures of this thesis. I am also very thankful for my fellow graduate students at McMaster University who contributed to my graduate student life.

I would like to thank my parents and my wife Dalia for their loving support and encouragement during my studies at McMaster University.

Finally, I dedicate this work with love to my son Adam.

TABLE OF CONTENTS

List of Figures and Tables

List of Abbreviations

EF	elengation factor
ER	endoplasmic reticulum
FGARAT	formylglycineamidine ribotide amidotransferase
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GDH	glutamate dehydrogenase
GSI	glutamine synthetase I
GSII	glutamine synthetase II
GSIII	glutamine synthetase III
HSP60	60 kDa heat shock protein
HSP70	70 kDa heat shock protein
LSU rRNA	large subunit rRNA
SSU rRNA	small subunit rRNA
Tcp-1	T-complex polypeptide 1

Preface to Thesis

This thesis is written in manuscript format with each chapter consisting of a separate manuscript (except Chapter 1 and 6 which are Background, and Conclusions and Future Directions, respectively). The references for Chapters 1 and 6 are combined and presented at the end of the thesis. Figure numbering for Chapters 1 and 6 are also continuous.

CHAPTER 1

BACKGROUND

Preface and Rationale

For a long time, the study of evolution was mainly based on morphological characteristics and thus it was restricted to plants and animals. But morphological criteria are inadequate when applied to studying the evolution of microorganisms as they exhibit simple morphology which is relatively uninformative phylogenetically. Advances in molecular biology, such as PCR and sequencing techniques, have enhanced our understanding of evolution. Molecular characteristics such as DNA and protein sequences are now mainly utilized to study evolution. Sequences of protein and DNA contain a vast number of characters which are highly informative for comparative phylogenetic studies of all extant species including microorganisms. DNA or protein molecules and their counterparts in all species can be compared even if distantly related. The shared molecular characters between sequences from various species reflect evolutionary relationship. Sequences of homologous genes or their corresponding proteins or RNAs are therefore useful models for evolutionary studies. The evolutionary origin as well as the evolutionary relationship among species can be inferred by comparing these sequences from various species. In this thesis the 70-kDa heat shock protein (HSP70) sequence has been utilized to investigate the evolution of prokaryotes, eukaryotic organelles, and early branching eukaryotes. HSP70 protein was chosen for its high degree of sequence conservation.

General Introduction

Understanding the evolutionary relationships among organisms is one of the major challenges in biological sciences. The aim is to trace the evolutionary pathways from the earliest organisms of about four billion years ago, to the great biodiversity we see today. The early studies of evolution were restricted to the classification of plants and animals based on their morphological characteristics (Cain, 1962; Priest and Austin, 1993). In the case of bacteria however the problem is different because of the small size of these organisms and lack of pronounced morphological variation. In 1684 Antonie van Leeuwenhock made the first illustrations of bacteria that suggested morphological variations amongst them (Priest and Austin, 1993). However, despite many years of work, a proper and generally acceptable classification of bacteria has not yet been developed. Much of the earlier work in microbiology was done on the physiology and pathogenicity of bacteria. Bacteria living in extreme conditions, which have little to do with pathogenic infection, were ignored. For comprehensive investigation into the question of evolution it is imperative to study all existing organisms. Further, bacteria hold a special phylogenetic position in the evolution of all extant species. Chloroplasts and mitochondria, the organelles of eukaryotic cells that harness the sunlight and utilize oxygen as an energy source, respectively, are of bacterial origin (Gray and Doolittle, 1982; Margulis, 1970). Knowing the evolution of bacteria provides insights into the evolution of eukaryotic cells and organelles. Furthermore, studying the evolution of bacteria that thrive in extreme conditions (i.e. high temperature and reducing anoxic atmosphere), similar to those of the early earth, is important for understanding the nature of the universal ancestor that gave rise to contemporary species.

The differences in cellular organization have led to the classification of all organisms into two primary groups: prokaryotes and eukaryotes. The evolutionary relationships within prokaryotes and

between the prokaryotes and eukaryotes could not be adequately resolved based on morphological data (Cain, 1962; Priest and Austin, 1993; Woese, 1991, 1992). The possibility of determining such relationships and constructing a phylogeny of organisms including bacteria became realistic with advances in sequencing methodologies. Sequences of genes and/or their products, RNAs and proteins, provide phylogenetic information so the study of evolution has shifted from the comparative studies based on morphological criteria to sequence data. An important development in this regard was the realization that organisms have in their genome sequence records of the changes that have occurred since the divergence from their common ancestor (Zuckerkandl and Pauling, 1965). The extent and nature of the differences among sequences provide insights into the phylogenetic relationships of all organisms. Consider for example comparing two protein sequences of 100 amino acids or their genes of 300 base pairs. The probability of having complete sequence identity is very small: $(1/20)^{100}$ for the proteins, and $(1/4)^{300}$ for the genes. For these molecules to have only 30% sequence identity, the probability value is still very low. However, many proteins/genes from different organisms show a high degree of identity (>30%) to each other. Because such identity is unlikely to occur by chance, it must be due to evolutionary relatedness (i.e. common ancestry) of these proteins/genes. The four character states (nucleotides) in genes or the 20 character states (amino acids) in proteins are discrete, and amenable to computer and numerical processing and therefore would directly and quantitatively reflect the evolutionary relationships among species. Molecular sequence data is currently the most widely used characteristic to construct phylogenetic trees.

Assessing the similarity among sequences involves three independent operations. First, establishing the optimal alignment of the investigated sequences. Second, choosing a scoring system to calculate similarities or identities among sequences. Third, determining the statistical significance

of the score. The alignment of sequences places characters (amino acids or nucleotides) in positional homology and therefore requires inserting gaps, which correspond to insertions or deletions, into one or more of the aligned sequences to place positions inferred to be homologous into the same column of the alignment. The best possible alignment between two sequences is the one in which the number of mismatches and gaps are minimum. The sequence alignments form the data for phylogeny reconstruction. Two of the most common approaches to phylogeny reconstruction are parsimony methods and distance matrix methods (Felsenstein, 1988). These methods interpret the sequence alignment in different ways and can therefore often give different phylogenies. A maximum parsimony method infers the number of evolutionary events implied by a particular topology and chooses a tree that requires the minimum number of these evolutionary events (Li and Graur, 1991). In contrast, distance methods use pairwise measures of differences to construct a phylogenetic tree (Li and Graur, 1991). One commonly used distance method is the Neighbour Joining method (Saitou and Nei, 1987) which constructs the tree by linking the least distant pairs of sequences together. When two species are clustered together they form a new taxon and lose their individual identities. The new distances are then calculated from the original matrix to correspond to distances between all remaining species and the newly formed taxon.

Since the evolution of the investigated gene is influenced by unknown evolutionary forces it is difficult to determine which method will work best for a particular gene sequence. Therefore it is necessary to estimate the reliability of the results of these methods. The bootstrap test estimates the confidence of the inferred phylogenetic relationships (Felsenstein, 1988). In the bootstrap test, columns in the aligned sequences are sampled randomly and replaced with columns from the original alignment to obtain new alignments containing the original number of columns. Thus, some columns

will not be included at all in a given bootstrap replication, others will be included once, and still others twice or more. If the sample data are in fact representative of the underlying population, the confidence interval associated with them will be reflected in the number of bootstrap replicons (i.e. bootstrap scores) which cluster a particular group of species. The inferred phylogeny or tree based on a given gene or protein sequence is only an approximation of the true phylogenetic tree. However, an agreement between the results of different phylogenetic methods would increase confidence in the inferred tree. The confidence in the inferred gene tree is also dependent on whether the investigated gene sequences are orthologous (i.e. diverged through speciation) or some have been acquired from different species by gene transfer (i.e. xenologous). Phylogenetic analysis must be based on orthologous sequences. A phylogenetic tree based on one particular gene or its protein sequence precisely means a tree of that particular gene, not the tree of the species from which this gene was isolated. Individual gene phylogeny will not necessarily represent the species phylogeny (Nei, 1987), because homologous genes in different populations generally diverge much earlier than population splitting (Takahata, 1989). However, the species phylogeny can be better estimated based on sequence data from many genes as well as including morphological, physiological and biochemical features of species.

The Use of Molecular Sequence in Evolutionary Studies

Significant similarity among DNA, RNA or protein sequences reflects phylogenetic relatedness. High similarity implies recent divergence from a common ancestor and low similarity implies early divergence from a common ancestor. The first sequences to be used in molecular phylogeny were those of cytochrome *c* and globins (Dayhoff *et al.*, 1972; Schwartz and Dayhoff, 1978). However, due to difficulty in protein sequencing, only limited inferences could be drawn for

evolutionary studies. Today, the ease with which DNA can be sequenced has led to DNA sequences accounting for the bulk of the data for phylogeny reconstruction. Very often the DNA sequences are translated by computer programs into protein sequences before analyses. This reduces bias in phylogenetic reconstruction stemming from variations in G+C content amongst different species. In proteins, each character has 20 states instead of the 4 in nucleic acid sequences. This provides more information per site in proteins and reduces the likelihood of mere chance alignment or errors due to multiple substitutions at a site. Also, by contrast to structural nucleic acids [viz. small subunit ribosomal RNA (SSU rRNA)], compositional changes in protein coding sequences occur predominantly in the third-codon positions, which because of the degeneracy of the genetic code has a minimal effect on the encoded amino acid sequence. Nevertheless, large number of molecular phylogenies have been based on the SSU rRNA (Leipe *et al.*, 1993; Olsen *et al.*, 1994; Sogin *et al.*, 1996; Woese, 1987, 1991). In the rRNA sequence analysis, conservation of the secondary structure of the molecule is utilized to obtain an alignment of the primary sequences. This means that the sequence alignment is carried out based on the assumptions that similar secondary structure constraints apply in very distantly related species. This assumption can introduce bias in the alignment which may lead to an incorrect phylogenetic relationship for certain species (Hasegawa *et al.*, 1993). Since such assumptions concerning secondary structure are not necessary for alignment of protein sequences, proteins provide a better tool for evolutionary studies. Furthermore, the size of the SSU rRNA is not consistent in all groups of organisms. Eukaryotes possess 18S rRNA whereas prokaryotes have 16S rRNA.

For a protein to be a good model for global evolutionary studies it should have the following characteristics: (i) be present in all forms of life; (ii) be sufficiently large in size to possess adequate

information to allow statistically significant conclusions; (iii) have a high degree of conservation of function and primary structure so that the corresponding positions from different species can be aligned and compared with minimal ambiguity; (iv) the orthologous, paralogous (i.e. diverged after gene duplication), and xenologous sequences should be clearly distinguished when more than one homolog is present in a species; (v) have distinctive sequence signatures that are useful in identifying specific groups of species as well as in establishing the relationship among them.

The Current Universal Tree of Life

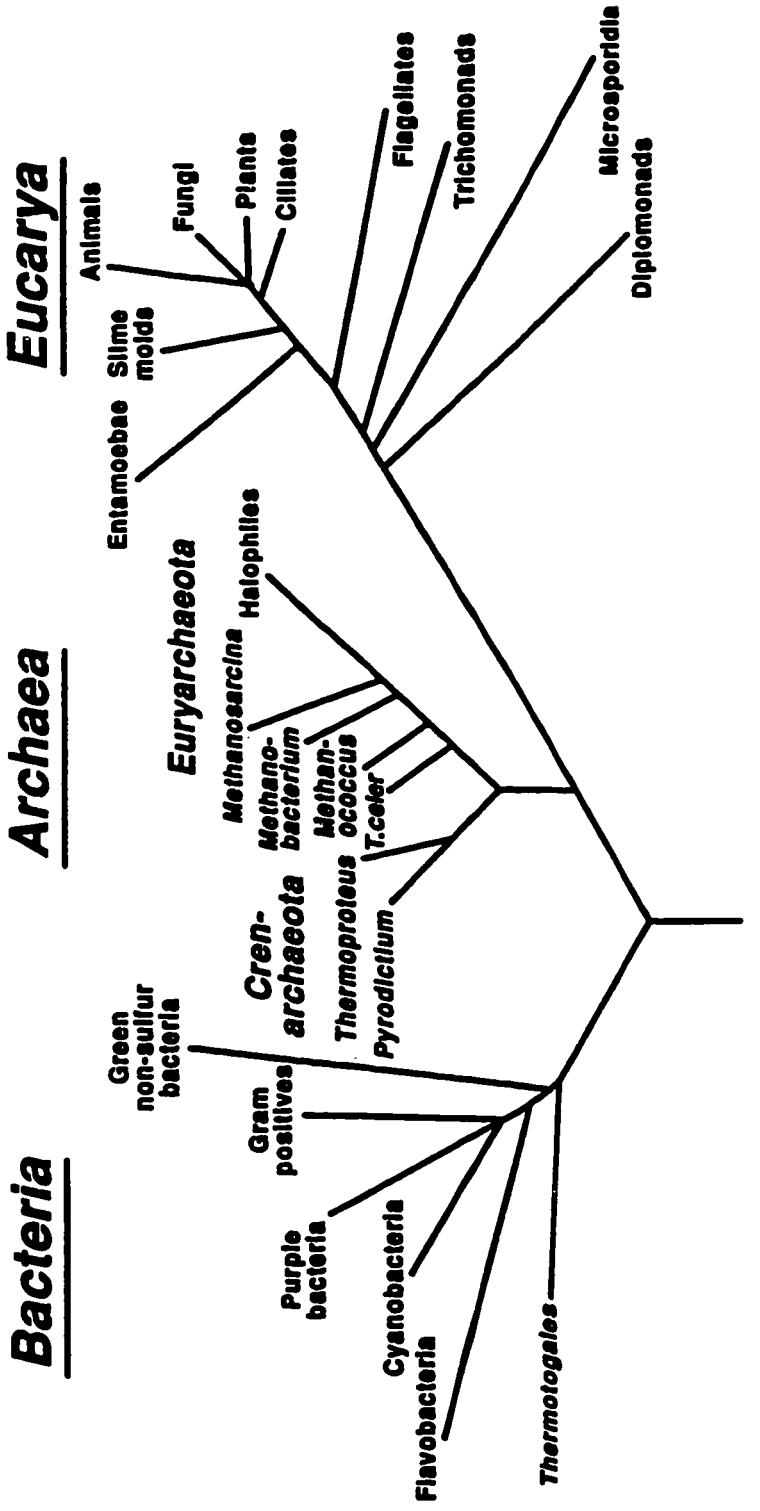
The earliest and most extensive molecular studies to understand the evolutionary relationship among all extant species have been carried out based on the sequence of SSU rRNA (Fox *et al.*, 1980; Leipe *et al.*, 1993; Olsen and Woese, 1993; Olsen *et al.*, 1994; Sogin *et al.*, 1996; Woese, 1987, 1991). These studies revealed the diversity of the prokaryotic and eukaryotic species and led to the construction of a universal tree of life based on sequence data. All organisms were put into three primary groups or domains: *Archaea* (Archaebacteria), *Bacteria* (Eubacteria), and *Eucarya* (Eukaryotes). In this universal tree of life, all eukaryotic species were proposed to belong to one distinct monophyletic domain. The prokaryotes, on the other hand were divided into two distinct monophyletic domains (Woese, 1987; Woese *et al.*, 1990).

Initially the universal tree of life was unrooted. That is, the earliest species or domain was not determined and therefore the evolutionary relationships among the three primary domains was not known. Phylogenetic trees are normally rooted by comparisons to distant outgroups. For example, the root of eukaryotic phylogeny can be inferred by assigning bacterial species as an outgroup. However, in the universal tree which includes all organisms no outgroups could be assigned. This difficulty was overcome later by comparing duplicated gene products which are

protein sequences related by a duplication that preceded the speciation of organisms. Thus, the evolutionary relationship of the three domains became clear and the ancestors could be postulated. Two independent studies have suggested the likely root of this universal tree by comparing pairs of protein sequences for EF-Tu(1α) and EF-G(2) (Iwabe *et al.*, 1989) as well as for α and β subunits of ATPase (Gogarten *et al.*, 1989). Because these pairs of gene duplicates are present in all three primary domains, it is assumed that their duplication occurred prior to the initial divergence of the domains from their common ancestor. Two subtrees were constructed, one for each of the gene duplicates, and subsequently one subtree was used as an outgroup for the other. The root of the inferred composite tree was then set at a point where the two genes diverged by gene duplication. By using this approach, the root of the universal tree was positioned between eubacteria on one hand and archaebacteria and eukaryotes on the other. A similar rooting of the universal tree has also been obtained based on sequence data for aminoacyl-tRNA synthetases (Brown and Doolittle, 1995). In the rooted form of the universal tree, eukaryotes are proposed to share a common evolutionary history specifically with archaebacteria, exclusive of eubacteria (see Figure 1). If true, this means that archaebacteria are specific relatives of the eukaryotes. The above tree has been widely accepted; however subsequent studies based on other protein sequences have presented a very different picture.

Figure 1. A universal tree based on SSU rRNA sequence and the rooting inferred from ancestral gene duplication.

This figure is adapted from Woese *et al.*, 1990.


Bacteria
Archaea
Eucarya
Animals
Fungi
Plants
Ciliates
Slime molds
Entamoebae
Flagellates
Trichomonads
Microsporidia
Diplomonads
Euryarchaeota
Halophiles
Methanosarcina
Methano-bacterium
Methan-ococcus
T.celer
Cren-archaeota
Thermoproteus
Pyrodictium
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales

Evolutionary Studies Based on HSP70

Recent studies in our laboratory, based on HSP70 sequences challenge the currently favoured view concerning the evolution of prokaryotes and eukaryotes. HSP70 performs an essential molecular chaperone function by binding to partially unfolded proteins and assisting both in their proper folding as well as in their transport across various intracellular membranes (Craig *et al.*, 1993; Craig *et al.*, 1990; Welch, 1993). The functions that HSP70 performs in protein metabolism are essential for cell survival under all conditions (Craig *et al.*, 1993; Hartl, 1996; Welch, 1992) which is in accordance with the universal distribution of this protein in all forms of life.

The use of HSP70 in evolutionary studies is highly justified for a number of different reasons (Gupta and Golding, 1993; Gupta *et al.*, 1994; Gupta and Singh, 1994): (i) the HSP70 gene has been found in all species examined, except in the archaebacterium *Methanococcus janaschii* (Bult *et al.*, 1996). It is also found in most eukaryotic cellular compartments, i.e. cytoplasm, endoplasmic reticulum (ER), mitochondria, and chloroplasts. Thus, HSP70 is useful for examining the origin and evolutionary relationships of organisms as well as eukaryotic organelles. (ii) It constitutes the most conserved protein known as the minimal amino acid sequence identity that is observed between HSP70 homologs from any two species is about 45% over the entire length. (iii) HSP70 has a large size ($\approx$ 70 kDa) so it possesses adequate information for comparative studies. (iv) HSP70 homologs from different species have functional equivalence and constancy facilitating the determination of relationships over a broad range of evolutionary distance. (v) HSP70 homologs from different species have similar size (600-650 amino acids) which allows alignment of most of their length with minimal ambiguity and minimizes statistical fluctuations inherent in the data sets. (vi) HSP70 homologs from eukaryotic species have been extensively characterized. Hence, HSP70 homologs

that are present in eukaryotic cells (i.e. orthologous, paralogous and xenologous sequences) are readily distinguished. (vii) HSP70, as a protein molecule, can accommodate changes in the nucleotide sequences stemming from differences in the G+C contents of species. (viii) HSP70 has sequence signatures that are distinctive of different groups of species. Given all of the above characteristics and considering its important function, HSP70 provides a very useful model for evolutionary studies.

The global alignments of HSP70 sequences from species representing all three domains have revealed several sequence signatures specific to certain groups of species (Boorstein *et al.*, 1994; Gupta *et al.*, 1997; Gupta and Golding, 1993; Gupta and Singh, 1994; Gupta and Singh, 1992). Phylogenetic analyses based on these sequences have also revealed novel evolutionary relationships among groups of species (Gupta *et al.*, 1997; Gupta and Golding, 1993; Gupta and Singh, 1992, 1994). The sequence signatures and phylogenetic analyses of HSP70 have shown a clear distinction between prokaryotic (eubacterial and archaebacterial) and eukaryotic homologs. Other sequence signatures as well as the overall branching pattern of HSP70-based phylogeny showed a clear distinction between eukaryotic cytosolic homologs and gram-negative bacteria on the one hand and archaebacteria and gram-positive bacteria on the other. Among the sequence signatures found in HSP70, the most striking one is the presence of a 23-27 amino acid segment in the amino terminal quadrant of HSP70. This segment is present in all HSP70 sequences from gram-negative bacteria and eukaryotes including mitochondria, chloroplast, cytosolic (nuclear), and ER, but not in any of the homologs from archaebacteria or gram-positive bacteria. However, there are no signature sequences found that are specific for the archaebacterial homologs. The eukaryotic homologs (cytosolic, ER) are distinctive in containing many signature sequences that are not found in any of the prokaryotic or organellar homologs. The phylogenetic branching of the HSP70 sequences also support the

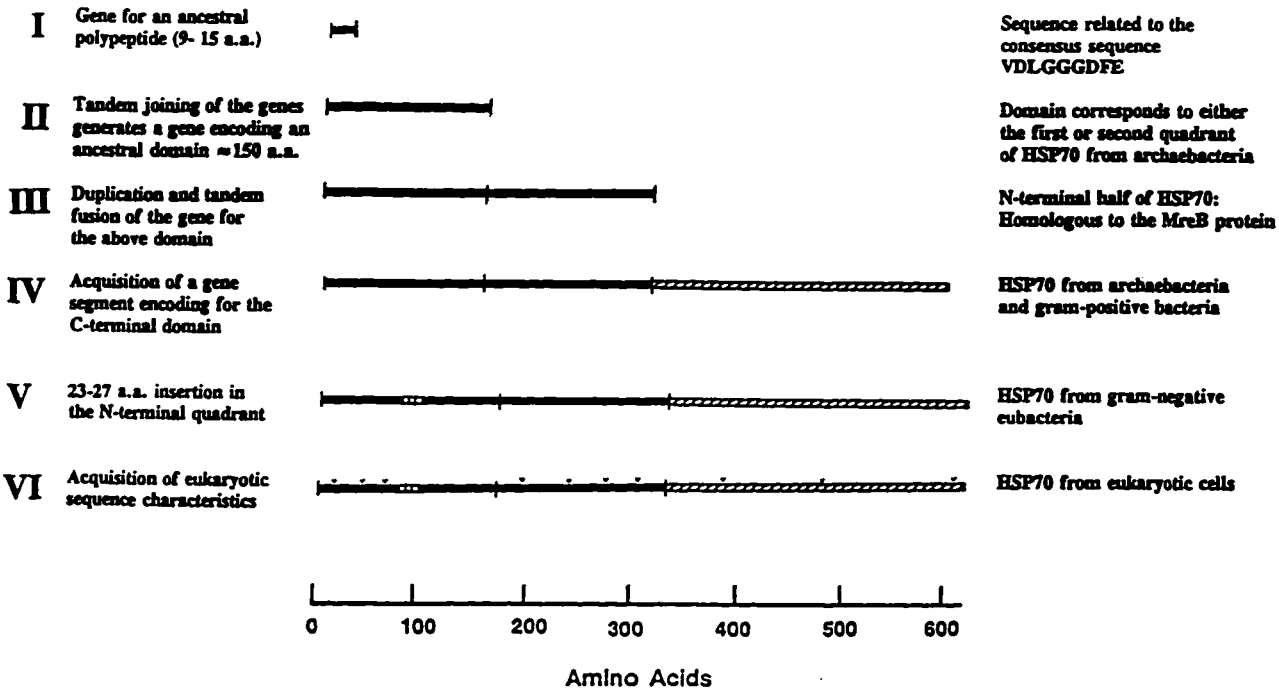
species grouping based on the sequence signatures. The eukaryotic species (cytosolic and ER species) branch as a monophyletic group with strong statistical support and their branching is quite distinct from organellar homologs. Archaebacterial and gram-positive eubacterial species cluster together in the HSP70 tree. Although sequence signatures are important for species identification, they do not affect the branching pattern of species in the phylogeny. The exclusion of the 23-27 amino acid segment from HSP70 species does not change the branching pattern of the species (Gupta and Golding, 1993). Therefore, the branching pattern of the species in the HSP70 tree is not dependent on the presence or absence of the segment but rather on the overall sequence similarity of the protein. Thus, inferences from HSP70-based phylogeny are robust and reliable, yet at considerable variance with the current view concerning the evolutionary relationship amongst species based on rRNA and some other protein sequences.

The studies based on HSP70 sequences also provide information concerning the evolution of the HSP70 gene. Significant sequence similarity is observed between the first and second quadrants of HSP70 homologs from archaebacteria and gram-positive eubacteria (Gupta and Singh, 1992). Such sequence similarity is not apparent in HSP70 homologs from eukaryotes and gram-negative bacteria. This observation suggests that the first and second quadrants arose by duplication of an ancestral gene, and that the 23-27 amino acid segment might have been inserted at a later time (Gupta and Singh, 1992). This view has received support from the similarity in the three dimensional structure between the first and second quadrants of HSP70 (Flaherty *et al.*, 1990) as well as by the observation that the 23-27 amino acid segment is located on the outside in the tertiary structure of the protein (Gupta and Singh, 1992). Additionally, the protein MreB of *E. coli* (gram-negative bacterium) and *Bacillus subtilis* (gram-positive bacterium) (Doi *et al.*, 1988; EMBL Database) and

the N-terminal half of HSP70 from gram-positive bacteria and archaebacteria exhibit significant sequence similarity (Gupta and Singh, 1992; Gupta and Golding, 1993). MreB is only about half the length of HSP70 and it very likely evolved from a predecessor of HSP70 before the acquisition of the C-terminal half (Gupta and Singh, 1992). The alignment of MreB with HSP70 proteins from gram-negative bacteria or eukaryotic cells requires a gap in the MreB sequence to adjust for the 23-27 amino acid segment whereas alignment of MreB with archaebacterial or gram-positive bacterial homologs requires no adjustments (Gupta and Singh, 1992). The absence of the 23-27 amino acid segment in MreB protein strongly indicates that the ancestral HSP70 lacked this segment and that it has been introduced at a later time. The evolution of MreB protein from an ancestral HSP70 by the loss of the C-terminal half is considered unlikely because MreB proteins from gram-negative and gram-positive bacteria lack the 23-27 amino acid insert (Gupta and Golding, 1993). HSP70 homologs from gram-positive bacteria and archaebacteria which lack the 23-27 amino acid insert therefore represent a more ancient version of the HSP70 protein (Gupta and Golding, 1993; Gupta and Singh, 1992). The N-terminal half of HSP70 has also been found to contain several repeats of a conserved segment of about a 9-15 amino acids (Gupta and Golding, 1993). Based on these observations a model (Figure 2) for the evolution of the HSP70 gene has been proposed (Gupta and Golding, 1993; Gupta and Singh, 1992). It is proposed that initially, there was a tandem joining of an ancestral gene for 9-15 amino acid polypeptide which is repeated in the N-terminal quadrant leading to the formation of a domain of about 150 amino acids. This ancestral gene then duplicated and fused together giving rise to the N-terminal half of the protein. The C-terminal domain of HSP70 was acquired at the next stage, leading to a protein corresponding to HSP70 homologs from archaebacteria and gram-positive bacteria. In the fifth stage, the 23-27 amino acid segment was

Figure 2. Hypothetical model showing the various stages in the evolution of the HSP70 gene.

Adapted from Gupta and Golding 1993.


I
Gene for an ancestral polypeptide (9- 15 a.a.)
Sequence related to the consensus sequence VDLGGGDFE
II
Tandem joining of the genes generates a gene encoding an ancestral domain ≈150 a.a.
Domain corresponds to either the first or second quadrant of HSP70 from archaebacteria
III
Duplication and tandem fusion of the gene for the above domain
N-terminal half of HSP70: Homologous to the MreB protein
IV
Acquisition of a gene segment encoding for the C-terminal domain
HSP70 from archaebacteria and gram-positive bacteria
V
23-27 a.a. insertion in the N-terminal quadrant
HSP70 from gram-negative eubacteria
VI
Acquisition of eukaryotic sequence characteristics
HSP70 from eukaryotic cells
0
100
200
300
400
500
600
Amino Acids

inserted into the amino terminal quadrant giving rise to a protein similar to that found in gram-negative bacteria. In the last stage, the eukaryotic sequence signatures were acquired to form the eukaryotic HSP70 homologs. It is postulated that the first four stages took place before the divergence of archaebacteria, eubacteria and eukaryotes (Gupta and Golding, 1993). The global alignment of HSP70 sequences indicates that the eukaryotic sequence signatures are specifically found in the cytosolic and ER homologs. This observation and the monophyletic nature of eukaryotes as indicated by HSP70-based phylogeny, strongly suggest that these sequence signatures have been introduced early in the evolution of eukaryotic cells (Gupta and Golding, 1993).

Evolution of Prokaryotes

Based on cytological distinctions, all living systems were placed into one of two categories, the prokaryotes and eukaryotes (Murray, 1962, 1974; Stanier, 1970). The prokaryotes which include all bacteria lack a membrane-bounded nucleus that distinguishes the eukaryotic cells. All prokaryotes have this cytological distinction and yet are diverse among themselves in cell shape, structure, physiology, metabolism, motility, and growth (Dworkin, 1992). Because of this diversity all prokaryotes are not necessarily closely related to one another phylogenetically. The classification of prokaryotes is therefore important to understand their phylogenetic relationships. Studies in comparative morphology in eukaryotes such as animals and plants were successful and led to a good understanding of their phylogenetic relationships. By contrast, prokaryotes do not possess complex and detailed morphologies on which phylogeny can be based. Therefore, such phylogenetic principles were not very successful in the classification of bacteria due to bacteria's simple morphology which may or may not reflect evolutionary relationship. Nevertheless, micromorphology has been utilized for this purpose. Stain characteristics and microscopic research on prokaryotic organisms identified three common cell shapes: cocci (spherical), bacilli (rodlike), and spiral (Joklik *et al.*, 1992). By the use of Gram stain, bacteria were classified into two major groups: gram-positive bacteria which stain blue, and gram-negative bacteria which stain red (Joklik *et al.*, 1992). This approach formed the basis for the well known *Bergey's Manual of Systematic Bacteriology* that is useful for bacterial identification (Sneath *et al.*, 1986). The Gram staining is dependent on the thickness of the peptidoglycan layer in the cell wall and the permeability properties of the intact cell envelope (Joklik *et al.*, 1992). However, this may not suit some bacteria with a cell envelope that is chemically and structurally different from that of gram-positive and gram-negative bacteria. The methanogeneic

archaebacteria, none of which contain peptidoglycan cell wall, in part stain gram-negative, in part gram-positive (Kandler and Konig, 1985). Halobacteria also lack peptidoglycan, but while *Halobacterium* reacts gram-negative, the *Halococcus* species are gram-positive (Kandler and Konig, 1985). Also, this approach would not be useful for the classification of wall-less bacteria (e.g. *Mycoplasma and Thermoplasma*). Thus, the cell envelope stain-based approach can not be applied for the classification of all bacteria.

Cavalier-Smith (1987) suggests that the gram-negative bacteria are the most primitive of all life forms based on cellular characteristics, evolving directly from a primitive progenitor which has an outer membrane and murein (peptidoglycan). This organism, through a series of mutational transformations and the loss of the outer membrane, gave rise to the gram-positive bacteria. Gram-positive bacteria in turn lost the murein and this led to the origin of archaebacteria. After the loss of the murein which is a component necessary for cell support, the organisms diverged into two lines of descent. One is the archaebacteria which developed a new type of rigid cell envelope built from ether-linked membrane lipids and without murein. The other cells are the eukaryotes which developed an internal proteinaceous cytoskeleton. This model is largely speculative and can not be used to precisely determine the branching order of the prokaryotic species or construct a phylogenetic tree.

The application of rRNA sequence analysis played a key role in suggesting the phylogenetic relationships within the prokaryotes and had a dramatic effect on the above views. Sequence comparison of the rRNA from thermophilic and halophilic bacteria revealed that their rRNAs were dissimilar from typical bacterial sequences. As indicated previously, this sequence comparison showed that there were not one large group of bacteria, the prokaryotes, but two: the eubacteria and

the archaebacteria (Woese, 1987; Woese *et al.*, 1990). The eubacterial phylogeny has been studied largely on the basis of the sequence of rRNA. These studies have led to the division of eubacteria into several main groups. These include thermotogales, green nonsulfur bacteria, deinococci, cyanobacteria, low-G+C and high-G+C gram-positive bacteria, proteobacteria or purple bacteria, spirochetes and chlamydiae, planctomycetes, green sulfur bacteria, flavobacteria and bacteroids (Olsen *et al.*, 1994; Woese, 1987, 1991). Some of these groups contain some unexpected members which are morphologically and physiologically different. For example, *Mycoplasmas* although they lack a cell wall, branch within the gram-positive group of bacteria. This species which is bounded only by a membrane, is believed to have lost the ability to synthesize a cell wall (Maniloff, 1992). Previously based on gram staining, *Mycoplasmas* were thought to be derived from gram-negative eubacteria, but now based on sequence data, it is clear that *Mycoplasmas* have arisen from a gram-positive bacterial ancestor (Olsen *et al.*, 1994; Woese, 1987; Woese *et al.*, 1980). In this eubacterial phylogeny, it is indicated that all gram-positive species branch in one group, while gram-negative species branch polyphyletically deeper and lower than gram-positive species. It is of interest in this regard that a number of species show some characteristics of both gram-positive as well as gram-negative bacteria. *Deinococcus*, for example, are orange to red bacteria that show gram-positive staining (Sneath *et al.*, 1986). However, a number of characteristics of this species, including the presence of an outer membrane and a fatty acid profile that is rich in palmitoleate and lacking any branched-chain members, indicate that they are more similar to gram-negative bacteria and that the gram-positive staining is probably due to the thickness of the peptidoglycan component (Brooks *et al.*, 1980; Sneath *et al.*, 1986; Murray, 1992). Based on rRNA sequence analysis, *Deinococcus* has been placed in a separate group which branches early in the evolution of eubacteria (Olsen *et al.*,

1994; Woese, 1987).

Given these examples, eubacterial species are better defined based on molecular data. Most eubacterial groups have been found to be coherent on the basis of rRNA sequence. With this framework, it would be useful to also include phenotypic features so that a bacterial group can be defined phenotypically as well as phylogenetically. Such phenotypic markers should be good indicators of phylogeny. A recent suggestion by Dr. R. S. Gupta (private communication) could be a better tool than gram staining for bacterial classifications. He suggests that all bacterial species could be classified on the basis of the number of their biological membranes. All gram-positive bacteria (including *Mycoplasma)* and archaebacteria (including *Thermoplasma*) have a single cellular membrane, whereas all "true" gram-negative bacteria have two membranes. Therefore, the number of membranes in prokaryotic species can provide a clear distinction between gram-positive bacteria and archaebacteria (one membrane) on one hand and gram-negative bacteria (two membranes) on the other. The presence of one membrane in gram-positive bacteria and archaebacteria is most likely a common evolutionary trait because of the close evolutionary relationship seen between these two groups based on molecular sequence data of HSP70 and other proteins (to be discussed later).

Comparative studies based on rRNA indicate that archaebacteria are a distinct group and are distantly related to eubacteria (Olsen *et al.*, 1994; Woese, 1987, 1996). This view was supported by many fundamental differences which distinguish archaebacteria and eubacteria (Zillig *et al.*, 1993). Archaebacterial RNA polymerases are of at least two types, and are structurally distinct from the single type found in eubacteria (Zillig *et al.*, 1985). Also, the structure and function of ribosomes are different from that in eubacteria and eukaryotes (Wittmann-Liebold *et al.*, 1990). Most antibiotics that inhibit transcription and translation in eubacteria have no effect in archaebacteria (Zillig *et al.*,

1988) and the tRNAs of archaebacteria show a characteristic pattern of modified bases (Edmonds *et al.*, 1991). One of the most distinctive features of the archaebacteria is the chemistry of the membrane lipids (DeRosa et al., 1986; Langworthy, 1985). These are branched-chain, ether-linked lipids, common and unique to all archaebacteria. Based on the above characteristics archaebacteria are proposed to be a distinct monophyletic group (Zillig, 1991; Zillig *et al.*, 1993).

Phenotypically archaebacteria comprise four main metabolic types: the methanogens, the extreme halophiles, sulfate-reducing archaebacteria and the sulfo-thermophiles (Jones *et al.*, 1987; Woese, 1996). However, these four distinct phenotypes do not define four separate phylogenetic groupings. Rather, they are phylogenetically intermixed. Based on rRNA phylogeny, the archaebacteria form two distinct lineages; the *Crenarchaeota* and the *Euryarchaeota* (Woese, 1991, 1996). The *Euryarchaeota* comprises a mixture of all archaebacterial phenotypes whereas the *Crenarchaeota* are exclusively sulfo-thermophiles (Woese, 1991, 1996; Woese *et al.*, 1990). The latter group of archaebacteria is considered to be more ancient (Woese, 1987) and referred to as "eocytes" in other studies (Lake, 1988, 1989). The *Crenarchaeota* grow at extremely high temperatures and require sulphur as an energy source (Woese, 1987). The role of sulphur in the metabolism together with the thermophilic character of the *Crenarchaeota* has led to the suggestion that they represent an ancient metabolic format (Woese, 1987, 1991). Woese *et al.* (1987,1990) consider archaebacteria to be distinctive with regard to molecular structure of their rRNA genes and therefore have given them equal status with the existing kingdoms of prokaryotes and eukaryotes. However, based on comparison of SSU rRNA sequences Lake (1988) has argued that archaebacteria are polyphyletic, and according to his proposal eocytes are more closely related to eukaryotes than to other archaebacteria and eubacteria, whereas methanogens and halophiles are closer to eubacteria.

The relationship suggested by Lake is strongly supported by the polypeptide insert found exclusively in EF-1α/Tu sequences from eukaryotes and eocytes (Rivera and Lake, 1992). Along this line, the composite trees of EF-1α/Tu and EF-2/G have given interesting results (Hashimoto and Hasegawa, 1996). In the EF-1α/Tu part of the composite tree, eocytes were the closest relatives of eukaryotes, whereas in the EF-2/G part, methanogens and halophiles were closer to eukaryotes than the eocytes. Thus the monophyletic or polyphyletic nature of archaebacteria and their relationship to eukaryotes need to be resolved.

Detailed phylogenetic analyses and sequence features of highly conserved proteins such as HSP70, glutamine synthetase (GS) and glutamate dehydrogenase (GDH) have shown a different view with regard to the evolution of archaebacteria and its relationship to other species. Based on these protein sequences, a specific relationship between archaebacteria and gram-positive bacteria has been found (Golding and Gupta, 1995). In the glutamine synthetase I (GSI) gene family sequence analysis, all gram-positive bacteria and various archaebacteria (except *Sulfolobus solfataricus* which is an eocyte) lack a 25 amino acid insert that is present in the gram-negative bacteria (Brown *et al.*, 1994). Based on the alignment of GDH sequences, the archaebacterial and gram-positive bacterial homologs exhibit significant sequence similarity and they share a three amino acid gap in a conserved region of the protein (Gupta, 1997). Besides these sequence signatures, phylogenetic analysis based on GSI and GDH sequences strongly supports the close evolutionary relationship between archaebacteria and gram-positive bacteria (Benachenhou-Lahfa *et al.*, 1993; Brown *et al.*, 1994; Golding and Gupta, 1995). In GDH-based phylogeny, archaebacteria are not a monophyletic group but branch polyphyletically with gram-positive bacteria (Benachenhou-Lahfa *et al.*, 1993).

In addition to the above, in HSP70-based phylogeny the branching of the archaebacteria

within the gram-positive species further suggests that archaebacteria are not a monophyletic group (Gupta and Golding, 1993; Gupta and Singh, 1992, 1994). Sequence analysis of HSP70 homologs from organisms representing all three domains showed that the four archaebacterial species examined (namely *Thermoplasma acidophilum, Methanosarcina mazei, Halobacterium marismortui, and Halobacterium cutirubrum)* bore all the distinguishing characteristics of eubacterial species, particularly those from the gram-positive bacteria (Gupta and Singh, 1994). However for the *Crenarchaeota* (eocytes), which are believed to be the closest relative to eukaryotes (Lake 1988, 1989), no HSP70 sequence is available at present. Nevertheless, the phylogenetic analysis based on all archaebacterial, eubacterial, and representative eukaryotic HSP70 sequences strongly supports a close and specific relationship between the archaebacteria and gram-positive group of bacteria (Gupta and Singh, 1994). In phylogenies using different methods, the methanogenic archaebacterium *M. mazei* branches with gram-positive species with low G+C content, whereas *Halobacteria* showed strong affinity for the group of gram-positive bacteria with high G+C content (Gupta and Singh, 1994). *T. acidophilum* which lacks the cell wall typical of either archaebacteria or of eubacteria, branches separately from the other archaebacteria, forming the outgroup of the clade of archaebacteria and gram-positive bacteria (Gupta and Singh, 1994). The paraphyletic branching of archaebacteria with gram-positive bacteria does not result from a clustering effect of species with different genomic G+C content (Gupta and Singh, 1994), as the exclusion of homologs from the gram-positive species with high G+C content did not affect the branching pattern. Furthermore, the branching of archaebacteria with gram-positive species is not dependent upon or affected by the point of the placement of the root within the phylogenetic trees (Gupta and Singh, 1994). Thus, based on HSP70 sequence data the close evolutionary relationship between archaebacteria and gram-positive

bacteria is robust and reliable.

With regard to the branching pattern of eubacterial groups, the phylogenies based on 16S rRNA sequences indicate that the gram-positive group of bacteria are a coherent group, but gram-negative bacteria are not (Olsen *et al.*, 1994; Woese, 1994). The gram-negative group of bacteria encompasses 10 distinct groups, each of which is an equivalent of the gram-positive group of bacteria. The groups thermotogales, green nonsulfur bacteria, deinococci and cyanobacteria generally show deeper branching than gram-positive species, whereas the other groups consisting of proteobacteria, planctomyces, green sulfur bacteria, spirochetes and chlamydiae, and falvobacteria, branch either more recently or in a similar position to gram-positive species (Olsen *et al.*, 1994; Woese, 1987, 1994). However, as has been noted (Gupta, 1997; Gupta *et al.*, 1997) in the 16S rRNA phylogenies, the bootstrap scores for many of the critical nodes leading to the gram-positive bacteria and those of deep branching species are in the range of 25-50%, indicating that their placements are not reliable (Eisen, 1995; Van de Peer *et al.*, 1994). Similar low bootstrap scores have been obtained for phylogenies based on 5S and 23S rRNA sequences (De Rijk *et al.*, 1995; Van den Eynde *et al.*, 1990). The current eubacterial phylogeny and branching order is mainly based on the rRNA sequences due to the availability of sequences representing a wide range of species and the relative ease of cloning and sequencing these genes even from unculturable species (Lane *et al.*, 1985). However, phylogenies based on other conserved protein sequences such as GroEL have shown that gram-positive and gram-negative bacteria are phylogenetically distinct groups (Gupta, 1995, 1996; Viale *et al.*, 1994). This is supported by the observation that homologs of GroEL from gram-negative eubacteria have a signature sequence of a one amino acid insert, which is absent from all the gram-positive species (Gupta *et al.*, 1997). Based on this sequence the various groups of

eubacteria could be arranged in a number of distinct groups from the most ancient to recent: gram-positive bacteria, cyanobacteria and chloroplast homologs, chlamydiae and spirochaetes, proteobacteria and finally mitochondrial homologs (Gupta, 1995, 1996).

In HSP70-based phylogeny the eubacterial species branch in two main groups, gram-negative and gram-positive bacteria (Gupta *et al.*, 1997). While gram-positive species branch polyphyletically with archaebacteria, gram-negative species branch with eukaryotic species (Golding and Gupta, 1995; Gupta *et al.*, 1997; Gupta and Singh, 1994). Interestingly, *Deinococcus* which show characteristics of both gram-negative as well as gram-positive staining, along with cyanobacteria constitute the deepest branching lineage within the gram-negative species (Gupta *et al.*, 1997). Thus, based on HSP70 and other protein sequences, three main prokaryotic groups could be recognized: archaebacteria, gram-positive bacteria and gram-negative bacteria. However, the division among them is not a clear-cut one. Many species based on sequence data have been found to branch with one group while having cellular characteristics of other distantly related groups. For example, cyanobacteria, which based on rRNA sequence analysis constitutes one of the deepest branching lineages within the gram-negative species, have a close affinity to the gram-positive group of bacteria. The deep branching of cyanobacteria was also seen in phylogenies based on GroEL and HSP70 sequences. In addition, some sequence signatures show a close relationship of cyanobacteria to gram-positive and archaebacteria. In the formylglycineamide ribotide amidotransferase (FGARAT) sequence, the cyanobacterial homolog shares two large deletions with all gram-positive and archaebacterial homologs (Gupta and Golding, 1996). *Thermotoga maritima* (thermotogale species), which branches the deepest in the rRNA tree is proposed to be one of the gram-negative bacteria (Olsen *et al.*, 1994). However, based on GSI sequence analysis it branches within the gram-positive

group of bacteria, and thus is indicated to be a gram-positive species (Brown *et al.*, 1994). While *Deinococcus* and cyanobacteria have been studied based on HSP70 sequence, which has established their early branching with the gram-negative bacteria, no HSP70 sequence is available from thermotogales. Cavalier-Smith (1992) proposed that thermotogales and gram-positive bacteria might be closely related, since both have a single membrane. The early branching of *Deinococcus* and cyanobacteria in the evolution of gram-negative bacteria based on HSP70 sequence analysis (Gupta *et al.*, 1997), and their close relationship to gram-positive bacteria and archaebacteria strongly indicate that these species might have been intermediates in the transition between the groups of gram-positive and gram-negative bacteria (Gupta, 1997). Further, the branching of archaebacteria within gram-positive bacteria based on HSP70, GDH, GSI, and FGARAT strongly suggests that either archaebacteria evolved from gram-positive bacteria or vice versa (Gupta, 1997). The gram-positive species *Aquifex pyrophilus* whose ribosomes resemble those of archaebacteria (Acca *et al.*, 1994) might have also been an intermediate in the transition between gram-positive bacteria and archaebacteria (Gupta, 1997). These observations indicate that the evolutionary changes that gave rise to the main prokaryotic groups might have been gradual and continuous.

Evolution of Eukaryotes

Eukaryotes, which comprise all multicellular organisms, as well as a broad range of single-cell organisms, are distinguished from prokaryotes by the presence of several internal structures. These include the nucleus, mitochondria, chloroplasts, endoplasmic reticulum (ER), and other membrane-bound compartments such as the lysosomes, peroxisomes, and the Golgi apparatus (Alberts *et al.*, 1994). Unlike most prokaryotes, which have a cell wall, eukaryotic cells have an internal structural support called "the cytoskeleton". These features indicate the complexity of eukaryotic cells,

compared to the simpler structure of prokaryotes, and suggest that they have been acquired during the evolution of the eukaryotic cell. The hallmark feature of all eukaryotes is the presence of the nucleus with a double membrane separating the chromosomes from the rest of the cell. As such, the genesis of the nucleus has often been the focus of efforts to explain the origin of the eukaryotic cell. The evolution of eukaryotes was initially studied based on interpretation of fossil evidence and comparisons of common features among contemporary organisms. According to the fossil records, prokaryotic cells were present on the earth 3.5 to 4.0 billion years ago, while eukaryotic cells did not appear until 1.8 to 2.0 billion years ago (Knoll, 1992). The fossil evidence in conjunction with the structural complexity of eukaryotes suggests that eukaryotes originated from structurally simpler prokaryotic ancestors.

With the advent of molecular data the differences between eukaryotes and prokaryotes have narrowed and an increasing number of similarities that links eukaryotes to prokaryotes began to emerge. However, questions as to how eukaryotic cells evolved from prokaryotes have been difficult to explain due to lack of intermediates from the transition of prokaryotes to eukaryotes. Since only the final eukaryotic product is available, the evolutionary relationship between prokaryotes and eukaryotes is studied using their common chemistry, physiology and metabolism as well as common molecules in their biochemical pathways.

According to Cavalier-Smith (1987), early diverging eukaryotes are suggested to have originated as transient intermediates in the evolutionary process between gram-positive bacteria and archaebacteria. After the loss of the murein from gram-positive bacteria, some of these organisms developed a cytoskeleton. It is at this point that the eukaryotic cells emerged. It is the development of the cytoskeleton and not the nucleus that is the hallmark in the evolution of eukaryotes according

to Cavalier-Smith (1987). When the eukaryotic cells developed an internal support, the membrane was kept fluid and developed the ability for endocytosis and exocytosis. The development of endocytosis might have led to invagination in the membrane and gave rise to both the nucleus and ER. It is also suggested that the first eukaryote would have been anaerobic and subsequently acquired aerobic free-living bacteria to give rise to aerobic eukaryotes, which have mitochondria. This theory was further supported with the identification of *Giardia lamblia*, a eukaryote lacking mitochondria (Adam, 1991; Cavalier-Smith, 1993). If this is true, *G. lamblia*-like organisms are probably the missing link between prokaryotes and eukaryotes, from which all eukaryotic cells have descended.

With regard to the evolution of the eukaryotic nucleus, yet another proposal was offered. Zillig (1991) has based his model on several molecular and metabolic features of all the major groups of organisms: eubacteria, archaebacteria and eukaryotes. Based on a large number of eubacterial and archaebacterial features that are present in eukaryotes, he suggested that the genesis of the eukaryotic cell is a fusion product of an archaebacterium and a eubacterium (Zillig, 1991). This model would have been much stronger if it had been based on quantitative measures (sequence data) rather than qualitative measures (cellular features).

With the molecular sequence data the origin of eukaryotic cells could be better assessed. The SSU rRNA sequences were exploited to determine how closely or distantly eukaryotes are related to prokaryotes (Leipe *et al.*, 1993; Sogin, 1991; Sogin *et al.*, 1996). These studies were used to determine which organisms represent the earliest lines of descent among the eukaryotes (Sogin *et al.*, 1989). *G. lamblia* was discovered to have a rRNA sequence that is very close to that of prokaryotic cells. This *G. lamblia* rRNA shares more of its sequence with prokaryotes (eubacteria and archaebacteria) than does the corresponding rRNA of any other eukaryote. This means that *G.*

lamblia is a member of the earliest emerging eukaryotic lineages.

Most contemporary eukaryotes have mitochondria which are believed to have descended from eubacteria by endosymbiosis. As indicated above, *G. lamblia*, which is a free living unicellular protist, has no mitochondria (Cavalier-Smith, 1993). Being a eukaryote and yet having no mitochondria, this protist is believed to have branched off in the earliest phases of eukaryotic evolution before the symbiotic origin of mitochondria occurred. Based on the rRNA tree (Leipe *et al.*, 1993; Sogin, 1991; Sogin *et al.*, 1989), protists lacking mitochondria branched as the earliest offshoots in the eukaryotic tree. Such species include *Vairimorpha necatrix* (Microsporidians), *Tritrichomonas foetus* (Trichomonads), and *G. lamblia* (Diplomonads). However, as indicated previously, this phylogenetic tree, inferred from a single gene of rRNA may not be reliable because of bias in base composition prevalent among species (Hasegawa and Hashimoto, 1993). For example, the rRNA of *G. lamblia* has very high G+C content (≈75%) which is nearly double that found in *V. necatrix* (Hashimoto and Hasegawa, 1996). Protein-based phylogenies, which are less affected by base variation, would produce robust estimation of the early divergence of eukaryotes. EF-1α/Tu and EF-2/G protein sequence analyses, in turn, have further confirmed that the mitochondrion-lacking species *Glugea plecoglossi* (Microsporidians) and *G. lamblia* (Diplomonads) are the earliest offshoots of the eukaryotes (Hashimoto and Hasegawa, 1996). The lack of mitochondria and the early branching of these eukaryotes may suggest that their nucleus and possibly the endoplasmic reticulum have primitive characteristics relative to those found in higher eukaryotes.

The evolution of eukaryotes has been further investigated based on rRNA sequences. Sogin and coworkers have defined five major eukaryotic groups plus two clusters consisting of a series of paraphyletic protist lineages (Leipe *et al.*, 1993; Sogin, 1991; Sogin *et al.*, 1996). All these

eukaryotic groups cluster in a distinct single domain. Furthermore, in rRNA-based phylogenies eubacteria and archaebacteria appear to be more closely related to each other than to eukaryotes (Woese, 1987,1992). This is the principal discrepancy between rRNA-based phylogenies and those inferred from comparisons of protein sequences of duplicated genes (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989). Protein based phylogenies, as indicated previously, position archaebacteria as specific relatives to the eukaryotes. To reconcile these discrepancies between rRNA and protein based phylogenies, Sogin (1991) proposed a model which describes a process leading to formation of a stable nucleus in a eukaryotic lineage. In the proposed model, a proto-eukaryotic lineage and a proto-bacterial lineage have evolved from a progenote. Archaebacteria and eubacteria were derived from the proto-bacterial lineage, while the eukaryotic cell emerged due to engulfment of an archaebacterial organism into a proto-eukaryotic lineage. The archaebacterial DNA formed a chimera with the existing proto-eukaryotic genome and led to the formation of the eukaryotic nucleus. In this model, it is proposed that the archaebacterial genome has contributed the majority of protein-coding regions found in contemporary eukaryotes, while the proto-eukaryotic lineage, which is proposed to be completely unlike contemporary organisms and contained an RNA-based metabolism, has contributed the cytoskeleton. Sogin (1991) hypothesizes that the eukaryotic nucleus has originated at least in part from a prokaryotic endosymbiont, but that the host organism was actually an RNA-based one, whose distinguishing feature was the possession of a cytoskeleton that conferred on it the ability to engulf other microorganisms.

In phylogenetic trees based on sequences of some conserved proteins a closer and specific relationship of the eukaryotic homologs to the archaebacterial counterparts is observed. This analysis led to the presently favoured view that the eukaryotic cells have directly descended from an

archaebacterial ancestor (Figure 1). The most compelling phylogenetic evidence is based on the sequence comparison of the elongation factors EF-1α/Tu and EF-2/G (Hashimoto and Hasegawa, 1996; Iwabe *et al.*, 1989), which is also supported by sequence signatures found only in homologs from eukaryotes and archaebacteia. Studies on EF-1α/Tu sequence alignment identified a 13 amino acid common insert that is present in homologs from eukaryotes and archaebacteria (Gupta, 1997), and another insert that is present only in homologs from one group of archaebacteria "the eocyte" and eukaryotes (Rivera and Lake, 1992). Similarly, the phylogenetic analysis of duplicated gene pairs for α- and β-subunits of ATPase supports the relatedness between these two groups (Gogarten *et al.*, 1989). T-complex polypeptide 1 (Tcp-1)-related chaperonins also show a specific relationship between eukaryotes and archaebacteria (Gupta, 1995). Further, the relationship between eukaryotes and archaebacteria is strongly supported by the presence of a eukaryotic histone in archaebacteria (Sandman *et al.*, 1990; Searcy, 1975, 1987), and by the structure and sequence similarity of DNA-dependent RNA polymerase (Schnabel *et al.*, 1982). The topology of the tree based on DNA-dependent RNA polymerase sequences agrees well with the phylogenies based upon the protein sequences of duplicated gene pairs (Puhler *et al.*, 1989). If the above phylogenies are correct, eukaryotes and archaebacteria must have shared a common evolutionary history exclusive of the eubacteria. Despite the support of the above studies to each other, the specific group of archaebacteria from which eukaryotes branched and the extent and nature of the genetic material that archaebacteria have contributed remain to be clarified. Woese (1991) has suggested that the last common ancestor of archaebacteria was a thermophilic, anaerobic, and sulfur-metabolizing organism. Because the *Crenarchaeota* (or eocyte) group is the most closely related to the archaebacterial common ancestor (Woese, 1991), it is expected that this group is more closely related to eukaryotes

than the other group of archaebacteria: *Euryarchaeota* (namely methanogens, and the extreme halophiles). Such a relationship between *Crenarchaeota* (eocytes) and eukaryotes was postulated based on rRNA sequences (Lake, 1988), and EF-1α/Tu sequence analysis (Hashimoto and Hasegawa, 1996; Rivera and Lake, 1992). The relationship of eukaryotes to this specific group of archaebacteria is further supported by the fact that the *Crenarchaeotal* translation apparatus shows a number of "eukaryotic-like" characteristics not seen in *Euryarchaeota* (Woese, 1991). Thus, it appears that archaebacterial species or more specifically *Crenarchaeota* (eocyte) have substantially contributed to the genetic make-up of the eukaryotic nuclear genome.

As sequence data on more proteins have become available, the above view has not been supported, and sequence features and phylogenies based on several proteins give very different view of the history of eukaryotes. Recent studies based on highly conserved proteins have proposed a close evolutionary relationship between the cytosolic eukaryotic homologs and gram-negative species, and that rather than with eukaryotes, the archaebacterial species grouped with gram-positive species. Most prominent of these proteins are glutamine synthetase (GS), glutamate dehydrogenase (GDH), glyceraldehyde 3-phosphate dehydrogenase (GAPDH), formylglycineamide ribotide amidotransferase (FGARAT), and HSP70.

Phylogenetic analysis based on GS has shown a close relationship between eukaryotes and gram-negative bacteria (Tiboni *et al.*, 1993). Homologs from this protein have been found in three distinctive types with different sizes and are located in different groups of organisms. Eukaryotes have a GSII homolog while GSI and GSIII occur in prokaryotes. Although GS has high degree of sequence conservation and is considered to be a good protein for the study of evolutionary relationships, only a small number of sequences are known and whether GSI and GSII are paralogs

or orthologs is not yet clear. Furthermore, the close relationship between eukaryotes and gram-negative bacteria could be due to lateral gene transfer between these two groups.

GDH is also one of the most conserved proteins known and provides a good tool for evolutionary studies (Benachenhou-Lahfa *et al.*, 1993; Golding and Gupta, 1995). Phylogenetic analysis based on GDH sequences have supported a relationship between eukaryotes and gram-negative bacteria. However to further delimit the origin of eukaryotes, more GDH sequences are needed especially from the archaebacteria; the methanogenes.

Another example of a protein sequence that demonstrates a close relationship between the eukaryotes and gram-negative bacteria is the enzyme FGARAT. From the sequence alignment of FGARAT sequences two indels (i.e. either insertions or deletions) are found to be uniquely shared by homologs from eukaryotes and gram-negative bacteria. A specific relationship between these groups is also strongly supported by phylogenetic analysis of the sequence data (Gupta and Golding, 1996).

The close relationship between eukaryotes and gram-negative bacteria has also been shown based on GAPDH sequence (Martin *et al.*, 1993). However it has been suggested that this relationship may not be reliable because the eukaryotic nuclear genes might have originated from the ancestor that gave rise to the organelles (i.e. chloroplasts and mitochondria).

Consequently, the notion one gets is that the inference concerning evolution of eukaryotes based on the above protein sequences is not that compelling due either to limited species representation and/or due to poor characterisation of the eukaryotic homologs. When comparing protein sequences from various species it is often very difficult to determine whether a particular eukaryotic homolog has originated from the nucleus (cytosolic) or acquired later by gene transfer

from the gram-negative bacterium that gave rise to mitochondria or chloroplast (organellar). If the sequences were acquired by such gene transfer, the phylogeny would give false estimates as to the origin of the eukaryotic cells. As discussed previously, HSP70 protein sequence provides the clearest and best studied example for studying the evolution of organisms. The eukaryotic cytosolic homologs of HSP70 are clearly distinguished from organellar homologs. This distinction has been based on the presence of specific signature sequences (Gupta and Golding, 1993; Gupta and Singh, 1992, 1994) as well as based on biochemical and cellular localization studies (Craig *et al.*, 1990). HSP70 homologs from the cytosol, ER, mitochondria and chloroplasts can be readily recognized (Gupta and Golding, 1993; Gupta and Singh 1992, 1994). The eukaryotic cytosolic and the ER-resident (cytosolic/ER) forms of HSP70s contain numerous sequence signatures that are not found in any of the prokaryotic or organellar homologs. Because of these signature sequences, the possibility that the eukaryotic cytosolic/ER homologs could be derived from organellar sequences is considered unlikely (Gupta and Singh, 1994; Gupta and Golding, 1996) and therefore the conclusion from HSP70-based phylogeny concerning the origin of eukaryotic cells can be drawn with confidence.

The eukaryotic cytosolic/ER and the organellar (i.e. mitochondrial and chloroplasts) homologs of HSP70 contain a number of sequence signatures, including the 23-27 amino acid insert, that are found in the gram-negative species but not in other prokaryotic homologs (Gupta and Golding, 1993; Gupta and Singh, 1992, 1994). This observation indicates that the cytosolic/ER and the organellar HSP70 sequences are derived from gram-negative species. Furthermore, the phylogenetic branching of the eukaryotic cytosolic/ER and the organellar homologs with gram-negative species supports their origin from this group of bacteria (Gupta and Golding, 1993; Gupta and Singh, 1994). In HSP70-based phylogeny, eukaryotic cytosolic/ER homologs branch in a monophyletic group which is

distantly related to the gram-negative and organellar species (Gupta and Golding, 1993; Gupta and Singh, 1994). The branching pattern of eukaryotic cytosolic/ER homologs in conjunction with their specific sequence signatures that is not found in any of the prokaryotic or organellar homologs further excludes the possibility of HSP70 gene transfer from the ancestor of either mitochondria or chloroplasts to the nucleus. In this context, it is of interest to characterize the sequence of HSP70 homologs from early diverging eukaryotes that lack mitochondria.

In summary, the conclusion from phylogenies based on HSP70 and other conserved protein sequences is that eukaryotic cells have evolved, in part at least, from a gram-negative ancestor. Although this conclusion is well supported by the bootstrap scores and the characteristics of HSP70 sequences, it is at variance with phylogenies based on the duplicated protein-coding genes for elongation factors and ATPase subunits as well as other protein sequences, which show eukaryotes as being more closely related to archaebacteria than to any other eubacterial group. Whether either of these very different views is true, or whether both are true and can be reconciled, needs further investigation. Detailed discussion on this issue will be saved to the last chapter as it concludes this thesis.

Evolution of Mitochondria

Eukaryotic organelles such as mitochondria and chloroplasts are among the cellular structures that distinguish the eukaryotic cell. These organelles display numerous features that are found in free living bacteria. The size, shape, protein synthetic system, electron-transport reactions and energy metabolism of mitochondria are very similar to those in bacteria (Alberts *et al.*, 1994). Furthermore, the α-proteobacteria *Paracoccus denitrificans*, has been found to possess a large number of mitochondrial features shown by its respiratory chain, oxidative phosphorylation system and membrane invagination which are remarkably similar to the cristae of mitochondria (Taylor, 1987). Because of such characteristics these bacteria were dubbed "free-living mitochondria" (John, 1987). Also, the protein synthesis system found in mitochondria and eubacteria shows a similar sensitivity (resistance) to inhibitors of translation and both employ fMet-tRNA^{Met} rather than Met-tRNA^{Met} as an initiator for protein synthesis (Gray, 1992). Based on these morphological, biochemical, and physiological characteristics an evolutionary relationship between these organelles and the eubacteria was proposed, and named "the endosymbiotic hypothesis" (Gray, 1992; Gray and Doolittle, 1982; Schwartz and Dayhoff, 1978; Taylor, 1987; Margulis, 1970). Specifically, mitochondria and chloroplasts are proposed to have originated from eubacterial ancestors by endosymbiotic events. Support for this hypothesis also came from sequence data. It has been found that certain genes for mitochondrial proteins show a closer relationship to eubacterial homologs than to nuclear or cytosolic homologs found in the same organism (Gupta and Golding, 1993; Yang *et al.*, 1985). The genome of mitochondria is much smaller than that of eubacteria (Alberts *et al.*, 1994), and therefore it is reasonable to assume that the transition from the endosymbiont (eubacterium) to integral cellular component (mitochondrion) must have been accompanied by a massive gene reduction and/or transfer

to the nucleus (Brennicke *et al.*, 1993; Smith *et al.*, 1992). The nuclear presence of genes for many mitochondrial proteins support the above hypothesis.

Evolutionary studies based on cytochrome *c* sequences have clearly shown a close relationship between mitochondrial homologs and the gram-negative eubacterial homologs of the α-subdivision group of proteobacteria (Dickerson, 1980; Schwartz and Dayhoff, 1978). Mitochondrial cytochrome *c* is an important protein in the electron transfer chain. In eukaryotes, it is coded in the nucleus but functions in the mitochondrion. It has three distinct sizes: small, medium and large. The mitochondrial cytochrome *c* sequences are of the medium size class similar to molecules present in the purple non-sulfur photosynthetic bacteria such as *Rhodopseudomonas virdis* and *Rhodomicrobium vannielii* (Dickerson, 1980). However, some of the other α-proteobacteria, e.g. *Paracoccus denitrificans*, *Rhodobacter sphaeroides*, and *Rhodobacter capsulatus* contain the large class of cytochrome *c* (Dickerson, 1980; John, 1987). Due to their large size variation and limited sequence conservation among distant species the cytochrome *c* sequences are not well suited for detailed phylogenetic analysis (Dickerson, 1980; Meyer *et al.*, 1986).

Although the endosymbiotic hypothesis for the origin of eukaryotic organelles from specific groups of eubacteria is supported by many different lines of evidence, the strongest evidence in this regard is based on phylogenetic analysis of 16S small subunit (SSU) and 23S large subunit (LSU) rRNA sequences (Gray, 1992; Gray *et al.*, 1989; Yang *et al.*, 1985). These analyses indicate that cyanobacteria and α-proteobacteria are the closest relatives of mitochondrial and chloroplast sequences, respectively. While rRNA provides a valuable chronometer for global phylogenies (Olsen *et al.*, 1994; Woese, 1987), its usefulness for establishing organelle phylogenies is somewhat compromised due to the large size variation in mitochondrial SSU and LSU rRNA sequences in

different eukaryotic organisms (Gray, 1988). Further, the G+C content of the mitochondrial SSU and LSU rRNA sequences varies considerably between different species (e.g. from 16.1% G+C in *Crithidia fasciculats* to 52.5% in wheat) and such differences in base composition can yield misleading phylogenetic trees (Hashimoto *et al.*, 1994; Steel *et al.*, 1993).

Recent phylogenetic analysis based on the sequence of the heat shock 60 kDa proteins (also known as chaperonins HSP60, or GroEL in eubacteria) strongly supports the endosymbiotic origin of mitochondria (Gupta, 1995, 1996; Viale and Arakaki, 1994). HSP60 is found in eubacteria and the eukaryotic organelles (mitochondria and chloroplast). Archaebacteria and eukaryotic cytosol have Tcp-1, a protein distantly related to HSP60 (Gupta, 1995). Phylogenetic studies based on this protein sequence are of particular relevance because of the concerns about rRNA phylogenies due to large differences in base compositions in different species. Phylogenetic analysis based on a large number of HSP60 homologs from eubacteria and organelles have clearly shown that the α-proteobacteria are the closest relatives of mitochondria (Gupta, 1995, 1996; Viale and Arakaki, 1994). Interestingly, the obligate intracellular pathogens *Ehrlichia chaffeensis* and *Rickettsia tsutsugamushi* are the closest and share a common ancestor with mitochondria. These eubacteria which are obligate parasites or endosymbionts of eukaryotic cells appear to have phenotypic traits of the mitochondrial ancestor.

It is of great importance to study the evolution of mitochondria based on the sequence analysis of HSP70 protein. HSP70 protein sequences from bacteria and organelles are well characterized and can be readily distinguished from each other as well as from the eukaryotic nuclear sequences. In contrast to cytochrome *c* and rRNA, HSP70 has very similar size in all different species. Furthermore, unlike HSP60, it is found in mitochondria as well as in the eukaryotic cytosol

and all bacterial species. Phylogenetic analysis of mitochondrial and various prokaryotic homologs of this highly conserved protein would provide important insight into the origin of mitochondria. In fact, homologs from proteobacteria should be included in this analysis and in particular those organisms that can live either as a free living bacterium or intracellularly in a symbiotic relationship with eukaryotic cells (Moulder, 1985). *Rhizobium meliloti*, *Agrobacterium tumefaciens* which can parasitize cell nodules of plant root, *Rickettsia tsutsugamushi*, and *Ehrlichia chaffeensis* which can parasitize animal cells are good candidates for this study.

Rationale and Goals of the Thesis

The sequence of HSP70s provides a unique view of the gene evolution and evolutionary relationships in general. Understanding the evolution of the most conserved protein known is necessary in order to understand the evolutionary relationships among eukaryotes and prokaryotes and how eukaryotes and their mitochondria and chloroplasts evolved throughout the history of life. The goals of this thesis were to study the following questions in evolution: (i) phylogenetic placement of *Mycoplasmas* within the prokaryotes, (ii) the branching position of *Giardia lamblia* within eukaryotes and the possible identification of endoplasmic reticulum in this organism, and (iii) the origin of mitochondria and its evolutionary relationship to *Rhizobium meliloti* and *Pseudomonas cepacia*. This entails sequencing all HSP70 genes from the indicated species. Sequencing these genes required designing primers, PCR amplification of the sequence, cloning PCR products and subsequently sequencing the complete genomic gene of HSP70. To analyse the sequences they were first aligned with available HSP70 sequences then the relationships and placement of these species in the phylogeny were determined. The sequences were also identified based on sequence signatures. The identity of ER in *G. lamblia* required specific antibodies to the ER homolog of HSP70 (referred to as Bip), therefore, the gene for Bip has been expressed and antibodies were raised against its purified protein. This antibody was used to localize the cellular distribution of Bip in *G. lamblia* using immunofluorescence and cryoelectron microscopy.

CHAPTER 2

Phylogenetic Analysis of *Mycoplasmas* Based on HSP70 Sequences: Cloning of the HSP70 *(dnaK)* Gene Region of *Mycoplasma capricolum*

In this chapter the phylogenetic placement of *Mycoplasma capricolum* in the global phylogeny is examined based on the protein sequence of HSP70. After the gene for HSP70 was cloned and sequenced from this species, the genome sequence of *Mycoplasma genitalium* was reported. Both the *Mycoplasma* homologs were therefore included with published sequences from prokaryotes, eukaryotes and organelles in the phylogenetic analysis. My contribution to this chapter was the complete sequencing of the HSP70 gene region (about 5.0 kb) including the PCR amplification, cloning of PCR products, and isolating and sequencing the genomic clones of HSP70. In addition, I was also involved in the phylogenetic analysis of the data and in the preparation of the manuscript. The *Mycoplasma capricolum* sequence has been deposited in the GenBank data base under the accession number U51235.

INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY, Jan. 1997, p. 38–45
0020-7713/97/$04.00+0

Vol. 47, No. 1

Phylogenetic Analysis of Mycoplasmas Based on Hsp70 Sequences: Cloning of the *dnaK* (*hsp70*) Gene Region of *Mycoplasma capricolum*

MIZIED FALAH AND RADHEY S. GUPTA*

Department of Biochemistry, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

A 5.0-kb region containing the *hsp70* (*dnaK*) gene was cloned from *Mycoplasma capricolum* and sequenced. In addition to the *hsp70* gene, this sequence region also contained the complete sequences for the *grpE* and *orfA* genes and partial sequences for the *clpB* and *dnaJ* genes. The order of the above gene sequences in the cloned fragment was found to be *clpB-orfA-grpE-hsp70-dnaJ*, which is similar to the order seen in various other gram-positive groups of bacteria. The Hsp70 homologs from two mycoplasma species, *Mycoplasma capricolum* and *Mycoplasma genitalium*, contain a number of sequence signatures, including the absence of a large insert in the N-terminal quadrant, that are characteristics of the homologs from gram-positive bacteria and archaebacteria. A detailed phylogenetic analysis based on Hsp70 sequences was also performed. In neighbor-joining and parsimony trees based on Hsp70 sequences, both mycoplasma species branched with the low-G+C-content gram-positive group of bacteria (e.g., *Lactobacillus* and *Erysipelothrix* species) in 87% and 96% of the bootstrap replicates, respectively, indicating their close evolutionary relationship to this group. The phylogenetic trees based on Hsp70 sequences show a polyphyletic branching of archaebacteria with the gram-positive species, which is statistically strongly favored.

Mycoplasma species are the smallest free-living microorganisms known. They lack a typical eubacterial cell wall, periplasmic space, and an outer membrane (20, 26, 27). Their genomes are far smaller than the typical genomes of prokaryotes (they are one-sixth to one-third the size of the *Escherichia coli* genome) and approach the theoretical size limit for genomes of free-living organisms (11, 20, 27). Mycoplasma species are distinct because of their genome size and their strict requirement for exogenous sterol (20, 26). The peculiar characteristics of mycoplasma species have aroused considerable interest concerning the evolution of these organisms. Phylogenetic studies based on 16S rRNA sequence data have indicated that the mycoplasmas branch with the low-G+C-content subgroup of gram-positive eubacteria, namely, the *Lactobacillus* group (20, 30, 32). Due to the limited genome size, many gene functions are absent from mycoplasmas; however, heat shock treatment of *Mycoplasma capricolum* cells has been reported to cause increased synthesis of seven distinct protein bands, including a 70-kDa protein which reacts with antibodies to the *E. coli* DnaK protein (5). In our recent work we have been utilizing sequence data for the Hsp70 family of proteins to examine the evolutionary relationships among different groups of species (9, 14–16). Because of their large size and high degree of sequence conservation (the most conserved proteins found in any species) the Hsp70 proteins are well suited for examining deep phylogenetic relationships. Phylogenetic analyses based on Hsp70 sequences and several other protein sequences have revealed a number of important differences from rRNA-based phylogenies (1, 2, 4, 12, 14, 15, 29). In particular, a close evolutionary relationship between archaebacteria and the gram-positive group of bacteria has been suggested by these studies (12, 15). In view of these results and the limited genome size of mycoplasmas, we cloned the *hsp70* gene from *M. capricolum* to confirm the evolutionary relationship of this organism with other species. In this paper we present the results of cloning and sequencing studies of *hsp70* and a number of adjoining genes from *M. capricolum*. Recently, the complete genomic sequence of another mycoplasma species, *Mycoplasma genitalium*, was reported (11). The evolutionary relationship of mycoplasmas to other prokaryotic species was examined in this study.

* Corresponding author. Phone: (905) 525-9140, ext. 22639. Fax: (905) 522-9033.

MATERIALS AND METHODS

PCR. Genomic DNA from *M. capricolum* was kindly provided by P. M. Gillevet, George Mason University, Fairfax, Va. Degenerate oligonucleotide primers with opposite orientations for two conserved regions of Hsp70 sequences were synthesized at the Central Facility of the Institute of Molecular Biology, McMaster University, Hamilton, Canada. A forward primer (5'-CARGCNAC NAARGAYGCNGG-3') and a reverse primer (5'-GCNACNGCYTCRTCNG GRTT-3') (where N = A, C, G, or T; Y = C or T; and R = A or G) were made for sequence regions QATKDAG and NPDEAVA, respectively (13). PCR amplification in which *M. capricolum* genomic DNA and the primers described above were used was carried out as described previously (13). After 30 cycles, an expected 0.65-kb fragment which showed specific amplification was cloned and sequenced to ascertain its Hsp70-related nature and uniqueness. The cloned PCR fragment was used as a probe in Southern blots and colony-screening studies. To isolate a genomic clone for Hsp70, the genomic DNA was digested with a number of restriction enzymes and resolved on a 1% agarose gel. The region which hybridized to the PCR probe in the Southern blots was excised, and a library made from this DNA was screened with the ^{32}P-labelled cloned *M. capricolum* PCR probe. The DNA was completely sequenced after nested sets of deletions were made with a T7 sequencing kit (Pharmacia).

Phylogenetic analyses. The GenBank accession numbers of various Hsp70 homologs have been published previously (15). A multiple alignment of the Hsp70 family of sequences has been described previously (9, 13, 16), and the sequences for mycoplasma Hsp70s were added to this alignment. A phylogenetic analysis was carried out by using the sequence region corresponding to the region between amino acid 6 (beginning with IIGID) and amino acid 512 (ending with KNKAE) in the *M. capricolum* sequence, which could be aligned with minimum ambiguity in all Hsp70 homologs (9, 14–16). Phylogenetic analyses of sequence data were performed by using the programs BOOT, NEIGHBOUR, PROTPARS, PROTDIST, and CONSENSE from the PHYLIP version 3.5 program package (10, 28).

RESULTS

Cloning of the *hsp70* gene. The *hsp70* gene region from *M. capricolum* was cloned by using a PCR-based approach that has been successfully employed in our laboratory for cloning *hsp70/dnaK* from a variety of prokaryotic and eukaryotic spe-

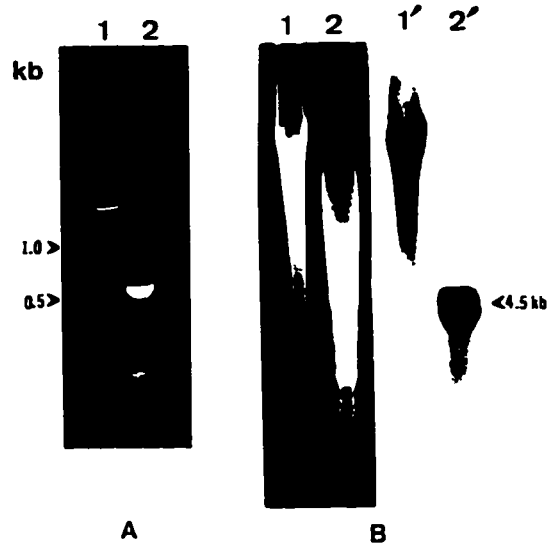


FIG. 1. Cloning of the *M. capricolum hsp70* gene. (A) Agarose gel electrophoresis of the PCR product of *M. capricolum* genomic DNA when degenerate oligonucleotide primers for conserved regions of *hsp70* genes were used. Lane 1, molecular size markers; lane 2, amplified 0.65-kb PCR fragment used as a probe in panel B. (B) Southern blot analyses of genomic DNA. Lanes 1 and 2, DNA digested with *Bam*HI and *Eco*RI, respectively, and stained with ethidium bromide; lanes 1′ and 2′, results of hybridization of lanes 1 and 2, respectively, with the PCR probe.

cies (9, 13–16). PCR amplification with primers for two highly conserved regions of Hsp70 led to the isolation of a 0.65-kb fragment (Fig. 1A), whose sequence was unique but showed a high degree of similarity to the Hsp70 family of proteins (data not shown). To isolate a genomic clone for Hsp70, *M. capricolum* genomic DNA was digested with *Bam*HI and *Eco*RI endonucleases and resolved on an agarose gel (Fig. 1B). Southern blot analyses in which the cloned PCR probe was used revealed specific hybridization to a 4.5-kb fragment for the *Eco*RI-digested DNA (Fig. 1B, lane 2′). To clone this fragment, the DNA region between 4 and 5 kb from *Eco*RI-digested DNA was subcloned in plasmid pGEM-7z(+). After transformation, the resultant library was screened by using the cloned PCR probe. The positive clones contained an approximately 4.5-kb insert which hybridized to the PCR probe. Sequencing of the insert revealed that some sequence from the 3′ end of the *hsp70* gene (200 bp) was missing. To clone the missing region, genomic DNA was digested with *Dra*I and hybridized to a probe from the 3′ end of the original clone (probe P2) (Fig. 2). A 1.8-kb fragment which hybridized to this probe was cloned, and it contained the 3′ end of the *hsp70* gene. The 4.5- and 1.8-kb fragments were completely sequenced. Translation of the 5.1-kb region that was spanned by these two fragments revealed that in addition to the *hsp70* gene sequence, this region also contained either complete or partial sequences for the following genes: *clpB*, *orfA*, *grpE*, and *dnaJ*. The nucleotide sequence and the open reading frames corresponding to the above proteins that are present in this fragment are shown in Fig. 3. This sequence has been deposited in GenBank database under accession number U51235. A partial sequence of some of these genes from *M. capricolum* was also reported recently by Bork et al. (3) as part of its genomic sequence.

The G+C content of the *M. capricolum* DNA sequence region was found to be 25.8 mol%, which is consistent with the previously published value (26). The codon usage in the coding region revealed that there was a strong preference for A or T at the third position. Another novel feature of the codon usage is that the codon UGA, which normally codes for a stop codon, was found to code for tryptophan. This unusual use of UGA, as well as a number of other codons (e.g., CGG is an unassigned codon instead of coding for arginine), in *M. capricolum* is well known from earlier studies (23, 25) and appears to have resulted from A/T-induced selection pressure.

Sequence alignment and phylogenetic analyses. By the time that we completed the cloning and sequencing of the *hsp70* gene region from *M. capricolum*, the sequence of another mycoplasma, *M. genitalium*, became available as a result of cloning of the entire genome of this species (11). *M. genitalium* is a member of the *Mycoplasma pneumoniae* group and is thought to have the smallest genome (580 kb) among eubacterial species (11, 20). Therefore, both mycoplasma Hsp70 sequences were utilized in further analyses. A multiple alignment of the

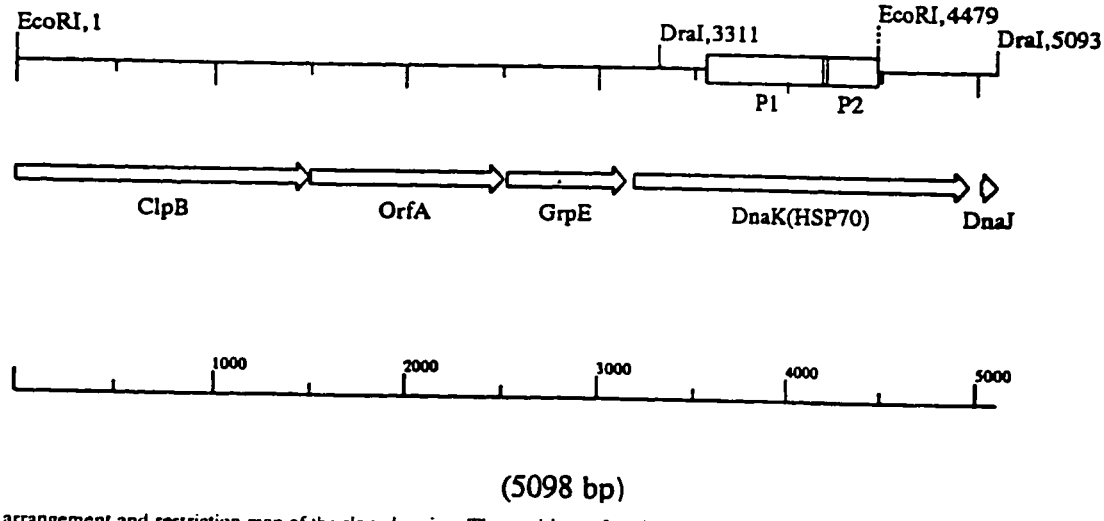


FIG. 2. Gene arrangement and restriction map of the cloned region. The positions of probes P1 and P2, which were used to isolate the *Eco*RI-*Eco*RI (4.5-kb) and *Dra*I-*Dra*I (1.8-kb) fragments, respectively, are shown. The arrows indicate the open reading frames corresponding to the *clpB*, *orfA*, *grpE*, *hsp70*/*dnaK*, and *dnaJ* proteins.

```
      Clp B
1     GA ATT CAT CAT AAT GCC TTA GTT TCA GCT GCA AAG TTA TCT AGT AGA TAT ATA ACT GAT AGA TAT TTA CCA GAC AAA GCT ATT GAT TTA
         Ile His His Asn Ala Leu Val Ser Ala Ala Lys Leu Ser Ser Arg Tyr Ile Thr Asp Arg Tyr Leu Pro Asp Lys Ala Ile Asp Leu

90    GTT GAT GAA GCT TGT GCT TCT ATT AAA ACA GAA TTA GCA AGT ATT CCA ATT GAA CTA GAT CAA GTA AAT AGA AAA GTA ATG CAA TTA GAA
      Val Asp Glu Ala Cys Ala Ser Ile Lys Thr Glu Leu Ala Ser Ile Pro Ile Glu Leu Asp Gln Val Asn Arg Lys Val Met Gln Leu Glu

180   ATT GAA ACT TCA GCT TTA GAA AAA GAA AAA GAT GAT AAA TCT AAA GAA AGA TGA CAA GAA GCT AAA AAA GAA TTA GAT AGT TTA AAA ATT
      Ile Glu Thr Ser Ala Leu Glu Lys Glu Lys Asp Asp Lys Ser Lys Glu Arg Trp Gln Glu Ala Lys Lys Glu Leu Asp Ser Leu Lys Ile

270   GAA CAA GCT ACT TTA AAT AAA AAG TGA GAA AAA GAA AAA GAA GAA TTA AGT AAA ATT AAT TTG GTA AAA TCA AGC ATT GAA AAT TTA AAA
      Glu Gln Ala Thr Leu Asn Lys Lys Trp Glu Lys Glu Lys Glu Glu Leu Ser Lys Ile Asn Leu Val Lys Ser Ser Ile Glu Asn Leu Lys

360   CAA GAA TTA GAA ACA GCT CAA AAC GAT GGA AAT TAT AAA AGA GCT GGA GAA ATT AAA TAC TCA TTA TTA CCA TCA CTT GAA AAA AGC TTA
      Gln Glu Leu Glu Thr Ala Gln Asn Asp Gly Asn Tyr Lys Arg Ala Gly Glu Ile Lys Tyr Ser Leu Leu Pro Ser Leu Glu Lys Ser Leu

450   GCT TTA TTT GAA ACA CAA ACT GGA GCA AAA ATG ATT TCA GAA GAA GTA ACT GAA CAA GAA ATT GCA AAA GTT GTT TCT AAA TCA ACA GGA
      Ala Leu Phe Glu Thr Gln Thr Gly Ala Lys Met Ile Ser Glu Glu Val Thr Glu Gln Glu Ile Ala Lys Val Val Ser Lys Ser Thr Gly

540   ATT TTA GTT GAT AAA TTA ATT TCT TCA GAA AAA GAA AGA CTT TTA AAT CTT GAA GAT CTA TTA AAA AAA TAT GTT AAA GGT CAA GAT CAA
      Ile Leu Val Asp Lys Leu Ile Ser Ser Glu Lys Glu Arg Leu Leu Asn Leu Glu Asp Leu Leu Lys Lys Tyr Val Lys Gly Gln Asp Gln

630   GCT ATT AAA GCA GTA ACT TCA GCA ATT ATG AGA AGT AGA AGT GGA ATT AAA AAT CCA GAT AAA CCA ATT GGT AGT TTT CTA TTT TTC GGA
      Ala Ile Lys Ala Val Thr Ser Ala Ile Met Arg Ser Arg Ser Gly Ile Lys Asn Pro Asp Lys Pro Ile Gly Ser Phe Leu Phe Phe Gly

720   CCA ACT GGA GTT GGA AAA ACT GAA GTT GCA AGA AGT TTA GCA GAT ATT TTA TTT AAT TCA CCG AAA AAA ATG ATC AGA CTT GAT ATG AGT
      Pro Thr Gly Val Gly Lys Thr Glu Val Ala Arg Ser Leu Ala Asp Ile Leu Phe Asn Ser Pro Lys Lys Met Ile Arg Leu Asp Met Ser

810   GAA TAT ATG GAA AAG CAT TCT GTT GCT AAA TTG ATT GGT GCT CCT CCT GGA TAT GTT GGG TAT GAA GAA GGA GGA AGA TTA ACT GAA GCT
      Glu Tyr Met Glu Lys His Ser Val Ala Lys Leu Ile Gly Ala Pro Pro Gly Tyr Val Gly Tyr Glu Glu Gly Gly Arg Leu Thr Glu Ala

900   GTA AGA AGA AAT CCT TAT TCA ATT ATT TTA TTT GAT GAA ATT GAA AAA GCT CAT AGT GAT GTG TTT AAC ATC TTA TTA CAA ATA CTA GAT
      Val Arg Arg Asn Pro Tyr Ser Ile Ile Leu Phe Asp Glu Ile Glu Lys Ala His Ser Asp Val Phe Asn Ile Leu Leu Gln Ile Leu Asp

990   GAT GGA AGA TTA ACA GAT TCA TTA GGA AAA ACT ATT GAT TTT AAA AAT ACA ATT ATT GTT ATG ACT TCA AAT ATA GCT AGT CAA TAT TTA
      Asp Gly Arg Leu Thr Asp Ser Leu Gly Lys Thr Ile Asp Phe Lys Asn Thr Ile Ile Val Met Thr Ser Asn Ile Ala Ser Gln Tyr Leu

1080  TTA ACT TCA GAT GAT TTT GTA CAA ATT GAT GAT CAA AAG ATT CAA GCA GAA TTA AAT CAA ACT TTT AGG CCT GAA TTT TTA AAT AGA ATT
      Leu Thr Ser Asp Asp Phe Val Gln Ile Asp Asp Gln Lys Ile Gln Ala Glu Leu Asn Gln Thr Phe Arg Pro Glu Phe Leu Asn Arg Ile

1170  GAT AAT ATT GTT TAT TTT AAT GCT TTA TCA GTA CAA ACA ATT GGC GAA ATT GTA GAT AAA CTT TTA GAT GAA TTA ATT ACA AGA TTA CAA
      Asp Asn Ile Val Tyr Phe Asn Ala Leu Ser Val Gln Thr Ile Gly Glu Ile Val Asp Lys Leu Leu Asp Glu Leu Ile Thr Arg Leu Gln

1260  GAT GAA CAA AAC TAT TTT ATT AAT TTT TCA GAA GAA GCT AGA AAT AAA ATT ATT AAT GAA GGT TAT GAT AGG TTA TTT GGA GCA AGA CTA
      Asp Glu Gln Asn Tyr Phe Ile Asn Phe Ser Glu Glu Ala Arg Asn Lys Ile Ile Asn Glu Gly Tyr Asp Arg Leu Phe Gly Ala Arg Leu

1350  TTA AAA GAT ATA TTG AAA AAA ATA ATT GAA ACT TTA ATA GCG CAT TAT ATT ATT AGT GGT TTG ATA AGT GAA AGT ACT AGA TAT CTA ATT
      Leu Lys Asp Ile Leu Lys Lys Ile Ile Glu Thr Leu Ile Ala His Tyr Ile Ile Ser Gly Leu Ile Ser Glu Ser Thr Arg Tyr Leu Ile
                                                                  Stop      OrfA
1440  GAT GTT AAA AAT AAC CAA TTT ATT TTA GAA GAA TTT AAA CAA TTT AAT TAA TTT T   ATG TTA ACT AAA AGA CAA GTT AAA ATT TTA CAA
      Asp Val Lys Asn Asn Gln Phe Ile Leu Glu Glu Phe Lys Gln Phe Asn ---       MET Leu Thr Lys Arg Gln Val Lys Ile Leu Gln

1528  ACT ATT GTT GAA GAA TTT ATT AAA ACT AAT CAA CCA GTT GGA TCG AAA AGA ATT TTA GAA CTA TTA GAT ATT AAA ATA TCT TCA GCA ACA
      Thr Ile Val Glu Glu Phe Ile Lys Thr Asn Gln Pro Val Gly Ser Lys Arg Ile Leu Glu Leu Leu Asp Ile Lys Ile Ser Ser Ala Thr

1618  ATT AGA AAT GAA TCA GCT ATT TTA GAA CAT GAA GGT TAT TTA GAA AAA CAA CAC ACT TCA AGT GGA AGA ACA CCT TCA ACT AAA GGC TAT
      Ile Arg Asn Glu Ser Ala Ile Leu Glu His Glu Gly Tyr Leu Glu Lys Gln His Thr Ser Ser Gly Arg Thr Pro Ser Thr Lys Gly Tyr

1708  AGA TAT TAT GTT GAT AAT ATT ATG AAA CTA GAT TCA GCT GAT TAT ACT AGA TTA AAA ATT TAT TTA AAC CAG TTA TTA GAT TTA AGA AAA
      Arg Tyr Tyr Val Asp Asn Ile Met Lys Leu Asp Ser Ala Asp Tyr Thr Arg Leu Lys Ile Tyr Leu Asn Gln Leu Leu Asp Leu Arg Lys

1798  TAT GAT ATT GAT AAA ACG ATT AAT TAT GCT AGT GAA ATT ATT AGT GAA TTA ACT AAA ATG ACA GCA GTT GTA ATT AAA AGC AAA AAT ATA
      Tyr Asp Ile Asp Lys Thr Ile Asn Tyr Ala Ser Glu Ile Ile Ser Glu Leu Thr Lys Met Thr Ala Val Val Ile Lys Ser Lys Asn Ile

1888  AAA AAT ATA AAG TTA AAA AAA ATT GAA CTA ATA TTA TTA TCA GAA TTT TTA GCA AGT GTA TTA TTT ATT TTT TCT GAT GGT GAT GTG CAA
      Lys Asn Ile Lys Leu Lys Lys Ile Glu Leu Ile Leu Leu Ser Glu Phe Leu Ala Ser Val Leu Phe Ile Phe Ser Asp Gly Asp Val Gln

1978  AAT AAA ATG TTT AAT TTA AAA GAT ATT TCT TTA TCT GAT TTA AAA ATT GCT ATT AAA TTA TTT TCA GAT TTT TTA GTT GAT GTT AAA TTA
      Asn Lys Met Phe Asn Leu Lys Asp Ile Ser Leu Ser Asp Leu Lys Ile Ala Ile Lys Leu Phe Ser Asp Phe Leu Val Asp Val Lys Leu

2068  GAT GAA ATA GAT CAA TAT TTA AAT GAT TTA AAG CAT CAA TTA TCT TTA AGT ATT AAA CAA TAT GAC TAT GTT TTA AAC ACA TTT ATA AAT
      Asp Glu Ile Asp Gln Tyr Leu Asn Asp Leu Lys His Gln Leu Ser Leu Ser Ile Lys Gln Tyr Asp Tyr Val Leu Asn Thr Phe Ile Asn
```

FIG. 3. Nucleotide sequence of the cloned region and deduced amino acid sequences of the various encoded proteins. The start points and the ends of different genes are indicated.

```
2158 ACT ATT TTA GAA TCA AAA AAT GAA CAA AAA GAA ACT CAT GGA ATG AGA TAT ATG TTA GAA AAT CCT GAG TTT AAT GAT ACT AAT AAA TTA
     Thr Ile Leu Glu Ser Lys Asn Glu Gln Lys Glu Thr His Gly Met Arg Tyr Met Leu Glu Asn Pro Glu Phe Asn Asp Thr Asn Lys Leu

2248 AAA AAT GCA GTT AAA TTA GTT GAA CAA TTA TCT CCT TTT GAT TGA TTT AAT ATT GCT TAT GAA TCT AAT AAA AAT ATG AAT AAA ATA GCA
     Lys Asn Ala Val Lys Leu Val Glu Gln Leu Ser Pro Phe Asp Trp Phe Asn Ile Ala Tyr Glu Ser Asn Lys Asn Met Asn Lys Ile Ala

2338 ATT AAA ATT GGT AAT GAA ATC GAC CAA ATA AAT GAT GAT ATT TCA ATG ATT GCT ACA GAA TTA AAA ATT GGT AAT TCT TCT ACT GTT TTA
     Ile Lys Ile Gly Asn Glu Ile Asp Gln Ile Asn Asp Asp Ile Ser Met Ile Ala Thr Glu Leu Lys Ile Gly Asn Ser Ser Thr Val Leu
                                                                                                                  Stop
2428 ACT TTA GTA GGT CCA AAA AGA GGT AGA TTA CAA CCA AGT AAA CCA GTT AAT GAA CTT AAT TAT TGA AAT TAT TAA TGCAAAGGAG AATTAAA
     Thr Leu Val Gly Pro Lys Arg Gly Arg Leu Gln Pro Ser Lys Pro Val Asn Glu Leu Asn Tyr Trp Asn Tyr ---
     GrpE
2520 ATG ACT GAA GAA TTA AAA AAT AAA AAA AAT AAT AAA AAC TAT TAT AGT CAA AAT AAA AAT AAA ACT AAA GCT GAA TTT CAA AAA CCT CAT
     Met Thr Glu Glu Leu Lys Asn Lys Lys Asn Asn Lys Asn Tyr Tyr Ser Gln Asn Lys Asn Lys Thr Lys Ala Glu Phe Gln Lys Pro His

2610 GTT AAA AAA AAT CAA TAT TTA AAG TTA AAA ACA AAA CTT GAT ACT GCT TTA TTA GAA GTT CAA AAT TTA AAA GAT TTA AAT GAA ACT TTA
     Val Lys Lys Asn Gln Tyr Leu Lys Leu Lys Thr Lys Leu Asp Thr Ala Leu Leu Glu Val Gln Asn Leu Lys Asp Leu Asn Glu Thr Leu

2700 AAA AAA GAT ATT GAA TCT GAA AGA CAA CTT AAT TTA GCT GAA ATT AGT AAT TTA ACT AAA AAG TAT AAT CAA AAA GAA ATA GAA ATT CAA
     Lys Lys Asp Ile Glu Ser Glu Arg Gln Leu Asn Leu Ala Glu Ile Ser Asn Leu Thr Lys Lys Tyr Asn Gln Lys Glu Ile Glu Ile Gln

2790 AAA TAT GGC GCT AGT AAA TTA GCA AGA GAC TTA ATT CAA CCT TTA GAA ATT TTA AAA AAA GTT GTT AAT GCT CCT AAT AAT AAT GAA GTA
     Lys Tyr Gly Ala Ser Lys Leu Ala Arg Asp Leu Ile Gln Pro Leu Glu Ile Leu Lys Lys Val Val Asn Ala Pro Asn Asn Asn Glu Val

2880 GTT CTT CGA AAT GTT AAA GGG TTT GAA ATG ATT GTT AGT CAA ATA AAT AAC GTT TTA GAA TCA CAT CAT ATT AAA GCT ATG AAT GTT AAA
     Val Leu Arg Asn Val Lys Gly Phe Glu Met Ile Val Ser Gln Ile Asn Asn Val Leu Glu Ser His His Ile Lys Ala Met Asn Val Lys

2970 GTT GGT GAT ATG TTT AAT CCA CAT CTT CAT GAT GCT AAT GAA GCA GTT GAA TCA GAT GAA TAT AAA ACA AAT CAA ATT GTT GGT GTT TTA
     Val Gly Asp Met Phe Asn Pro His Leu His Asp Ala Asn Glu Ala Val Glu Ser Asp Glu Tyr Lys Thr Asn Gln Ile Val Gly Val Leu
                                                                                                                  Stop
3060 AGT GAT GGG TAT ATG ATC CAT GAT AAA GTA CTA ATT TAT GCA ATA GTC AAA GTT GCC AAA ATA AAT AAA AAC AAT AAT TAA AATAAAAAAA
     Ser Asp Gly Tyr Met Ile His Asp Lys Val Leu Ile Tyr Ala Ile Val Lys Val Ala Lys Ile Asn Lys Asn Asn Asn ---
                                                 dnaK
3151 TAGTAAAAAA AGATACAAGG AAGAAAATAA AACC   ATG GCA AAA GAA AAA ATT ATT GGA ATA GAT TTA GGA ACT ACT AAT TCA GTT GTT TCA GTT
                                             MET Ala Lys Glu Lys Ile Ile Gly Ile Asp Leu Gly Thr Thr Asn Ser Val Val Ser Val

3245 ATT GAA GGT GGT CAA CCT ATT ATT TTA GAA AAT CCT GAA GGT CAA AGA ACT ACA CCA AGT GTT GTT GCT TTT AAA AAT TCA GAT ATT ATT
     Ile Glu Gly Gly Gln Pro Ile Ile Leu Glu Asn Pro Glu Gly Gln Arg Thr Thr Pro Ser Val Val Ala Phe Lys Asn Ser Asp Ile Ile

3335 GTT GGT GGA GCT GCT AAA CGT CAA GCT GTA ACT AAT CCA AAT GTT GTT CAA TCA ATA AAA TCA AAA ATG GGA ACT ACT TCT AAA GTT AAT
     Val Gly Gly Ala Ala Lys Arg Gln Ala Val Thr Asn Pro Asn Val Val Gln Ser Ile Lys Ser Lys Met Gly Thr Thr Ser Lys Val Asn

3425 TTA GAA GGA AAA GAT TAT AGT CCA GAA CAA ATT TCA GCT GAA ATT TTA AGA TAT ATG AAA AAT TAT GCT GAA GCT AAA TTA GGA CAA AAA
     Leu Glu Gly Lys Asp Tyr Ser Pro Glu Gln Ile Ser Ala Glu Ile Leu Arg Tyr Met Lys Asn Tyr Ala Glu Ala Lys Leu Gly Gln Lys

3515 GTA ACT AAA GCT GTT ATT ACA GTT CCT GCT TAT TTT AAT GAT GCA CAA AGA AAA GCT ACA AAA GAT GCT GGA ACA ATT GCA GGA TTA CAA
     Val Thr Lys Ala Val Ile Thr Val Pro Ala Tyr Phe Asn Asp Ala Gln Arg Lys Ala Thr Lys Asp Ala Gly Thr Ile Ala Gly Leu Gln

3605 GTT GAA AGA ATT ATT AAT GAA CCA ACT GCC GCT GCT TTA GCT TAT GGA CTA GAT AAA CAA GAT AAA GAA GAA ACA ATT TTA GTT TAT GAT
     Val Glu Arg Ile Ile Asn Glu Pro Thr Ala Ala Ala Leu Ala Tyr Gly Leu Asp Lys Gln Asp Lys Glu Glu Thr Ile Leu Val Tyr Asp

3695 TTA GGA GGA GGA ACT TTT GAT GTT TCT ATT CTA GCT ATT GGT GGT GGA AGT TTT GAT GTT ATT GCA ACT AGT GGA AAT AAT AAA TTA GGT
     Leu Gly Gly Gly Thr Phe Asp Val Ser Ile Leu Ala Ile Gly Gly Gly Ser Phe Asp Val Ile Ala Thr Ser Gly Asn Asn Lys Leu Gly

3785 GGA GAT AAT TTT GAT GAA GAA ATT ATC AAA TGA TTA CTA GGT AAA ATT AAA GCT GAA TAC AAT ATT GAT TTA TCT AAA GAA AAA ATG GCT
     Gly Asp Asn Phe Asp Glu Glu Ile Ile Lys Trp Leu Leu Gly Lys Ile Lys Ala Glu Tyr Asn Ile Asp Leu Ser Lys Glu Lys Met Ala

3875 TTA CAA AGA TTA AAA GAT GAA GCT GAA AAA GCA AAA ATT AAT TTA TCT AGC CAA TTA GAA GTT GAA ATT AAT TTA CCA TTT ATT GCA ATG
     Leu Gln Arg Leu Lys Asp Glu Ala Glu Lys Ala Lys Ile Asn Leu Ser Ser Gln Leu Glu Val Glu Ile Asn Leu Pro Phe Ile Ala Met

3965 AAT GAA AGT GGA CCA ATT TCT TTT GCA ACT CTA ACA AGA AGT GAA TTT AAC AAA ATT ACA AAA CAT TTA GTT GAC TTG ACT ATT CAA CCA
     Asn Glu Ser Gly Pro Ile Ser Phe Ala Thr Leu Thr Arg Ser Glu Phe Asn Lys Ile Thr Lys His Leu Val Asp Leu Thr Ile Gln Pro

4055 GTT AAA GAT GCT TTA AGT GCT GCT AAA AAA ACT CCA AGT GAA ATT AAT GAA GTT TTA TTA GTT GGT GGG TCA ACA AGA ATA CCT GCT GTT
     Val Lys Asp Ala Leu Ser Ala Ala Lys Lys Thr Pro Ser Glu Ile Asn Glu Val Leu Leu Val Gly Gly Ser Thr Arg Ile Pro Ala Val

4145 CAA GAA TTA GTA AAA AGT TTA TTA AAT AAA GAA CCA AAT AGA TCA ATT AAT CCA GAT GAA GTT GTT GCT ATG GGT GCT GCT GTG CAA GGT
     Gln Glu Leu Val Lys Ser Leu Leu Asn Lys Glu Pro Asn Arg Ser Ile Asn Pro Asp Glu Val Val Ala Met Gly Ala Ala Val Gln Gly

4235 GGA GTT TTA GCT GGT GAA GTT ACT GAT ATT TTA TTA TTA GAT GTA ACC CCA TTA TCA TTA GGT ATT GAA ACA ATG GGT GGA GTT ATG ACA
     Gly Val Leu Ala Gly Glu Val Thr Asp Ile Leu Leu Leu Asp Val Thr Pro Leu Ser Leu Gly Ile Glu Thr Met Gly Gly Val Met Thr
```

FIG. 3—*Continued.*

two mycoplasma Hsp70 sequences, various prokaryotic Hsp70 sequences, and a few representative eukaryotic Hsp70 sequences was obtained. This sequence alignment was very similar to that reported previously, with all of the Hsp70 homologs exhibiting extensive similarity throughout their lengths (9, 16). *M. capricolum* Hsp70 exhibited >50% sequence identity with various prokaryotic homologs, and the minimum level of identity observed for the eukaryotic homologs was in the range of

```
4325 AAA TTA ATT GAA AGA AAT ACT ACA ATT CCA GCA AAA AGA ACT CAA ATT TTT TCA ACT GCA ACA GAT AAT CAA CCA GCT GTT GAT ATT AAT
     Lys Leu Ile Glu Arg Asn Thr Thr Ile Pro Ala Lys Arg Thr Gln Ile Phe Ser Thr Ala Thr Asp Asn Gln Pro Ala Val Asp Ile Asn

4415 GTT TTA CAA GGT GAA AGA GCA ATG GCA GCT GAT AAT AAA TCA TTA GGT CAA TTC CAA CTA ACA GGA ATT CAA CCA GCT CCT AGA GGT ATT
     Val Leu Gln Gly Glu Arg Ala Met Ala Ala Asp Asn Lys Ser Leu Gly Gln Phe Gln Leu Thr Gly Ile Gln Pro Ala Pro Arg Gly Ile

4505 CCA CAA ATT GAA GTT ACT TTT GAA ATT GAT GCT AAT GGT ATT GTA AGT GTT TCA GCA AAA GAT AAA AAT ACT AAT GAA GAA AAA ACT ATT
     Pro Gln Ile Glu Val Thr Phe Glu Ile Asp Ala Asn Gly Ile Val Ser Val Ser Ala Lys Asp Lys Asn Thr Asn Glu Glu Lys Thr Ile

4595 ACT ATT TCA AAT TCA GGA AAT TTA AGT GAA GCT GAA GTT GAA AGA ATG ATA AAA GAA GCT CAA GAA AAT GCT GCA AAT GAT GAA GCT AAG
     Thr Ile Ser Asn Ser Gly Asn Leu Ser Glu Ala Glu Val Glu Arg Met Ile Lys Glu Ala Gln Glu Asn Ala Ala Asn Asp Glu Ala Lys

4685 AAA AAA AAT ATT GAA TTA AAA AAT AAA GCT GAA AAC TAT ATT AAC ATT ATT GAA ACT TCA CTA TTA CAA GCT GGT GAT AAA ATT AGT GCT
     Lys Lys Asn Ile Glu Leu Lys Asn Lys Ala Glu Asn Tyr Ile Asn Ile Ile Glu Thr Ser Leu Leu Gln Ala Gly Asp Lys Ile Ser Ala

4775 GAA CAA AAA GAA CAA TCA CAA AAA ATG ATT GAT GAA ATT AAA GAA CTA GTT AAA AAT GAA AAC TAT GAA GCT TTA GAA CAA AAA ATG GCT
     Glu Gln Lys Glu Gln Ser Gln Lys Met Ile Asp Glu Ile Lys Glu Leu Val Lys Asn Glu Asn Tyr Glu Ala Leu Glu Gln Lys Met Ala

4865 GAA CTA GAA CAA GCA ATG GCT GCA GCT GCT GAA TTT GCT AAC AAA CAT AAT GAT TCA GAT TCA AAT AAT AAT TCA TCA GAA CAA AAT AAT
     Glu Leu Glu Gln Ala Met Ala Ala Ala Ala Glu Phe Ala Asn Lys His Asn Asp Ser Asp Ser Asn Asn Asn Ser Ser Glu Gln Asn Asn
     Stop                                                                                                    dnaJ
4955 TAA TCACACTAAA TATATAAAAA CTAACATTGG TTAGTTTTTG TTTTTATAGA AAAGAAGAGA GAAT ATG AAA AAA AAG GAT TAT TAT GAA GTT TTA GGT
     ---                                                                        MET Lys Lys Lys Asp Tyr Tyr Glu Val Leu Gly

5055 GTA TCA AAA ACT GCT AGT GAA CAA GAA ATC AGA CAA GCT TTA AA 5098
     Val Ser Lys Thr Ala Ser Glu Gln Glu Ile Arg Gln Ala Leu
```

FIG. 3—*Continued.*

46 to 47% over the entire length of the protein. A number of Hsp70 family protein sequence signatures that are characteristic of the major groups of species have been identified previously (13–16). One prominent sequence signature is a large insert of 23 to 27 amino acids that is present in the N-terminal quadrant of all gram-negative bacteria and eukaryotic Hsp70, but absent in the homologs from archaebacteria and gram-positive bacteria. A portion of the sequence alignment that contains this sequence region is shown in Fig. 4. As Fig. 4 shows, as in all of the gram-positive bacteria and archaebacteria, this insert was also absent in the Hsp70 homologs from *M. capricolum* and *M. genitalium*. The lack of this insert in the mycoplasma Hsp70 homologs strongly indicates that there is an evolutionary linkage of these species to the gram-positive eubacteria and archaebacteria. To further verify the evolutionary relationships between mycoplasma and other species, phylogenetic analyses of Hsp70 sequence data were carried out. In these studies 31 Hsp70 sequences were used, and these sequences included 3 sequences from archaebacteria, 10 sequences from gram-positive eubacteria, 7 sequences from gram-negative eubacteria, 3 sequences of eukaryotic organelle homologs, and 6 sequences of eukaryotic cytosolic and endoplasmic reticulum resident homologs.

A neighbor-joining tree based on these sequences is shown in Fig. 5A. The bootstrap scores for various nodes (based on 100 replicates) are also indicated in Fig. 5A. As in our earlier work (9, 14, 15), the consensus tree based on Hsp70 sequences shows a clear distinction between the eukaryotic and prokaryotic species (in 100% of the bootstrap replicates) and a fairly good distinction (in 87% of the bootstrap replicates) between the eukaryotes and gram-negative bacteria on one end and the archaebacteria and gram-positive bacteria on the other end (Fig. 5A). The two mycoplasma species grouped together, and they branched with the low-G+C-content gram-positive bacteria in 85% of the bootstrap replicates.

Parsimony analysis of the sequences described above yielded a single most parsimonious tree requiring a total of 3,965 steps (Fig. 5B). Bootstrap scores for some of the important nodes in this tree are shown in Fig. 5B. As Fig. 5B shows, branching of the archaebacteria with the gram-positive species and branching of the eukaryotic cytosolic homologs with the gram-negative bacteria were strongly supported in 100% of the bootstrap replicates. The two mycoplasma species again showed a strong affinity for the low-G+C-content gram-positive bacteria, and the bootstrap score of this branching (96 of 100) indicated that this relationship was robust. In the parsimony tree the closest affiliation of the mycoplasma species was with *Erysipelothrix rhusiopathiae*, which is a human and animal pathogen similar to many mycoplasma species (19). Another difference between the neighbor-joining and parsimony trees is the difference in the branching positions of the eukaryotic cytosolic homologs. The parsimony tree indicates that there is a specific relationship between these homologs and the cyanobacterium-chloroplast group (which has also been observed previously [14, 15], whereas in the neighbor-joining tree a deeper branching of the latter group was observed.

DISCUSSION

In this paper we describe the cloning and a phylogenetic analysis of an Hsp70/DnaK homolog from *M. capricolum*. The 5.0-kb region surrounding the *hsp70* locus that was sequenced in this study has been found to contain either partial or complete sequences of a number of different genes. The genes that are present in the cloned fragment occur in the following order: *clpB* (1.5 kb; partial)-*orfA* (1.0 kb)-*grpE* (0.62 kb)-*hsp70* (1.8 kb)-*dnaJ* (≈0.1 kb; partial). The observed organization of these genes in *M. capricolum* is similar to the organization seen in several gram-positive bacteria (viz., *Bacillus subtilis* [31], *Clostridium acetobutylicum* [22], *Lactococcus lactis* [7], and *Staphylococcus aureus* [24]) and archaebacteria (viz., *Methanosarcina mazei* [19];]). In gram-positive bacteria such as *B. subtilis*, the last four genes (i.e., *orfA-grpE-hsp70* [*dnaK*]-*dnaJ*) form an operon (31), whereas in the gram-negative eubacteria (e.g., *E. coli*), an *orfA* gene homolog has not been identified.

The presence of an *hsp70* gene in *M. genitalium* and *M. capricolum*, which have minimal genomes (in *M. genitalium* a maximum of 470 protein-encoding sequences are present [11]), is consistent with our understanding that Hsp70 plays an essential role in all cell types (2, 21). In addition to Hsp70, the other genes that have been identified in the cloned region in *M. capricolum* (viz., *dnaJ*, *grpE*, *orfA*, and *clpB*) are also present in

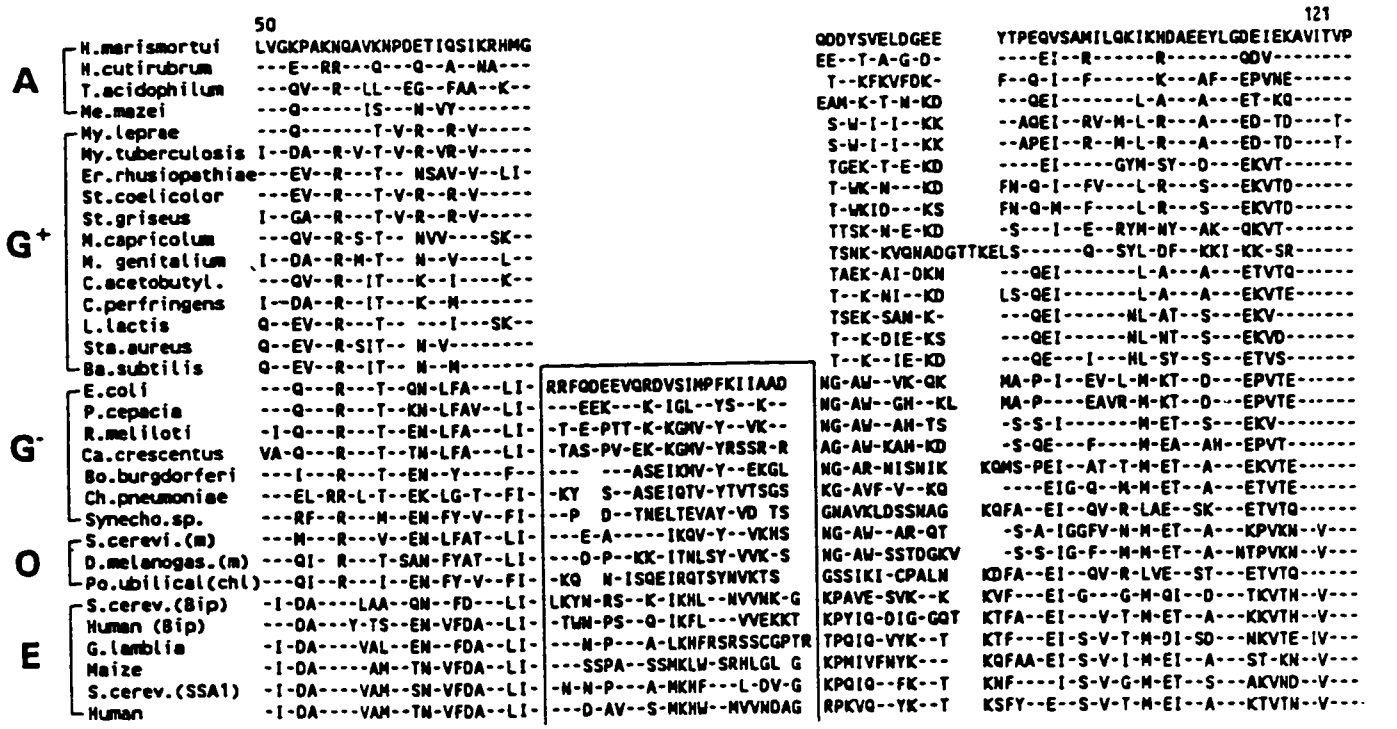


FIG. 4. Part of the Hsp70 sequence alignment showing the relatedness of mycoplasma species to the gram-positive bacteria and archaebacteria. A, E, G$^-$, G$^+$, and O refer to sequences from archaebacteria, sequences from eukaryotes (cytosolic), sequences from gram-negative bacteria, sequences from gram-positive bacteria, and eukaryotic organellar sequences (viz., mitochondrial and chloroplast sequences), respectively. The numbers at the top refer to positions in the *Halobacterium marismortui* sequence. The region enclosed in a box is the large insert in the N-terminal quadrant present in all Hsp70 homologs from gram-negative bacteria and eukaryotes. Dashes indicate identity with the amino acid in the top line. Abbreviations: Ba., *Bacillus*; Bo., *Borrelia*; C., *Clostridium*; C. acetobutyl., *Clostridium acetobutylicum*; Ca., *Caulobacter*; Ch., *Chlamydia*; D. melanogas., *Drosophila melanogaster*; E., *Escherichia*; Er., *Erysipelothrix*; G., *Giardia*; H., *Halobacterium*; L., *Lactococcus*; M., *Mycoplasma*; Me., *Methanosarcina*; My., *Mycobacterium*; P., *Pseudomonas*; Po., ubilical, *Porphyra umbilicalis*; R., *Rhizobium*; S. cerev., *Saccharomyces cerevisiae*; St., *Streptomyces*; Sta., *Staphylococcus*; Synecho. sp., *Synechococcus* species; T., *Thermoplasma*; m, chl, and Bip, mitochondrial, chloroplast, and endoplasmic reticulum resident forms of Hsp70, respectively.

M. genitalium (11). However, in contrast to the contiguous arrangement of these genes in *M. capricolum* and other bacterial species, these genes have been reported to be widely dispersed in the *M. genitalium* genome, with no two genes exhibiting close proximity (11). The possible significance of this difference in gene organization in *M. genitalium* is not clear at this time.

The sequence features of Hsp70 and the phylogenetic analysis based on the sequence data provide strong evidence that mycoplasma species are evolutionarily closely related to the gram-positive bacteria. Earlier studies on Hsp70 sequences have identified a large insert (between 23 and 27 amino acids long) in the N-terminal quadrant that is present in all of the homologs from gram-negative bacteria and eukaryotic species, but absent in various gram-positive bacteria and archaebacteria (14–16). The sequence alignment of Hsp70 homologs shows that like the Hsp70 homologs of various other gram-positive bacteria and archaebacteria, the Hsp70 homologs of both mycoplasma species examined also lack this insert. This observation strongly indicates that the mycoplasmas are related to the gram-positive bacterium-archaebacterium group of species.

Detailed phylogenetic analysis based on Hsp70 sequences revealed that the mycoplasma species branch with the low-G+C-content gram-positive bacterial group. The bootstrap scores for this affiliation in the neighbor-joining and parsimony trees indicate that this relationship is reliable. In the parsimony tree, the pathogenic bacterium *Erysipelothrix rhusiopathiae* was the closest relative of the mycoplasmas. The observed branching of the mycoplasmas with the low-G+C-content gram-positive eubacteria is consistent with the branching observed previously on 16S and 5S rRNA trees (6, 17, 20, 30, 32), the RecA protein sequences (8, 18), and a number of other characteristics (20). The fact that different gene and protein sequences, which differ in other important respects (8, 12, 5, 20, 30, 32), all support a similar evolutionary placement of mycoplasma species strongly indicates that the observed relationship is reliable and that it likely represents the species phylogeny.

One interesting and important difference between the Hsp70 phylogeny and the phylogeny based on 16S rRNA is that the former shows a polyphyletic branching of the archaebacteria with the gram-positive species (Fig. 5). We have previously shown that the observed polyphyletic branching pattern of archaebacteria is statistically strongly supported by results obtained with different phylogenetic methods (14, 15). Furthermore, as discussed in earlier work (12–14), the possibilities that the observed pattern could be due to the use of paralogous sequences, horizontal gene transfer, or tree reconstruction artifacts are considered highly unlikely. Since a similar branching of archaebacteria with gram-positive species is obtained with a number of other conserved proteins, including glutamate dehydrogenase (1) and glutamine synthetase I (4, 29), we think that these results may not be unusual but may point to a specific relationship between these two groups. Future studies with additional gene sequences should be helpful in clarifying this relationship.

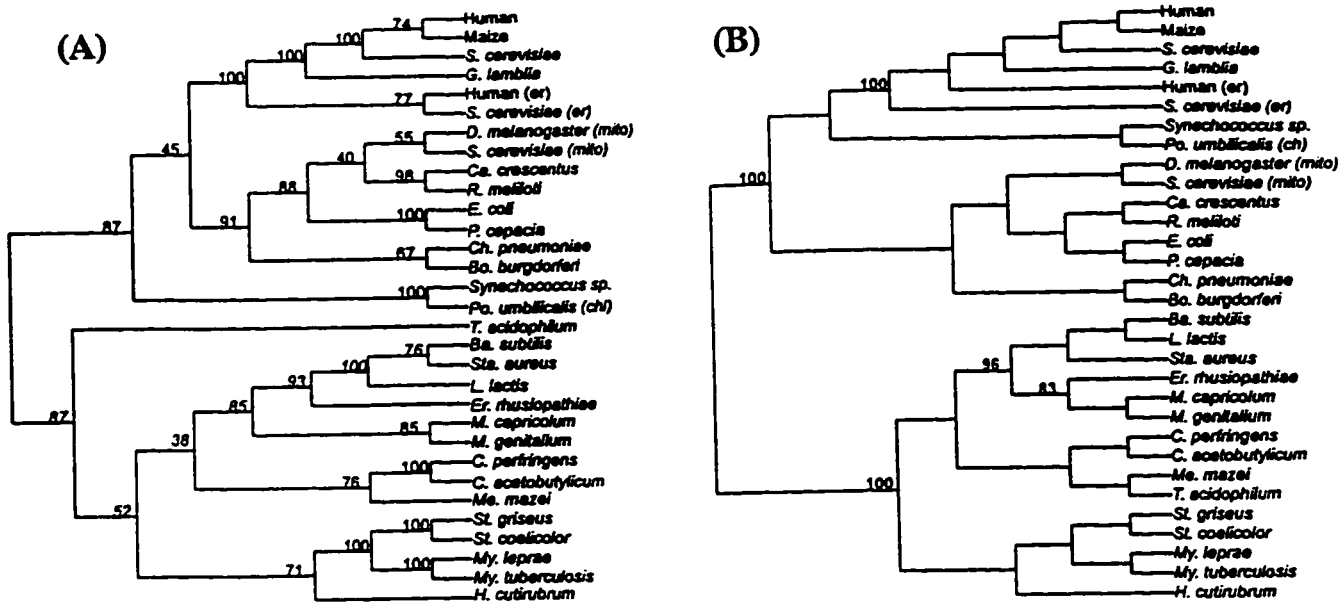


FIG. 5. Phylogenetic analysis based on Hsp70 protein sequences. (A) Neighbor-joining distance tree based on the sequence data. The distances were determined by using the PAM 250 matrix. The numbers at the forks indicate the percentages of times that the species to the right grouped together in 100 bootstrap replicates. For abbreviations see the legend to Fig. 4. (B) Maximum-parsimony tree based on the sequences. The bootstrap scores (out of 100) for some of the important nodes are shown. mito, chl, and er refer to the mitochondrial, chloroplast, and endoplasmic reticulum resident forms of Hsp70, respectively.

REFERENCES

1. **Benachenhou-Lahfa, N., P. Forterre, and B. Labedan.** 1993. Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching pattern of the archaebacteria in the universal tree of life. J. Mol. Evol. 36:335–346.
2. **Boorstein, W. R., T. Zeigelhoffer, and E. A. Craig.** 1994. Molecular evolution of the HSP70 multigene family. J. Mol. Evol. 38:1–17.
3. **Bork, P., C. Ouzounis, G. Cesari, R. Schneider, C. Sander, M. Dolan, W. Gilbert, and P. M. Gilbert.** 1995. Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. Mol. Microbiol. 16:955–967.
4. **Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle.** 1994. Evolutionary relationships of bacterial and archael glutamine synthetase genes. J. Mol. Evol. 38:566–576.
5. **Dascher, C. C., S. K. Poddar, and J. Maniloff.** 1990. Heat shock response in mycoplasma, genome-limited organisms. J. Bacteriol. 172:1823–1827.
6. **DeWachter, R., E. Huysmans, and A. Vandenberge.** 1985. 5S ribosomal RNA as a tool for studying evolution, p. 115–141. *In* K. H. Schleifer and E. Stackebrandt (ed.), Evolution of prokaryotes. Academic Press, Inc., New York, N.Y.
7. **Eaton, T., C. Shearman, and M. Gasson.** 1993. Cloning and sequence analysis of the *dnaK* gene region of *Lactococcus lactis* subsp. *lactis*. J. Gen. Microbiol. 739:3253–3264.
8. **Eisen, J. A.** 1995. The RecA protein as a model molecule for molecular systematics studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. 41:1105–1123.
9. **Falah, M., and R. S. Gupta.** 1994. Cloning of the *hsp70/dnaK* gene from *Rhizobium meliloti* and *Pseudomonas cepacia*: phylogenetic analyses of mitochondrial origin based on a highly conserved protein sequence. J. Bacteriol. 176:7748–7753.
10. **Felsenstein, J.** 1993. PHYLIP manual, version 3.5. University of Washington, Seattle.
11. **Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, J. C. Bult, A. R. Kerlavoge, G. Sutton, J. M. Kelly, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bottom, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison III, and J. C. Venter.** 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403.
12. **Golding, G. B., and R. S. Gupta.** 1995. Protein based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. 12:1–6.
13. **Gupta, R. S., and B. Singh.** 1992. Cloning of the HSP70 gene from *Halobacterium marismortui*: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model for the evolution of HSP70 gene. J. Bacteriol. 174: 4594–4605.
14. **Gupta, R. S., and G. B. Golding.** 1993. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria and eukaryotes. J. Mol. Biol. 37:573–582.
15. **Gupta, R. S., and B. Singh.** 1994. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biol. 4:1104–1114.
16. **Gupta, R. S., K. Aitken, M. Falah, and B. Singh.** 1994. Cloning of *Giardia lamblia* HSP70 homologs: phylogenies based on HSP70 and HSP90 sequences point to gene duplication events accompanying the origin of eukaryotic cells. Proc. Natl. Acad. Sci. USA 91:2895–2899.
17. **Hori, H., M. Sawada, S. Osawa, K. Murao, and H. Ishikura.** 1981. The nucleotide sequence of 5S rRNA from *Mycoplasma capricolum*. Nucleic Acids Res. 9:5407–5410.
18. **Karlin, S., G. M. Weinstock, and V. C. Brendel.** 1995. Bacterial classifications derived from RecA protein sequence comparisons. J. Bacteriol. 177:6881–6893.
19. **Macario, E. C., C. B. Dugan, and A. J. L. Macario.** 1994. Identification of a *grpE* heat shock gene homolog in the archaeon *Methanosarcina mazei*. J. Mol. Biol. 240:95–101.
20. **Maniloff, J., R. N. McElhaney, L. R. Finch, and J. B. Basemann (ed.).** 1992. Mycoplasmas: molecular biology and pathogenesis. American Society for Microbiology, Washington, D.C.
21. **Morimoto, R. I., A. Tissieres, and C. Georgopoulos (ed.).** 1994. The biology of heat shock proteins and molecular chaperones. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
22. **Narberhaus, F., K. Giebeler, and H. Bahl.** 1992. Molecular characterization of the *dnaK* gene region of *Clostridium acetobutylicum*, including *grpE*, *dnaJ*, and a new heat shock gene. J. Bacteriol. 174:3290–3299.
23. **Oba, T., Y. Andachi, A. Muto, and S. Osawa.** 1991. CGG: an unassigned or nonsense codon in *Mycoplasma capricolum*. Proc. Natl. Acad. Sci. USA 88:921–925.
24. **Ohta, T., K. Saito, M. Kuroda, K. Honda, H. Hirata, and H. Hayashi.** 1994. Molecular cloning of two new heat shock genes related to the *hsp70* genes in *Staphylococcus aureus*. J. Bacteriol. 176:4779–4783.
25. **Osawa, S., A. Muto, T. Ohama, Y. Andachi, R. Tanaka, and F. Yamao.** 1990. Prokaryotic genetic code. Experientia 46:1097–1106.
26. **Razin, S., and E. A. Freundt.** 1984. The mycoplasmas. p. 740–793. *In* N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. The Williams and Wilkins Co., Baltimore, Md.
27. **Razin, S.** 1985. Molecular biology and genetics of mycoplasmas. Microbiol. Rev. 49:419–455.
28. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method of reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.
29. **Tiboni, O., P. Cammarano, and A. M. Sanangelantoni.** 1993. Cloning and sequencing of the gene coding glutamine synthetase I from the archaeum

Pyrococcus woesi: anomalous phylogenies inferred from analysis of archael and bacterial glutamine synthetase I sequences. J. Bacteriol. 175:2961–2969.

30. **Weisburg, W. G., J. G. Tully, D. L. Rose, J. P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T. G. Lawrence, J. V. Etten, J. Maniloff, and C. R. Woese.** 1989. A phylogenetic analysis of mycoplasmas: basis for their classification. J. Bacteriol. 171:6455–6467.

31. **Wetzstein, M., U. Völker, J. Dedio, S. Löbau, U. Zuber, M. Schiesswohl, C. Herget, M. Hecker, and W. Schumann.** 1992. Cloning, sequencing, and molecular analysis of the *dnaK* locus from *Bacillus subtilis*. J. Bacteriol. 174:3300–3310.

32. **Woese, C. R., J. Maniloff, and L. B. Zablen.** 1980. Phylogenetic analysis of the mycoplasma. Proc. Natl. Acad. Sci. USA 77:494–498.

CHAPTER 3

Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: Implications regarding origin of eukaryotic cells and of endoplasmic reticulum

This chapter determines the phylogenetic placement of *Giardia lamblia* in the HSP70-based phylogeny. This entailed the cloning of HSP70 sequences from this species. Using a PCR-based approach, two HSP70-related fragments from *G. lamblia* DNA were isolated. I have worked on the isolation of a full-length genomic clone for one of the genes and the sequencing of both strands. Subsequent sequence analysis showed that the gene I sequenced corresponds to the eukaryotic cytosolic HSP70. The other gene that was sequenced by Karen Aitken and Dr. B. Singh based on its targeting and retention sequences is the ER form of HSP70 (Bip). The sequences studied in this chapter have been deposited in the GenBank data base under the accession nos. U04874 (cytosolic HSP70) and U04875 (ER HSP70).

Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 2895–2899, April 1994
Evolution

Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: Implications regarding origin of eukaryotic cells and of endoplasmic reticulum

(GRP78/Bip/molecular chaperones/archaebacteria/Gram-negative bacteria)

RADHEY S. GUPTA*, KAREN AITKEN, MIZIED FALAH, AND BHAG SINGH

Department of Biochemistry, McMaster University, Hamilton, ON, Canada, L8N 3Z5

Communicated by Ellis Englesberg, December 20, 1993 (received for review October 1, 1993)

ABSTRACT **The genes for two different 70-kDa heat shock protein (HSP70) homologs have been cloned and sequenced from the protozoan *Giardia lamblia*. On the basis of their sequence features, one of these genes corresponds to the cytoplasmic form of HSP70. The second gene, on the basis of its characteristic N-terminal hydrophobic signal sequence and C-terminal endoplasmic reticulum (ER) retention sequence (Lys-Asp-Glu-Leu), is the equivalent of ER-resident GRP78 or the Bip family of proteins. Phylogenetic trees based on HSP70 sequences show that *G. lamblia* homologs show the deepest divergence among eukaryotic species. The identification of a GRP78 or Bip homolog in *G. lamblia* strongly suggests the existence of ER in this ancient eukaryote. Detailed phylogenetic analyses of HSP70 sequences by boot-strap neighbor-joining and maximum-parsimony methods show that the cytoplasmic and ER homologs form distinct subfamilies that evolved from a common eukaryotic ancestor by gene duplication that occurred very early in the evolution of eukaryotic cells. It is postulated that because of the essential "molecular chaperone" function of these proteins in translocation of other proteins across membranes, duplication of their genes accompanied the evolution of ER or nucleus in the eukaryotic cell ancestor. The presence in all eukaryotic cytoplasmic HSP70 homologs (including the cognate, heat-induced, and ER forms) of a number of autapomorphic sequence signatures that are not present in any prokaryotic or organellar homologs provides strong evidence regarding the monophyletic nature of eukaryotic lineage. Further, all eukaryotic HSP70 homologs share in common with the Gram-negative group of eubacteria a number of sequence features that are not present in any archaebacterium or Gram-positive bacterium, indicating their evolution from this group of organisms. Some implications of these findings regarding the evolution of eukaryotic cells and ER are discussed.**

The ubiquitous 70-kDa heat shock protein (HSP70) family performs an essential "molecular chaperone" role in the intracellular trafficking of proteins and in diverse other cellular functions (see refs. 1–3). We have recently presented data showing that HSP70 constitutes the most conserved protein known to date that is present in all organisms (4). Global comparison and phylogenetic analyses of HSP70 sequence data have revealed an unexpectedly close evolutionary linkage between archaebacteria and the Gram-positive group of bacteria on one hand and the Gram-negative eubacteria and eukaryotic homologs on the other (4, 5). These studies also identified numerous shared sequence characteristics (or signatures) in HSP70s that either are distinctive of the main groups of organisms (e.g., prokaryotes vs. eukaryotes) or are uniquely shared by homologs from some of these groups (namely, common in Gram-positive bacteria and archaebacteria or between Gram-negative species and eukaryotic homologs). The presence of a number of unique sequence signatures in various known eukaryotic cytoplasmic homologs suggested to us that these were introduced at a very early stage in the origin of eukaryotic cells (4). If this hypothesis is correct then these signature sequences should also be found in HSP70 homologs from the earliest diverging eukaryotic organisms.

To test this hypothesis we undertook to clone HSP70 homologs from the protozoan *Giardia lamblia*, which lacks mitochondria (6) and which, on the basis of 16S rRNA phylogeny, constitutes the earliest diverging member within the eukaryotic lineage (7). In the present paper we describe the cloning and sequence comparison of two HSP70 homologs from *G. lamblia*.† One of these homologs apparently corresponds to the cytoplasmic form of the protein, whereas the other bears various characteristics of the form present in the endoplasmic reticulum (ER) (see refs. 3 and 8). Detailed comparison and phylogenetic analyses of these and other eukaryotic and prokaryotic HSP70 sequences presented here provide important insight into the origin of eukaryotic cells and of ER.

MATERIALS AND METHODS

G. lamblia Portland 1 strain was purchased from the American Type Culture Collection (catalogue no. 2088) and grown in Keister's modified TYI-S-33 medium (9) supplemented with 10% (vol/vol) inactivated fetal bovine serum. *G. lamblia* HSP70 genes were cloned by the procedures described previously (5, 10). The sequence data for the cytosolic and ER HSP70 clones have been deposited in the GenBank data base; the sequences for various other HSP70 homologs except for the *Thermoplasma acidophilum* sequence, which constitutes unpublished results of R.S.G. and B.S., were obtained from GenBank and Swiss-Prot data bases. The sequence alignment was carried out as described (4, 5). The phylogenetic analyses of sequence data were performed by using the programs BOOT, NEIGHBOR, CONSENSE, and PROTPARS from PHYLIP version 3.5 program package (11).

RESULTS

Cloning of *G. lamblia* HSP70 Homologs. The approach used to clone the HSP70 gene(s) involves PCR amplification of a segment of DNA that is flanked by two highly conserved regions in HSP70 (5, 10). The PCR fragments, after cloning and characterization, are then used as probe(s) for isolating

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: ER, endoplasmic reticulum; HSP70, 70-kDa heat shock family of proteins.

*To whom reprint requests should be addressed.

†The *G. lamblia* sequences reported in this paper have been deposited in the GenBank data base [accession nos. U04874 (cytosolic HSP70) and U04875 (ER HSP70)].

the genomic clone(s) (5, 10). In the case of *G. lamblia*, when individual clones containing the amplified fragments were sequenced, two different types of clones (representative HSG-4 and HSG-11), which showed extensive similarity to the HSP70 family of proteins, were obtained (Fig. 1). The cloned fragments of the two types were used to isolate clones from size-fractionated *Hin*dIII genomic library of *G. lamblia* DNA.

The clone HSG-12, obtained by using HSG-4 as the probe, contains a 1986-bp open reading frame, whose deduced amino acid sequence is shown in Fig. 1, line a. This protein contains a number of small insertions (at positions 120, 204, and 268) not found in any other homolog, indicating its more divergent character. There are several features indicating that this protein corresponds to the ER-resident form of HSP70 (commonly referred to as GRP78 or Bip; refs. 3 and 8). First, similar to other proteins imported into ER, it contains an N-terminal hydrophobic leader sequence before the highly conserved sequence GIDLGTTYSCV (positions 17–27 in Fig. 1, line a), present in all HSP70 homologs. Second, all ER-resident proteins contain a conserved sequence, XDEL, where X is usually K or H at the extreme C-terminal end (see ref. 8; signature c in Fig. 1). Since the last four amino acids in the open reading frame in HSG-12 are KDEL, they provide a strong indication that this protein corresponds to the ER homolog. Third, the sequence comparison presented in Fig. 1 shows that the ER-resident forms of HSP70s (including those whose sequences are not shown) differ from their cytoplasmic counterparts in a number of different positions, which can be used to distinguish between the two. As seen (see signatures c in Fig. 1), the protein encoded by HSG-12 possesses all unique amino acid substitutions that are characteristic of the ER HSP70 homologs. In view of these characteristics, this protein henceforth will be referred to as ER HSP70 or GRP78.

A second genomic clone, GL-15, obtained upon screening with the probe HSG-11 contains an open reading frame of 1992 bp (after ATG) encoding a protein 664 amino acids long. The deduced amino acid sequence of this clone is shown in Fig. 1, line f. Based on a number of different characteristics of this sequence, this apparently corresponds to the cytosolic form of HSP70. First, this sequence shows greater similarity to the cytosolic HSP70 than to either mitochondrial or ER homologs (not shown). Second, all eukaryotic cytosolic HSP70s contain a sequence related to the consensus sequence GPTIEEVD at their extreme C-terminal end (Fig. 1, signature d, and unpublished data). The protein encoded by GL-15 ends with the sequence GPSVDDLD, which is very similar to the above sequence. Third, the protein encoded by GL-15 contains various characteristic amino acids in different positions that distinguish cytosolic HSP70 homologs from those present in other compartments—namely, ER and mitochondria (Fig. 1, signatures a, c, and d). In view of these characteristics, this *G. lamblia* protein henceforward will be referred to as Cyto HSP70.

Sequence Comparisons and Signatures. A global comparison of *G. lamblia* HSP70 sequences with all known eukaryotic, eubacterial, and archaebacterial homologs was carried out. The results of these studies from a few representative species are presented in Fig. 1. The sequences shown include five ER homologs, four cytosolic HSP70s, one mitochondrial homolog, and one each from Gram-negative bacteria, Gram-positive bacteria, and archaebacteria. As seen, various HSP70 homologs exhibit extensive sequence similarity throughout their length, except in a small stretch near the C-terminal end. Previously, based on global comparison of HSP70 sequences, we have identified a number of sequence signatures (marked a in Fig. 1) that provide clear distinction between eukaryotic cytoplasmic and prokaryotic homologs (4, 5). As seen from Fig. 1, both of the *G. lamblia* HSP70 homologs contained all of the eukaryotic-specific signature sequences, thereby supporting our contention that these signatures are unique characteristics of all eukaryotic cytoplasmic homologs. It should be noted that mitochondria and chloroplast homologs display prokaryotic rather than eukaryotic sequence characteristics, which is in accordance with their endosymbiotic origin from eubacteria (5, 12).

An important sequence characteristic of HSP70 homologs is the presence of an insert of 23–27 amino acids in the N-terminal quadrant (boxed in Fig. 1), which is present in all eukaryotic homologs as well as those from Gram-negative eubacteria and eukaryotic organelles but is absent in HSP70s from archaebacterial and Gram-positive eubacterial species (4, 5). We have previously presented arguments that the species lacking this insert constitute the ancestral organisms (4, 5). Since this sequence feature has important implications concerning the origin of eukaryotic cells, a comparison of this region from a number of additional species, which include all of the known archaebacterial and Gram-positive HSP70 sequences, is presented in Fig. 2. As seen, this relatively conserved insert is present in all of the eukaryotic cytoplasmic homologs (including ER, heat shock-induced, and cognate forms), eukaryotic organeller homologs (namely, mitochondria and chloroplasts), and Gram-negative eubacterial homologs at precisely the same position in the sequence. In contrast, none of the HSP70 homologs from various archaebacteria (including a methanogen, halophile, and thermoacidophile) or Gram-positive eubacteria contained this sequence. In addition to this insert, the homologs from Gram-negative eubacteria and eukaryotes also shared other sequence features (marked b in Figs. 1 and 2) that are not present in the other two groups. The presence of these signature sequences in all eukaryotic HSP70 homologs is strongly indicative of their evolution from the Gram-negative group of bacteria.

Phylogenetic Analyses. To understand the relationships between the cytosolic and ER homologs and to determine the phylogenetic position of *G. lamblia* within the eukaryotic lineage, detailed phylogenetic analyses on HSP70 sequences were carried out. The analyses were carried out on a subset of 25 sequences, which included 11 ER HSP70s, 12 cytoplasmic HSP70s (both groups containing representatives from plant, animal, fungi, and protist species) and 1 each from bacteria and archaebacteria. The sequences from the latter species, because of their ancestral nature, could be used to root the eukaryotic HSP70 trees. A boot-strap, neighbor-joining consensus tree based on HSP70 sequences is shown in Fig. 3*A*. To obtain this, the sequences were "bootstrapped" 100 times, which is a method to place confidence limits on phylogenics (11); for each sequence a neighbor-joining tree was found (13), and finally a consensus tree based on these analyses was obtained. As seen, the ER homolog from *G. lamblia* showed the deepest branching amongst the eukaryotic homologs, and it formed the root of the eukaryotic tree. Following this, the rest of the eukaryotic homologs were separated into two distinct groups, one consisting of all the cytoplasmic homologs (100 times out of 100) and the other containing the various ER HSP70s. Cyto HSP70 from *G. lamblia* formed the deepest branch of the former group. Within each of these two groups, the various plants and animal species showed completely distinct branching (100 out of 100) from each other, but the relative branching orders of species within these groups were not reliable. Maximum parsimony analysis of the HSP70 sequence data gave rise to a single, most parsimonius tree requiring 2801 amino acid substitutions. The parsimony tree (Fig. 3*B*) is very similar to the boot-strap neighbor-joining tree (except for much earlier branching of the yeast ER homolog) and a clear distinction between ER and cytoplasmic homologs is again observed. The *G. lamblia* homologs again showed deepest branching within the two groups. These results strongly indicate that the

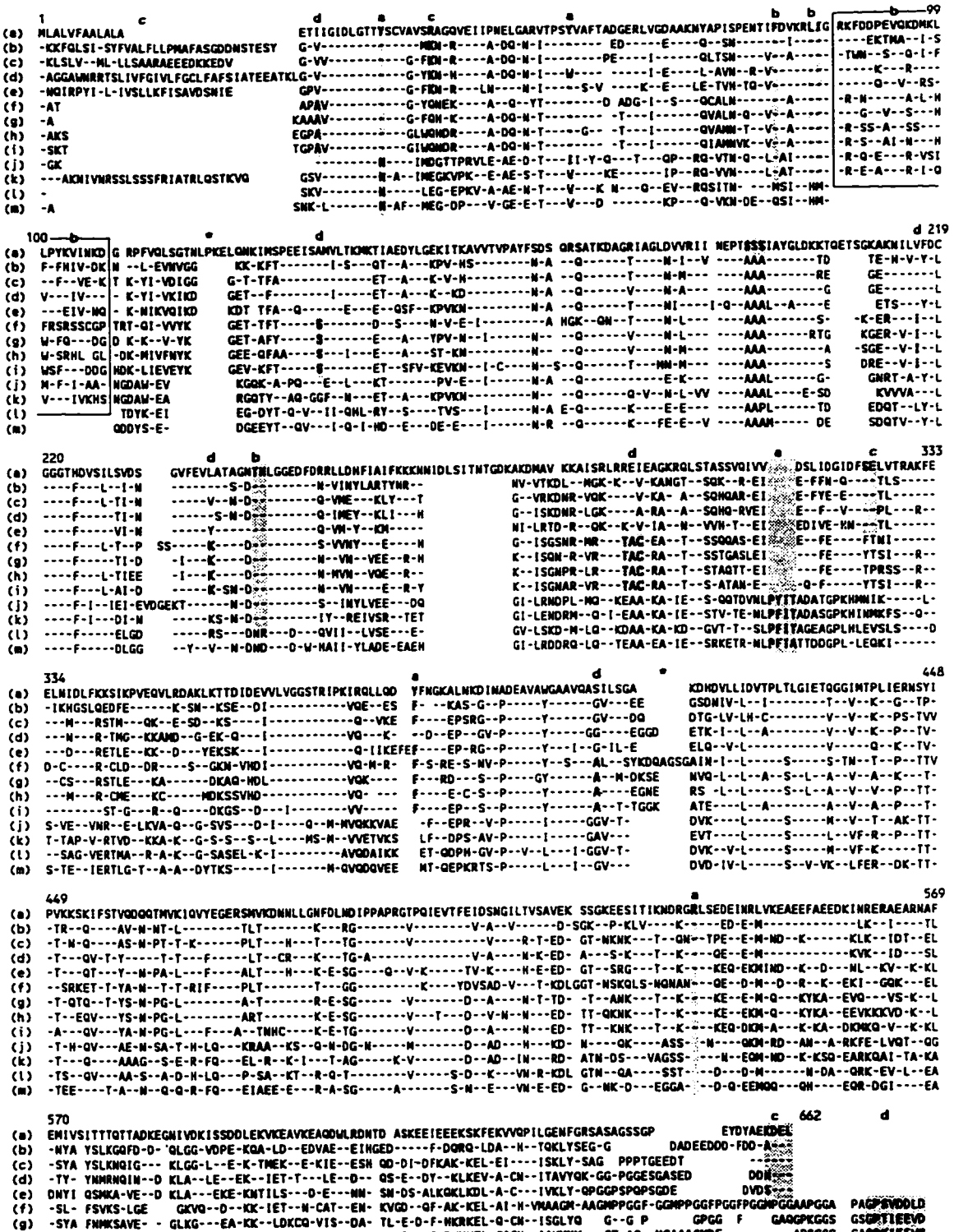


FIG. 1. Alignment of *G. lamblia* HSP70 sequences with sequences from other representative species. Rows: a–e, ER-resident forms of HSP70 (GRP78); f–h, cytoplasmic homologs. Species in rows: a, *G. lamblia*; b, *Saccharomyces cerevisiae*; c, human; d, tobacco; e, *Plasmodium falciparum*; f, *G. lamblia*; g, human; h, *Entamoeba histolytica*; i, maize; j, *Escherichia coli*; k, mitochondrial HSP70, *S. cerevisiae*; l, *Bacillus subtilis*; m, *Halobacterium marismortui*. a, b, c, and d refer to various signature sequences (shaded) that are characteristics of different groups of HSP70 homologs as follows: a, signature sequences providing distinction between eukaryotic and prokaryotic homologs; b, signature sequences shared by eukaryotic homologs and Gram-negative eubacteria (the large insertion in the N-terminal quadrant that is common between these species is boxed); c, signatures distinctive of ER HSP70 homologs; and d, signatures specific for eukaryotic cytoplasmic HSP70 homologs. Asterisks mark the positions where variable insertions (1–10 amino acids) are found in various eukaryotic homologs. The numbers above the top row refer to the *G. lamblia* GRP78 sequence positions; residues identical to the *G. lamblia* GRP78 sequence (top row) are denoted by dashes.

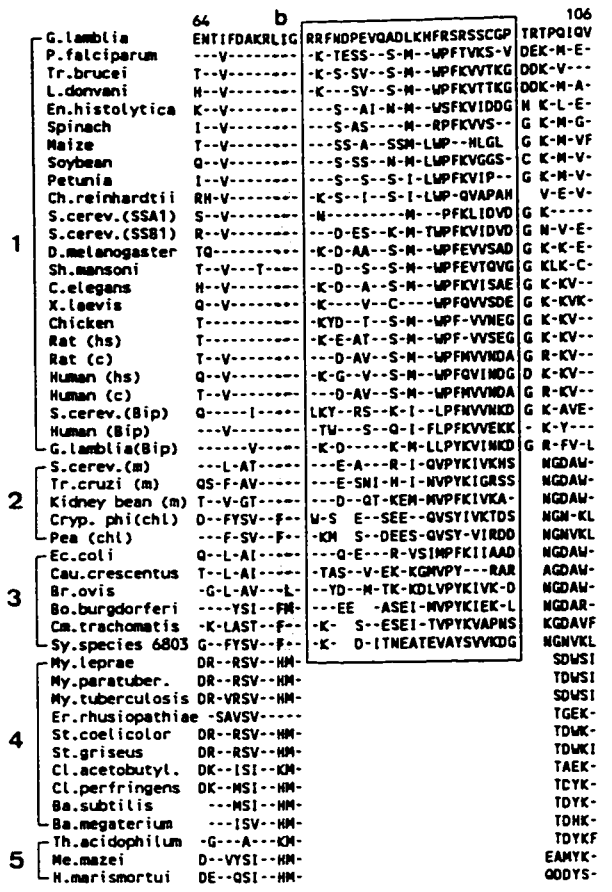


FIG. 2. Comparison of HSP70 sequences from different species in the region corresponding to the insert in the N-terminal quadrant. The numbers 1, 2, 3, 4, and 5 refer, respectively, to eukaryotic, organellar, Gram-negative eubacteria, Gram-positive eubacteria, and archaebacterial species or homologs. The large insert in groups 1–3 is boxed, other sequence features shared by these species (b) are shaded. The abbreviations for genes or species not previously cited in Fig. 1 are: Tr., *Trypanosome*; L., *Leischmania*; En., *Entamoeba*; Ch., *Chlamydomonas*; S. cerev., *S. cerevisiae*; D., *Drosophila*; Shi., *Shistosoma*; C., *Caenorhabditis*; X., *Xenopus*; Ec., *Escherichia*; Cau., *Caulobacter*; Br., *Brucella*; Bo., *Borellia*; Cm., *Chlamydia*; Sy., *Synechocystis*; Cryp., *Cryptomonas*; My., *Mycobacterium*; Er., *Erysiothrix*; St., *Streptomyces*; Cl., *Clostridium acetobutylicum*; Ba., *Bacillus*; Th., *Thermoplasma*; Me., *Methanosarcina*; and H., *Halobacterium*. The residues identical to the top row are denoted by dashes (-). The notations (c), (hs), (m), (Bip), and (chl) in parentheses refer to cognate, heat-induced, mitochondrial, ER, and chloroplast homologs. SSA1 and SSB1 are names of specific homologs. The numbers at the top indicate the position in the *G. lamblia* Cyto HSP70 sequence.

ER and cytoplasmic homologs branched off at a very early time in the evolution of eukaryotic cells and that these have since then evolved independently.

DISCUSSION

This paper describes the cloning and sequencing of two different HSP70 homologs from the protozoan *G. lamblia*. One of these genes corresponds to the cytosolic form of the protein, while the other, based on the presence of characteristic N-terminal signal sequence, C-terminal ER retention sequence (KDEL), and other signatures, is the equivalent of the GRP78 or Bip protein, which is an ER-resident protein in eukaryotic species (3, 8). This report identifies an ER-resident protein in *G. lamblia*, which is regarded as a primitive eukaryotic "archaezoan" lacking mitochondria and other organelles such as peroxisomes (6, 14). Although some structures analogous to rough ER have been observed in electron micrographs (15, 16), the definitive identification of ER in *G. lamblia* has not yet been made.

Phylogenetic trees based on both cytoplasmic HSP70 as well as ER HSP70 (GRP78) homologs, independently provide strong support to the inference based on 16S rRNA (7) regarding the deepest divergence of *G. lamblia* among various eukaryotic species including plant, animal, fungi, and protists. The possibility that the observed deep branching of *G. lamblia* is a consequence of unequal rate effect (17) is considered unlikely but cannot be entirely excluded. The phylogenetic trees based on HSP70 sequences also show a deep branching of the yeast *S. cerevisiae*, which raise questions concerning the phylogenetic position of fungi among eukaryotes and adds to the debate in this regard (see refs. 18 and 19). In composite phylogenetic trees based on HSP70 sequences, the distinct branching of the cytosolic and ER homologs strongly indicates that they diverged from each other at a very early stage in the evolution of eukaryotic cells. The early divergence of cytoplasmic and ER HSP70s has also been proposed earlier (20). Since both these families of homologs contain a number of unique signature sequences (autapomorphic characteristics) that are not found in any of the prokaryotic or organellar homologs, one could exclude the possibility of lateral gene transfer occurring at a later time from either of these sources in their origin. We have previously postulated that these autapomorphic sequence signatures were introduced in the eukaryotic common ancestor during an early stage in its evolution before the genome became stabilized (see ref. 4). The presence in all eukaryotic homologs of these autapomorphic signature sequences, in fact, provides strong evidence regarding the monophyletic nature of the eukaryotic lineage. In view of the above, to explain the presence of these two types of HSP70 homologs in all eukaryotic organisms one has to invoke that they arose by a gene duplication event that took place at a very early stage in the eukaryotic cell history.

The HSP70 family of proteins, whose gene duplication is postulated here, carry out a highly conserved "molecular chaperone" function in the transport of other "passenger" proteins across intracellular membranes and in the maintenance of proper protein conformations (see refs. 1–3). This function of HSP70 is apparently essential in all cellular compartments as suggested by the presence of distinct homologs in different compartments such as cytoplasm, ER, mitochondria, and chloroplasts. The current view is that some of the other compartments that are present in cells are either topologically equivalent to cytosol (namely, nucleus), or to the ER (namely, Golgi, endosomes, and lysosomes) (21, 22). With regard to the evolution of ER, the close resemblance in the composition of ER membrane and nuclear outer envelope and the continuity of nuclear outer membrane with ER membrane (21, 22) strongly suggest that both of these structures evolved at about the same time. The evolution of ER (or alternatively nucleus) in the ancestral eukaryotic cell thus created a new intracellular compartment that has to import and export proteins from outside, and this presumably necessitated the duplication of gene(s) for some of the molecular chaperone proteins that are essential in this regard. If the present hypothesis is correct, then the genes for some of the other chaperone proteins that interact with HSP70 and are essential for its function are also expected to be duplicated at the same time and should be present in both cytosol and ER.

Lastly, from the sequence comparisons presented in Figs. 1 and 2, it is clear that all of the eukaryotic HSP70 homologs, including those from *G. lamblia*, share a number of sequence features in common with the Gram-negative group of bacteria. The most prominent of these features is the presence in these species of a relatively conserved insert of between 23

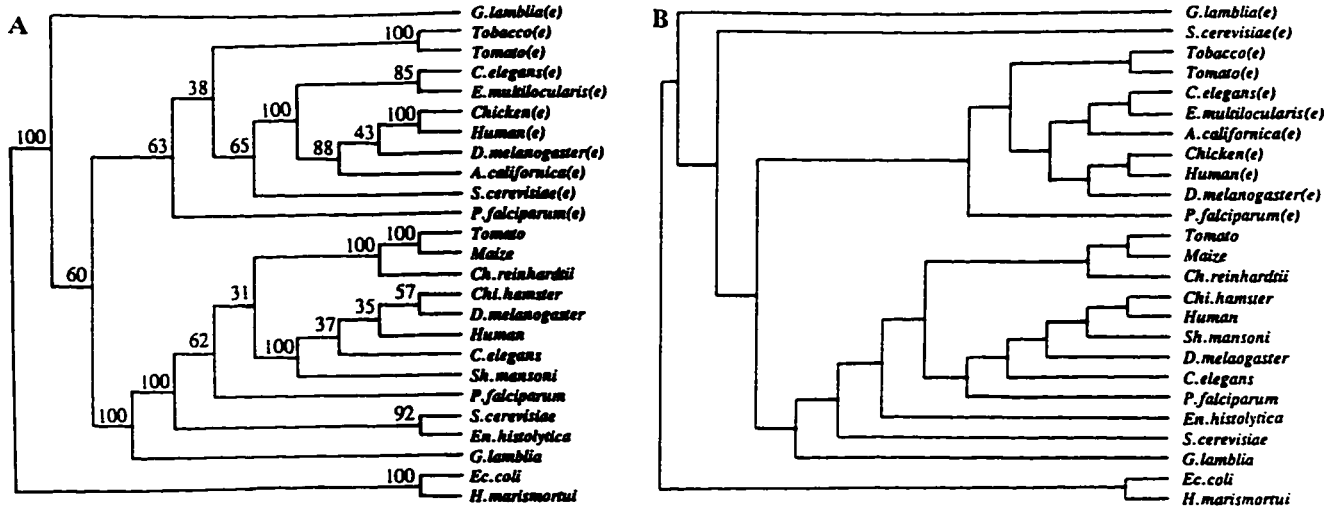


FIG. 3. Phylogenetic trees based on HSP70 sequences. (*A*) Neighbor-joining consensus tree obtained after 100 boot straps. The numbers at the forks indicate the number of times the species, which are to the right of that fork, grouped together out of 100 trees. Phylogenetic analysis was carried out on the sequence region corresponding to that between amino acid 14 (beginning with TIIGDL) and amino acid 570 (ending with RNAFE) in the *G. lamblia* GRP78 sequence (Fig. 1), which could be properly aligned in all species (5). (*B*) Maximum parsimony tree based on the same sequence data. The abbreviations, in addition to those described in the Fig. 2 legend, are: *E.*, *Echinococcus*; *A.*, *Aplysia*; *Chi.*, Chinese hamster. The (e) in parentheses refers to the ER form of HSP70 homologs. For *S. cerevisiae*, the homolog SSA4 was employed in these analyses.

and 27 amino acids in the N-terminal quadrant following the sequence KRLIG, which is not found in any of the homologs from archaebacteria or Gram-positive bacteria (see Fig. 2). The presence of various unique, shared sequence features between these species and detailed phylogenetic analyses of the HSP70 sequence data (the results of which are not affected by excluding this region) provide strong evidence that the eukaryotic HSP70 homologs have evolved from a Gram-negative eubacterial ancestor (4). This view is at variance with the one currently favored based on a limited number of proteins (namely, EF-1, EF-2, RNA polymerase A and C subunits, F- and V-type ATPases), which indicate closer similarity between archaebacteria and eukaryotic homologs (23–27). To explain this dilemma, we have proposed (4) a chimeric model for the origin of the eukaryotic cell in which the eukaryotic cell nucleus is formed by the primary fusion between an archaebacterium and a Gram-negative eubacterium. Based on the observations of Rivera and Lake (27), the archaebacterial partner was very likely a member of the thermoacidophilic (i.e., eocyte) group. After the fusion and during early stages of the formation of the pro-eukaryotic genome, an assortment or selection of genes from the two fusion partners took place. Depending upon the resolution of the duplicate genes in the resulting eukaryotic genome, the genes from the present-day eukaryotic cells will exhibit greater similarity to either one or the other parent (i.e., either the archaebacterium or Gram-negative eubacterium). The chimeric model for the origin of eukaryotic cells is strongly supported by our recent studies in which detailed phylogenetic analyses of all protein sequences in the data bases for which information is available from the main groups of organisms (namely, archaebacteria, Gram-positive and Gram-negative bacteria, and Eukaryote) are carried out (G. B. Golding and R.S.G., unpublished results). Results of these analyses indicate that in addition to HSP70, for several other protein sequences the eukaryotic homologs show a closer phylogenetic relationship to the Gram-negative eubacteria as compared with the archaebacteria. The observed relationship of eukaryotic homologs either to the archaebacteria or to the Gram-negative bacteria seems specific and not just by chance because in no case did homologs from the Gram-positive group of bacteria exhibit closer relationships to the eukaryotic homologs.

We thank Drs. G. B. Golding and B. J. Soltys for helpful discussions and Dr. Golding for access to his computer for phylogenetic analyses. The work was supported by a research grant from the Medical Research Council of Canada to R.S.G.

1. Gething, M. J. & Sambrook, J. (1992) *Nature (London)* **355**, 33–45.
2. Hartl, F.-U., Martin, J. & Neupert, W. (1992) *Annu. Rev. Biophys. Biomol. Struct.* **21**, 293–322.
3. Craig, E. A., Gambill, B. D. & Nelson, R. J. (1993) *Microbiol. Rev.* **57**, 402–414.
4. Gupta, R. S. & Golding, G. B. (1993) *J. Mol. Evol.* **37**, 573–582.
5. Gupta, R. S. & Singh, B. (1992) *J. Bacteriol.* **174**, 4594–4605.
6. Cavalier-Smith, T. (1987) *Nature (London)* **326**, 332–333.
7. Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. (1989) *Science* **243**, 75–77.
8. Pelham, H. R. B. (1989) *Annu. Rev. Cell Biol.* **5**, 1–23.
9. Keister, D. B. (1983) *Trans. R. Soc. Trop. Med. Hyg.* **77**, 487–488.
10. Galley, K. A., Singh, B. & Gupta, R. S. (1992) *Biochim. Biophys. Acta* **1103**, 203–208.
11. Felsenstein, J. (1993) *PHYLIP Manual, Version 3.5* (Herbarium, Univ. of California, Berkeley).
12. Gray, M. W. & Doolittle, R. F. (1982) *Microbiol. Rev.* **46**, 1–42.
13. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
14. Cavalier-Smith, T. (1989) *Nature (London)* **339**, 100–101.
15. Reiner, D. S., McCaffery, M. & Gillin, F. D. (1990) *Eur. J. Cell Biol.* **53**, 142–153.
16. Jacobson, L. M. & Band, R. N. (1987) *J. Protozool.* **34**, 83–86.
17. Felsenstein, J. (1978) *J. Syst. Zool.* **27**, 401–428.
18. Loomis, W. F. & Smith, D. W. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9093–9097.
19. Wainwright, P. O., Hinkle, G., Sogin, M. L. & Stickel, S. K. (1993) *Science* **260**, 340–342.
20. Nicholson, R. C., Williams, D. B. & Moran, L. A. (1990) *Proc. Natl. Acad. Sci. USA* **86**, 1159–1163.
21. Dillon, L. S. (1981) *Ultrastructure, Macromolecules, and Evolution* (Plenum, New York).
22. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1989) *Molecular Biology of the Cell* (Garland, New York), 2nd Ed., pp. 408–409.
23. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9355–9359.
24. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6661–6665.
25. Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H. P., Lottspeich, F., Garrett, R. A. & Zillig, W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4569–4573.
26. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579.
27. Rivera, M. C. & Lake, J. A. (1992) *Science* **257**, 74–76.

CHAPTER 4

Identification of endoplasmic reticulum in the primitive eukaryote *Giardia lamblia* using cryoelectron microscopy and antibody to BiP

The work presented in this chapter describes the existence of endoplasmic reticulum (ER) in *Giardia lamblia*, an early diverging eukaryote. The existence of ER in this organism has been questioned and we addressed this issue by immunoelectron microscopy using antibodies to an ER protein to identify ER. This was done by raising antibodies against the ER HSP70 homolog, Bip, and using this antibody to determine the immunolocalization of the marker protein in *G. lamblia* cells. My contribution to this chapter was the expression of *G. lamblia* Bip in bacteria, purification of the recombinant protein, raising polyclonal antibody to the recombinant Bip and immunofluorescence visualization. Raising antibody to giardial Bip was a key development. Dr. B. J. Soltys has utilized the Bip antibody in the cryoelectron microscopy and immunogold labeling experiments to definitively establish the existence of ER and to distinguish between different types of endomembranes in *G. lamblia*.

Journal of Cell Science 109, 1909-1917 (1996)
Printed in Great Britain © The Company of Biologists Limited 1996
JCS8009

Identification of endoplasmic reticulum in the primitive eukaryote *Giardia lamblia* using cryoelectron microscopy and antibody to BiP

Bohdan J. Soltys, Mizied Falah and Radhey S. Gupta*

Department of Biochemistry, McMaster University, Hamilton, ON Canada L8N 3Z5

*Author for correspondence

SUMMARY

***Giardia lamblia* trophozoites contain a complex endomembrane system as demonstrated by fluorescence and cryoelectron microscopy. The endomembrane system was weakly detected in live cells using the fluorescent membrane dye 3,3′-dihexyloxacarbocyanine iodide. The definitive identification of endoplasmic reticulum required the development of a molecular label. We expressed Giardial Bip in *Escherichia coli* and raised a polyclonal antibody to the purified protein. In western blots, the antibody was specific for Giardial Bip and did not react with human, monkey and rodent homologs. By immunofluorescence microscopy in methanol fixed cells the antibody visualized tubular structures and other subcellular components that required characterization by electron microscopy. Using cryotechniques we directly demonstrate the presence of a complex endomembrane system at the ultrastructural level. In conjunction with Bip immunogold labeling of cryosections we identify: (1) endoplasmic reticulum cisternae and tubules; (2) stacked perinuclear membranes; and (3) Bip presence in the nuclear envelope. Both the endoplasmic reticulum and nuclear envelope were found either with or without a cleft region suggesting each may contain common specialized sub-regions. In stacked perinuclear membranes, which may represent either multilamellar endoplasmic reticulum or a Golgi apparatus, Bip labeling was restricted to peripheral layers, also suggesting specialized sub-regions. Labeled endomembrane systems could be observed associated with microtubule structures, including axonemes and the adhesive disk. The presence of an extensive endomembrane system in *Giardia lamblia*, which represents one of the earliest diverging eukaryotic species, supports the view that both the nucleus and endomembrane system co-evolved in a common ancestor of eukaryotic cells.**

Key words: Bip, Endoplasmic reticulum, Golgi apparatus, Giardia

INTRODUCTION

Giardia lamblia represents one of the deepest branching or most primitive eukaryotes in existence (Sogin et al., 1989; Gupta et al., 1994), and is further distinguished because it lacks many of the subcellular organelles characteristic of higher eukaryotes, including mitochondria and peroxisomes (Friend, 1966; Feely et al., 1990; Adam, 1991). As such, this flagellated protozoan may provide our closest glimpse into the origins of the eukaryotic cell. The endoplasmic reticulum (ER) was first discovered and characterized by Keith Porter in mammalian cells (e.g. Porter, 1953). To this date our appreciation of ER structure and distribution has been derived from studies primarily on mammalian cells and there is a paucity of information on endomembranes in primitive eukaryotes. Although the fine structure of Giardia was first described 30 years ago (Friend, 1966), the very existence of ER has been doubted, even recently (Feely et al., 1990; Meyer, 1994). On the one hand, a lack of ER would appear to violate a central dogma in the evolution of the eukaryotic cell, namely that the endomembrane system and nucleus co-evolved in the same evolutionary event and that all eukaryotic cells would possess both (Alberts et al., 1994). On the other hand, if ER is present it may be present in its simplest evolutionary form and thus the study of Giardia may provide fundamental insights into intracellular organization.

This laboratory recently obtained the first molecular evidence that ER must be present in Giardia after isolating the gene for Bip (Gupta et al., 1994). Bip is a hsp70 homolog which resides in the lumen of the ER in higher eukaryotes. In higher eukaryotes, including yeast, Bip functions as a molecular chaperone in protein folding and in the translocation of proteins across the ER membrane (Craig et al., 1993). Giardial Bip was found to contain a classic C-terminal KDEL ER retention signal (Gupta et al., 1994), implying that ER membranes would be present in this organism. Since the KDEL signal acts in the retrieval of ER proteins that have entered the Golgi (Alberts et al., 1994), then the existence of a Golgi complex is also probable. We have now raised an antibody that is specific for Giardial BiP and have localized the protein at both the light and electron microscopic levels. The definitive identification of ER, however, also required demonstrating the presence of endomembranes in Giardia. The endomembranes of Giardia have eluded most past investigators. Most ultrastructural studies to date have dealt primarily with the organization of the cytoskeleton (see Adam, 1991; Soltys and Gupta, 1994a). Friend (1966) using conventional

EM methods originally reported that the only intracellular membranes that are present in Giardia are the nuclear envelope and the peripheral vacuoles that underlie the dorsal surface. The peripheral vacuoles are currently thought to be either secretory organelles (Friend, 1966; Reiner et al., 1990), endocytic vacuoles (Bockman et al., 1968; Tai et al., 1993) or to be lysosome-like (Feely and Dyer, 1987; Lindmark, 1988; McCaffery and Gillin, 1994). Structures with resemblance to ER, however, have been observed in electron micrographs (Reiner et al., 1990; McCaffery and Gillin, 1994) but were not identified with a definitive molecular marker.

To identify ER membranes in Giardia and to characterize their morphology and subcellular distribution, we have in this paper first of all adopted the cryotechniques of Tokuyasu (1986) to demonstrate the presence of an extensive endomembrane system in Giardia. The advantages of cryosections prepared by these procedures include enhanced membrane preservation and, since cells are not embedded in plastic nor osmicated to fix/contrast membranes, there is high epitope reactivity during antibody labeling. Second, to obtain an ER marker, we report the expression of recombinant Giardial BiP in *E. coli* and the development and characterization of an antibody to it. We have used this antibody in immunogold labeling of cryosections to provide the first definitive identification of ER in Giardia using a molecular marker that is specific for the ER network. We have also identified stacked perinuclear membranes in trophozoites that, based on morphology and partial labeling with BiP antibody, may represent the Golgi apparatus. An overview of the ER network at the light microscopic level was also obtained by visualizing endomembranes in living trophozoites with $DiOC_6(3)$ cyanine dye labeling (Terasaki et al., 1984; Soltys and Gupta, 1992), or by immunofluorescence labeling in fixed cells with Bip antibody. The results indicate Giardia has an unexpectedly extensive endomembrane system, the characteristics of which are described.

MATERIALS AND METHODS

Cell culture

G. lamblia WB (ATCC 30957) trophozoites were grown in glass culture tubes in TY1-S-33 medium supplemented with bile (Keister, 1983). Trophozoites growing as monolayers were harvested by washing and chilling in PBS (containing 8 g NaCl, 0.2 g KCl, 1.15 g Na_2HPO_4 and 0.085 g KH_2PO_4 per liter, at pH 7.2), followed by centrifugation. Human fibroblasts, B-SC-1 African green monkey kidney cells and Chinese hamster ovary cells were grown according to previously described procedures (Soltys and Gupta, 1992). All cells were free of mycoplasma contamination.

Bacterial expression of Giardial Bip

Based on the nucleotide sequence of Giardial Bip, forward and reverse oligonucleotide primers which flanked the coding sequence for the protein and containing unique restriction enzyme sites (*Eco*RI and *Hin*dIII, respectively) were custom synthesized (Mobix - The Institute for Molecular Biology and Biotechnology, McMaster University, Hamilton, ON). The nucleotide sequence of the primers employed were as follows: (forward) 5′-GAATTCCATGCTCGCTCTTGTCTTTGCC-3′ and (reverse) 5′-GGATCCAAGCTTAGAGTTCATCTTTTTCTGC-3′. These primers were employed in polymerase chain reaction (PCR) in conjunction with DNA from the plasmid HSG-12, which contained the cloned Giardial Bip gene (Gupta et al., 1994). After 35 PCR cycles, specific amplification of a 1.9 kb fragment was observed, as expected based on the positions of the primers. The fragment was isolated and after digestion with *Eco*RI and *Hin*dIII, it was subcloned in the pTrcHis expression plasmid, digested with the same enzymes. The sequencing of the plasmid (following transformation of *E. coli* JM109 cells) confirmed that the insert was in the correct reading frame. The DNA from this plasmid (designated pGLBip-1) was employed to transform *E. coli* Top10 cells for expression of the recombinant protein. However, no significant expression of the recombinant protein was observed using this construct.

The construct used in the above experiment contained the entire coding sequence of the Giardial Bip, including the N-terminal ER targeting presequence, which is not present in the mature protein. Since this sequence could adversely affect the expression, or even may be causing secretion of the recombinant protein, the plasmid pGLBip-1 was digested with the enzyme *Aval* and then religated. This led to removal of a portion of the 5′-end sequence corresponding to the N-terminal 47 amino acids. The resulting clone (designated pGLBip-2) contained 615 amino acids of the Giardial Bip sequence and it showed very good expression of the recombinant protein.

Expression and purification of recombinant protein

An overnight culture of *E. coli* cells harbouring pGLBip-2 plasmid was diluted 1:100 in fresh growth medium containing 100 μg/ml ampicillin and grown at 37°C. After 1-2 hours when the OD_{660} of the culture was between 0.4-0.6, isopropyl-β-D-thiogalactopyranoside (IPTG) was added to the culture at a final concentration of 0.5 mM. No IPTG was added to a parallel control culture. At various times afterwards, samples of induced and uninduced cultures were taken out and expression of the recombinant protein was examined by SDS-PAGE. For large scale preparation of the recombinant protein, the culture was induced with IPTG for 3 hours and the washed cell pellet was resuspended in 20 volumes of lysis buffer (6 M guanidine hydrochloride, 20 mM sodium phosphate and 500 mM sodium chloride, pH 7.8). After clearing the lysate by centrifugation (5,000 rpm for 15 minutes), the supernatant was applied to a prewashed Probond Nickel column (Invitrogen Corp., San Diego, CA). The recombinant protein which contains a poly-His_6 tag binds to the Ni column. After washing the column with buffers of decreasing pH (as per the supplier's protocol), the bound protein was eluted using a buffer containing 8 M urea, 20 mM sodium phosphate, 500 mM NaCl, pH 4.0, and desalted. SDS-PAGE analysis of the eluted protein showed that it consists mostly (>90%) of a ≈70 kDa protein and a few other minor protein bands of lower molecular masses, which presumably were proteolytic degradation products. Nevertheless, for antibody production, the ≈70 kDa protein band was excised from preparative SDS-PAGE gels and the eluted protein was lyophilized and used as antigen.

Polyclonal antibody to the recombinant Bip was raised by injecting about 1 mg of the recombinant protein emulsified with 1 ml of Titre Max (Cedarlane Laboratories, Hornby, ON) subcutaneously in a female rabbit. After 4-5 weeks, when the animal showed an immune response against the injected antigen, a second booster dose of the antigen in Titre Max was given. The animal was bled within the next 2-3 weeks. In most of the experiments described the polyclonal antibody was further affinity purified using the recombinant protein.

Gel electrophoresis and western blots

Giardia cellular extracts for polyacrylamide gel electrophoresis and western blots were prepared by boiling a pellet of PBS-washed cells for 10-15 minutes in Laemmli sample buffer containing 2% SDS, 100 mM dithiothreitol, 60 mM Tris-HCl, pH 6.8, and 0.001% bromophenol blue. To prevent proteolysis in sample preparation, ice-cold sample buffer was added to the pellet on ice, the pellet was resuspended by repeated pipetting, then the tube was transferred directly from ice to a boiling water bath. For tissue culture cells, Laemmli sample buffer was added to PBS-washed monolayer cultures and boiling was for 3 minutes. Electrophoresis was in 10% sodium dodecyl sulfate-polyacrylamide gels (SDS-PAGE), as described previously (Gupta and

Dudani, 1987). Proteins were transferred electrophoretically from polyacrylamide gels to nitrocellulose sheets. The blots were blocked with 3% bovine serum albumin (BSA) in saline (0.9% NaCl, 10 mM Tris-HCl, pH 7.4), then reacted with affinity purified rabbit polyclonal antibody to Bip. Visualization of polyclonal antibody was with horseradish peroxidase conjugated secondary antibody directed against rabbit IgG (Bio-Rad Lab. Ltd, Mississauga, ON) and color development with 4-chloro-1-naphthol (Bio-Rad Lab. Ltd, Mississauga, ON).

Immunocytochemistry and microscopy

For light microscopy a concentrated suspension of trophozoites in ice-cold TYI-S-33 culture medium was applied to ethanol-prewashed glass coverslips and trophozoites were allowed to adhere for 5-10 minutes at 37°C in a humidified incubator. For cyanine dye staining of living cells, $DiOC_6(3)$ (3,3'dihexyloxacarbocyanine iodide) (Sigma, St Louis, MO) was used at 5 μg/ml in Giardia culture medium for 40 seconds, as described previously for mammalian cells (Soltys and Gupta, 1992). For immunofluorescence labeling, trophozoites on coverslips were fixed by quickly plunging into −20°C methanol. After fixation for 15 minutes at −20°C, cells were rinsed with PBS, then antibody labeled. Immunofluorescence procedures and microscopy were otherwise described previously (Soltys and Gupta, 1992).

For electron microscopy, PBS washed cells on ice were fixed as a loose pellet of cells with 0.5% glutaraldehyde in 0.1 M sucrose, 0.1 M cacodylate, pH 7.3, at room temperature for 15 minutes (Soltys and Gupta, 1994a,b). To quench unreacted glutaraldehyde, cells were washed and incubated for 15 minutes in 100 mM ammonium chloride in sucrose-cacodylate buffer. Pellets of cells were infiltrated with 2.3 M sucrose for 3 hours (Soltys and Gupta, 1996). The general cryomicrotomy procedures of Tokuyasu (1986) were used. Ultrathin cryosections were cut on a Reichert-Jung ultra cut E ultramicrotome with the FC 4E cryosectioning attachment (knife −85°C; specimen −90°C; chamber −110°C). Antibody labeling of cryosections when using 10 nm gold labels was carried out using a three-stage immunolabeling procedure to amplify labeling intensity (Tokuyasu, 1983; Soltys and Gupta, 1994b, 1996). Sections were preabsorbed at room temperature with 50% fetal calf serum in 0.1 M Tris-HCl, pH 7.5 (carrier buffer). Sections were then reacted with affinity-purified polyclonal antibody in carrier buffer for 1.5 hours at 37°C in a humidified incubator. Washing of sections was for 30 minutes with 5% BSA in 0.1 M Tris-HCl, pH 7.5. Sections were then reacted with a 1:40 dilution of goat anti-rabbit IgG (Jackson Immunoresearch, West Grove, PA) in carrier buffer for 1 hour at 37°C, washed again, then reacted with a 1:5 dilution (A_{520}=0.5) of rabbit anti-goat IgG 10 nanometer gold conjugate (Sigma Chemical Co., St Louis, MO) in carrier buffer for 4 hours at 37°C. When 18 nm gold particles were used, a two-stage procedure was used and the second antibody was a goat anti-rabbit IgG 18 nm gold conjugate (BioCan Sci., Mississauga, ON). After washing, including a high salt wash with 0.5 M KCl in carrier buffer followed by washes with H_2O, cryosections were stained with 2% neutral uranyl acetate, then embedded in methylcellulose containing 0.1% acidic uranyl acetate. Sections were examined at 80 kV with a JEOL 1200 EX transmission electron microscope.

Fig. 1. Immunoblot detection of Bip in *G. lamblia* trophozoites. (a) SDS-PAGE. Coomassie blue staining of the gel. (b) Western blot analysis of trophozoites (lane 1) compared with human fibroblasts, B-SC-1 monkey kidney cells and Chinese hamster ovary cells (lanes 2-4), respectively. Lane 0 in b is a blank lane. Immunoblot in b was stained with affinity purified polyclonal antibody to Giardial Bip and a secondary horseradish peroxidase-conjugated antibody.

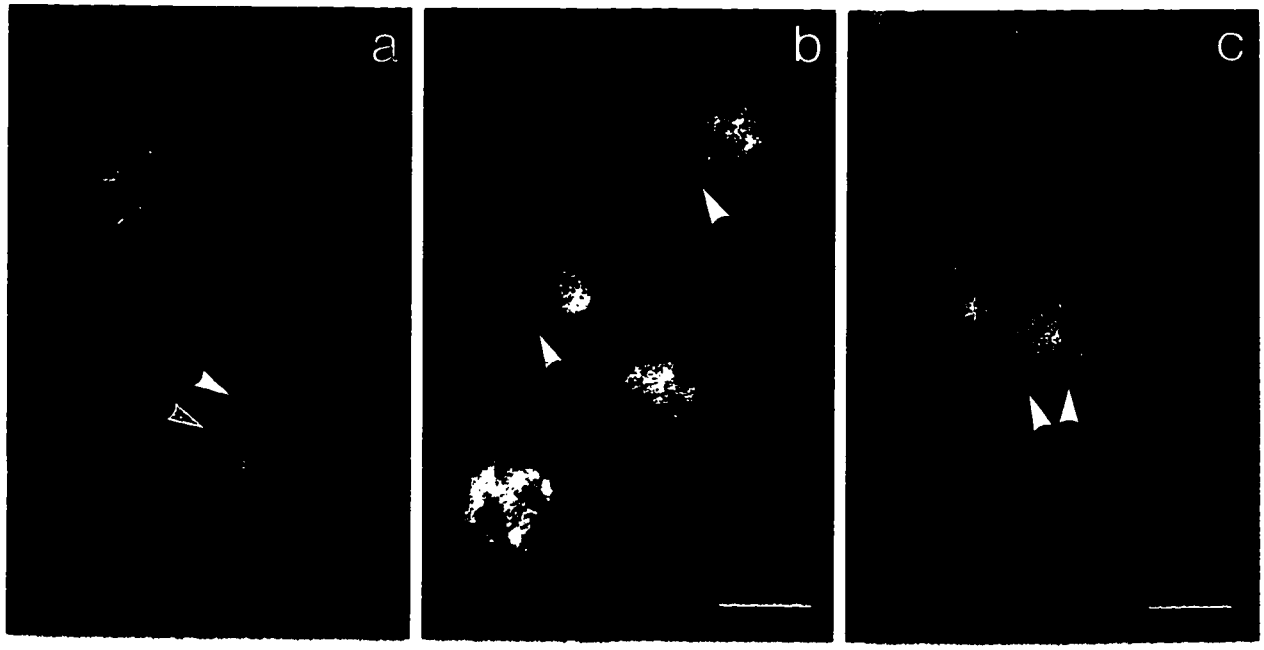


Fig. 2. Fluorescence visualization of endomembranes in *G. lamblia* trophozoites. (a) Cyanine dye $DiOC_6(3)$ staining in living cells. Fluorescein channel. (b and c) Bip antibody labeling in fixed cells. Fluorescein channel. Staining of the nuclear envelope is most prominent in a (see arrowheads). Cytoplasmic tubular structures are prominent in b and c (see arrowheads). The indicated tubular structure in c appears to be greater than 3 μm in length. Bars: (b) 10 μm; (c) 5 μm.

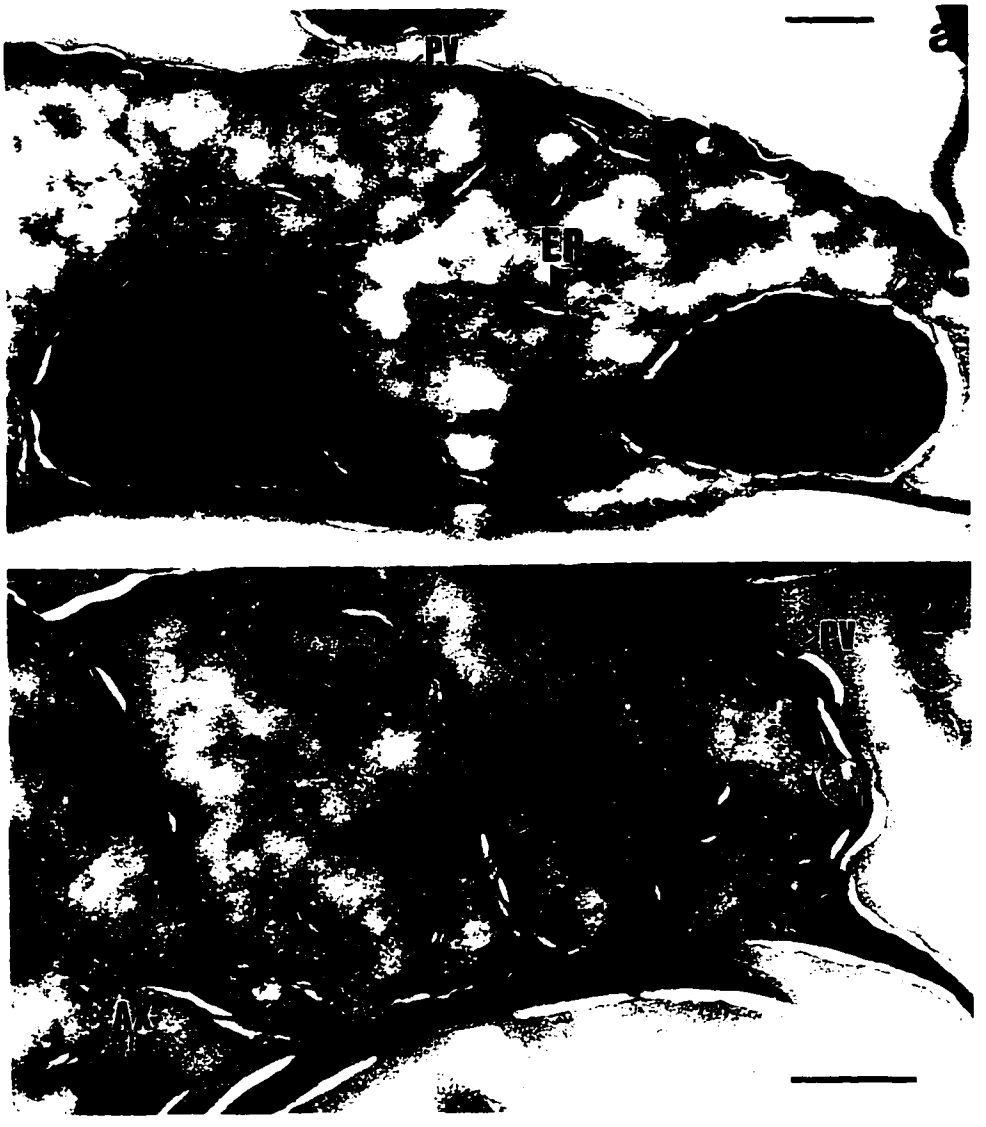


Fig. 3. Cryoelectron microscopy of *G. lamblia* trophozoites and immunogold labeling of endomembranes with antibody to Giardial Bip. Low magnification overviews; 10 nm gold conjugates used, which would be observed with a low power magnifier. ER, endoplasmic reticulum as identified by Bip labeling; AD, adhesive disk, AX axoneme, PV, peripheral vacuole; N, nucleus. Bars, 0.5 μm.

RESULTS

Previous morphological studies have shown that Giardia has a very primitive cytoplasm that lacks organelles such as mitochondria and peroxisomes. We asked whether ER is present and how does it compare morphologically with ER in higher eukaryotes. The definitive identification of any endomembrane components as being ER requires a molecular label, so we produced an antibody to the ER resident molecular chaperone Bip. The Bip gene from *G. lamblia* has been cloned and sequenced in our laboratory (Gupta et al., 1994). In the present study we expressed histidine tagged recombinant Bip protein in *E. coli*, purified it on a nickel column, then raised a rabbit polyclonal antibody to the purified protein (see Materials and Methods for details).

The specificity of the Bip antibody was evaluated in *G. lamblia* whole cell extracts and compared with other species. Fig. 1a and b are polyacrylamide gels and immunoblots, respectively, of *G. lamblia* (lane 1) compared with human fibroblasts, B-SC-1 monkey kidney cells and Chinese hamster ovary cells (lanes 2-4, respectively). A 70 kDa protein was strongly labeled in Giardia. No protein bands were labeled in the other cell types. In Giardia, faint lower molecular mass bands were also detected, probably representing minor proteolysis. Thus, the Bip antibody is species specific and does not react with rodent or primate cells.

We first evaluated endomembranes in live cells using fluorescence dye labeling. The cyanine dye $DiOC_6(3)$ has previously been used to visualize ER and mitochondria in live or fixed mammalian cells (Terasaki et al., 1984; Soltys and Gupta, 1992). In the absence of mitochondria in Giardia this dye might be expected to visualize primarily ER. Fig. 2a shows $DiOC_6(3)$ staining of live trophozoites. There is strong fluorescence staining of the nuclear envelope/periphery (arrowheads) and there are other fluorescent structures distributed throughout the cytoplasm. These structures appear to have both linear and punctate shapes. $DiOC_6(3)$ therefore is likely labeling endomembranes and these are presumably ER. The signal intensity is low, however, and the shape and distribution of these subcellular components is less amenable to characterization than ER in mammalian cells (Terasaki et al., 1984; Soltys and Gupta, 1992). It should be noted, however, that these were living cells and there were significant subcellular movements and photobleaching of fluorescence during photography, causing image loss. Fig. 2b and c in comparison show immuno-

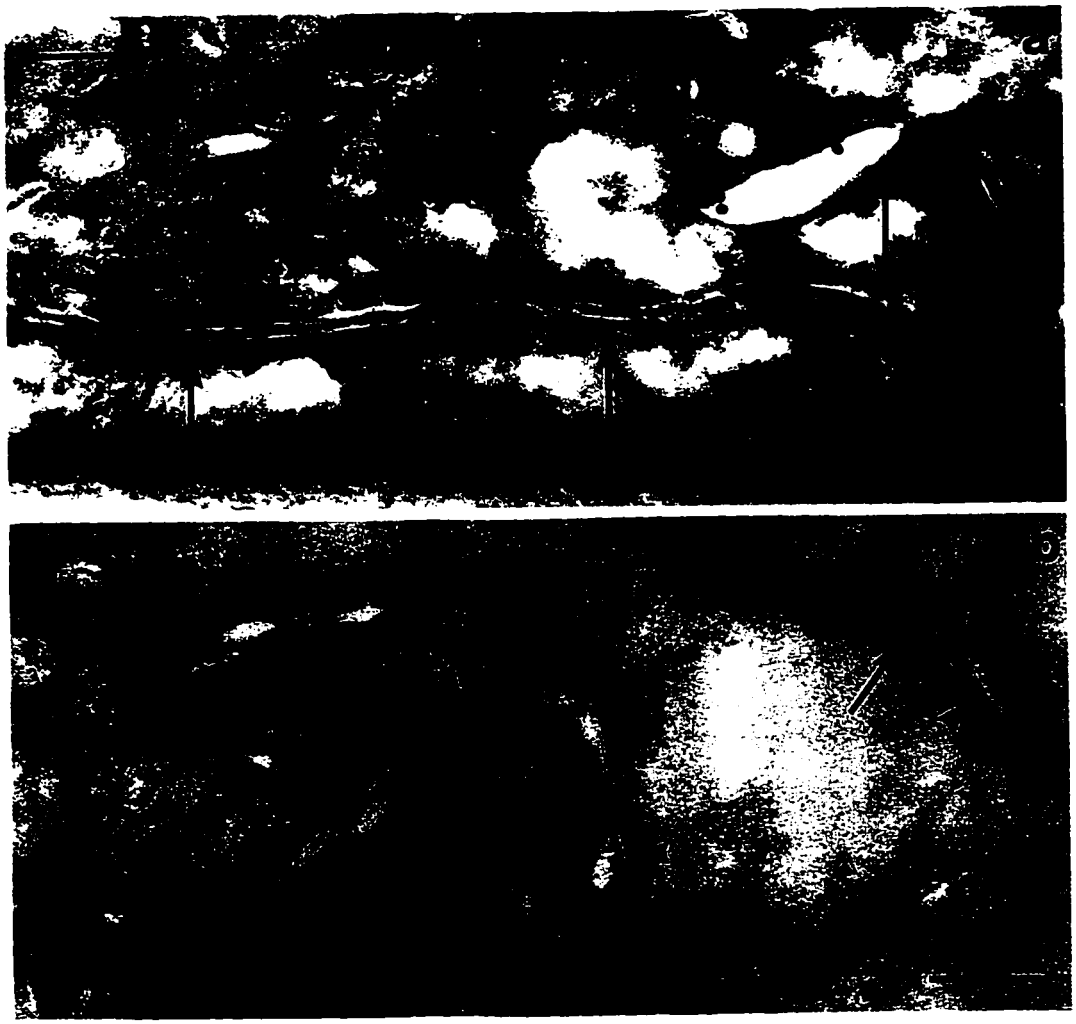


Fig. 4. Immunogold labeling of *G. lamblia* cryosections with Bip antibody. High magnification micrographs of ER membranes: (a) 18 nm gold labels; (b) 10 nm gold labels. In a, arrows denote a single ER tubule which is approximately 2 μm in length. A cytoplasmic cleft is also present in which a limiting membrane is detected along its upper border and is also gold labeled (arrowheads). In b, labeling of ER membranes is increased when 10 nm gold particle procedures are used. ER membranes with cytoplasmic clefts (arrowheads) and without (arrows) are present and both are labeled. These ER regions appear continuous with each other. Bars, 0.2 μm.

fluorescence labeling of fixed cells using Bip antibody. A fluorescent array throughout the cytoplasm is also detected by this approach, showing both tubular (arrowheads) and punctate or vesicle-like characteristics (Fig. 2b, cell in lower left corner), and there appears to be less labeling of the nuclear envelope compared with Fig. 2a. Control cells reacted with preimmune serum showed no labeling (not presented). The cell shown in Fig. 2c is at higher magnification compared with Fig. 2b. The indicated tubular structure, assuming it is a singular entity, is more than 3 μm in length. Thus both dye staining and fluorescent antibody labeling detect similar subcellular components that are present throughout the cytoplasm. While these results suggest the presence of an extensive system of endomembranes, the identification and characterization of ER membranes must be made at the ultrastructural level.

While previous morphological studies had difficulty preserving Giardia endomembranes, we were able to detect ER-like membranes by applying cryotechniques. Fig. 3a and b show low magnification overviews of two different regions of the cell. Fig. 3a is a central section through the two nuclei while Fig. 3b is a peripheral region of the cell. A prominent endomembrane component in both sections is the previously described peripheral vacuoles that underlie the dorsal surface of the cell. The peripheral vacuoles are thought to be secretory organelles (Friend, 1966; Reiner et al., 1990), endocytic vacuoles (Bockman et al., 1968; Tai et al., 1993) or to be lysosome-like (Feely and Dyer, 1987; Lindmark, 1988; McCaffery and Gillin, 1994). In addition, cytoplasmic clefts (labeled ER) are present that are apparently devoid of contents. These clefts were first described by Friend (1966) who did not observe them to contain membranes. These cytoplasmic clefts are abundant and present throughout the cytoplasm. Contrary to previous reports, we found in cryosections that these cytoplasmic clefts are membrane limited (see below). In addition we have also observed equally abundant ER-like membranes that do not contain an empty cleft. In addition to these membranous structures, we have also observed numerous vesicles throughout the cytoplasm that may be transport vesicles or cross-sections of ER tubules, and also novel stacked multilamellar membranes (see below).

We labeled cryosections with antibody to Giardial Bip to identify ER. Fig. 4a and b show high magnification micrographs of membranes labeled with 20 nm and 10 nm gold labels, respectively. In Fig. 4a, a single ER-like tubule can be tracked along its length for a distance of approximately 2 μm and it is labeled along the entire observed length. A cytoplasmic cleft is also present (top right quadrant) in which a membrane is clearly identified on its top edge and it is also antibody labeled. Therefore both these membranous structures are ER membranes. Gold labeling intensity of these membranes is

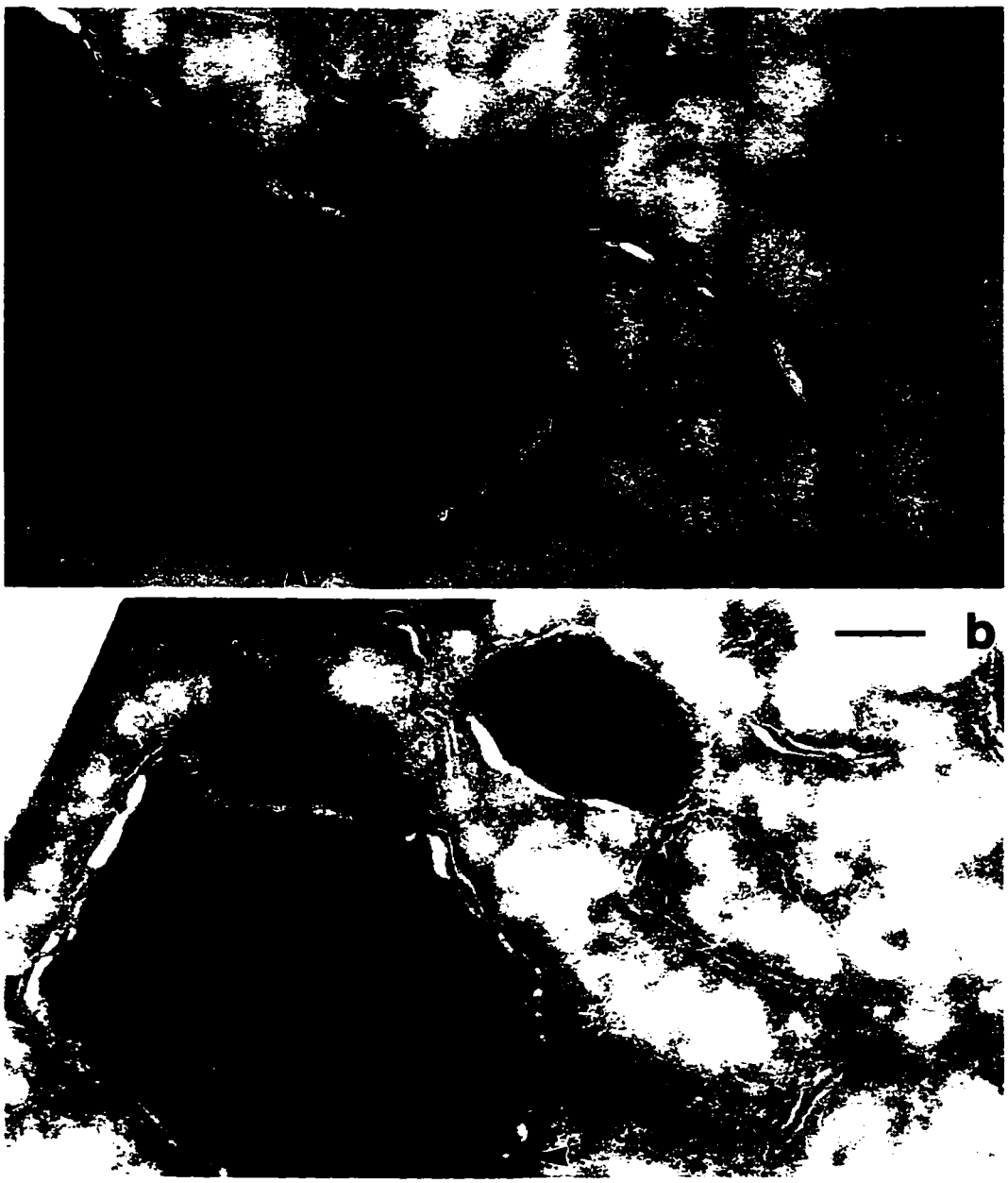


Fig. 5. Bip immunogold labeling of the nuclear envelope and juxtanuclear ER tubules. Some of the sites of the nuclear envelope that are labeled are indicated with arrowheads. Nuclear pores are observed and the nuclear envelope has several cleft regions. The nuclear envelope and the adjacent ER are morphologically similar. Note that the ER tubule in a immediately to the right of the nucleus has both cleft and non-cleft regions. 10 nm gold labels. Bar, 0.25 μm.

increased using our 10 nm gold labeling procedures as shown in Fig. 4b. Labeling again is demonstrated to be on membranes bordering cytoplasmic clefts and in the more typical ER-like membranes lacking a cleft. This is therefore proof that Giardia does indeed have ER. Control sections reacted with preimmune serum showed no labeling (not presented). We have not detected any difference in antibody labeling between the cytoplasmic clefts and the more typical ER membranes. The cleft itself may arise from a component that is readily extracted, so a compositional difference between these sets of membranes is likely. However, cytoplasmic clefts were often found to be continuous with non-cleft regions (Fig. 4b). The low magnification cryosections presented in the preceding Fig. 3 were in fact also antibody labeled (10 nm gold) and an overview of the distribution of gold particles on membranes throughout the cell can be obtained by use of a low power magnifier.

We have also obtained Bip labeling of the nuclear envelope. Fig. 5a and b show two nuclear regions at high magnification. Ten nanometer gold particles were used. ER tubules in the cytoplasm in close proximity to the nuclei are labeled. The nuclear envelopes are also labeled at several locations as indicated by arrowheads. This nuclear labeling is consistent with the view that ER membranes form the nuclear envelope and that Bip has functions at this location also.

We have observed novel multi-layered stacks of membranes in trophozoites, usually in a juxtanuclear location, and these have also been found to contain Bip. Fig. 6a and b show two examples of such stacks labeled with 10 nm gold particles. The stacks extend over a large area, often up to 1-2 μm in width. The intermembrane spacing in these stacks is approximately twofold less than in the general cytoplasmic ER. Note that Bip labeling is restricted to particular layers, suggesting functional specializations in different layers. The structural, positional and compositional properties of these stacked membranes suggest

Fig. 6. Bip immunogold labeling of stacked membrane systems in vegetatively growing trophozoites. Bip labeling is restricted to specific layers, suggesting specialized subregions. The membrane-to-membrane spacing of these stacked cisternae is compressed relative to ER tubules or the nuclear envelope. 10 nm gold labels. Bars, 0.2 μm.

they are different from ER and could possibly represent a primitive Golgi apparatus (see Discussion).

We have found that endomembranes in Giardia are also associated with microtubule structures including the adhesive disk and axonemes. Fig. 7a shows stacked membranes labeled with Bip antibody in association with microtubules of the adhesive disk. Microtubules in the adhesive disk are viewed in cross section (arrowheads), which are at the bottom of so called 'microribbons' composed of the protein giardin (see Adam, 1991). In Fig. 7b, ER tubules (arrows) are found in close association with an intracytoplasmic axoneme. Associations of endomembranes with microtubule-based structures is a hallmark of higher eukaryotes.

DISCUSSION

To develop a molecular marker that can identify ER in *G. lamblia*, we expressed Giardial BiP in *E. coli* and raised an antibody to the purified protein. The specificity of this antibody was established by western blots and by the specific labeling of endomembranes by immunocytochemical procedures. This antibody was found to be species specific and does not react with rodent or primate cells.

By immunofluorescence microscopy using anti-Bip antibody, we obtained an overview of the distribution of ER membranes in whole cells. The ER network was found to be extensive and present throughout the cell. This network shares with higher eukaryotes tubular membrane profiles and there were also nondescript patches of membranes that were labeled. These components were also visualized in living trophozoites using the cyanine dye $DiOC_6(3)$. With either approach, however, the ER network had low signal intensity and was not as clearly defined as in previous fluorescence microscopy of mammalian cells (Terasaki et al., 1984; Soltys and Gupta, 1992). This finding of an extensive ER network in Giardia may not have been expected because of the primitive evolutionary position of this species (Sogin et al., 1989; Gupta et al., 1994).

By applying cryotechniques we were able to preserve giardial endomembranes at the ultrastructural level. This permitted us to identify the ER by immunogold labeling. Cisternae and tubular forms of ER were identified and were found throughout the cell body, often associated with other structural elements including the nucleus, peripheral vacuoles and complex microtubule-based structures such as intracytoplasmic axonemes and the adhesive disk. Labeled vesicles were also observed that may represent transport vesicles or cross sections of ER-tubules. The tubular ER could be distinguished into two subregions, namely those with or without a cleft region. Friend (1966) also observed cytoplasmic clefts and, although limiting membranes were not preserved in his specimens, he considered these clefts to be unique to Giardia among all protozoa and formed by extraction of an unknown component(s) during specimen preparation. Thus the ER in Giardia could have two subregions with different compositions, although both regions contain Bip. The nuclear envelope is considered to be a subregion of ER and we also observed BiP labeling at this site. Cleft regions were also found in the nuclear envelope, suggesting further compositional similarities. These findings of ER subregions and the pervasiveness of ER membranes throughout the cell body are con-

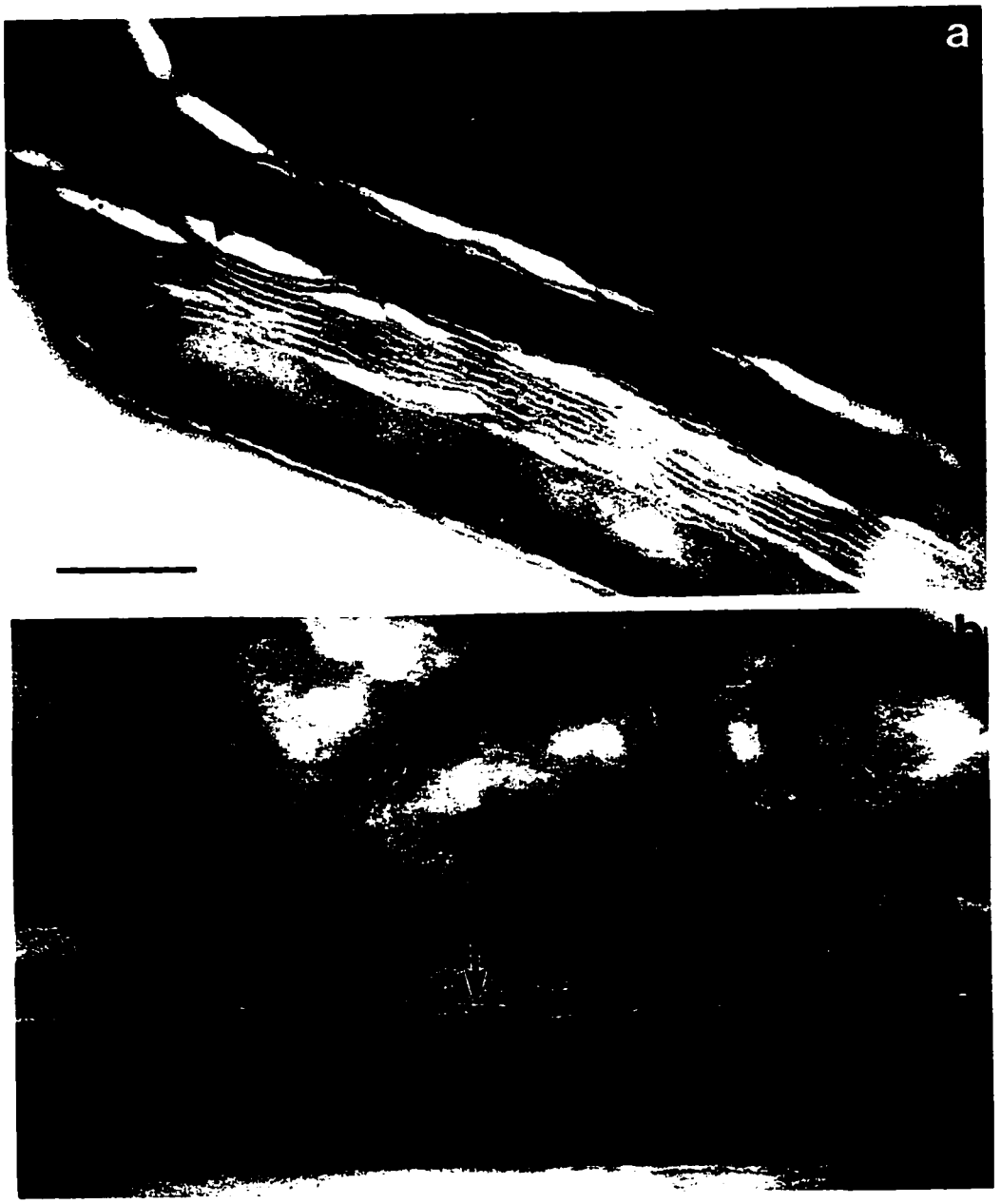


Fig. 7. Bip immunogold labeling of endomembranes associated with microtubule-based structures. (a) Stacked membranes associated with the adhesive disk (AD) in which microtubules (arrowheads) are viewed in cross section. N, nucleus. 10 nm gold labels. (b) ER tubules indicated by arrows are associated with an intracytoplasmic axoneme (AX). 18 nm gold labels. Bars, 0.2 μm.

sistent with current appreciation of the complex organization of ER in higher eukaryotes (Sitia and Meldolesi, 1992).

Our study has been of vegetatively growing trophozoites and in these we have also identified new juxtanuclear membrane stacks which were only partially labeled with BiP antibody. These membranes may represent either annulate lamellae (Kessel, 1992), karmellae (multilamellar ER) (Wright et al., 1988; Wanker et al., 1995) or a post-ER compartment such as the Golgi apparatus. Annulate lamellae are least favoured because the membrane stacks do not have the pore complexes that characterize such structures. If these membranes were a form of karmellae, which appear in cells overproducing ER membrane proteins (Wright et al., 1988; Wanker et al., 1995), then one would expect more uniform Bip labeling, as seen in other ER membranes in Giardia and in previous studies (e.g. Preuss et al., 1991). Alternatively, these flattened cisternae may represent a primitive Golgi apparatus - BiP should localize to both ER and Golgi membranes because Giardial BiP has been shown to contain a KDEL retrieval signal (Gupta et al., 1994) which in higher eukaryotes is recognized by a *cis*-Golgi receptor that functions to send BiP back to the ER (e.g. Lewis and Pelham, 1992). We also found BiP labeling was restricted to particular layers in the juxtanuclear stacks, in contrast with the more consistent labeling of all unstacked ER membranes. Based on the criteria of morphology, subcellular distribution and Bip localization only within a subregion, this membrane system is suggested to be the Golgi apparatus. This putative Golgi apparatus extends over a large area, which is somewhat different from the small compact juxtanuclear mass usually seen in mammalian cells. However, large extensive Golgi systems have also been observed in certain mammalian cells (see Rambourg and Clermont, 1990). Further characterization of these stacked membranes in Giardia trophozoites will require additional molecular markers which are specific for the Golgi (Banting et al., 1995).

ER and Golgi-like membranes may now join peripheral vacuoles and the nuclear envelope (Friend, 1966) as identified endomembranes in Giardia trophozoites. The encystation stage of Giardia's life cycle can be artificially induced in culture (Reiner et al., 1990) and studies of this stage have provided evidence for developmental changes in endomembranes. A Golgi apparatus structure was previously reported to form de novo uniquely during encystation, correlating with biochemical evidence for the appearance of regulated secretion (Reiner et al., 1990; Lujan et al., 1995). The de novo formation of a Golgi apparatus from ER, although possible, is considered a slow process and 'a strong selective disadvantage in a competitive environment' (Warren and Wickner, 1996). Encystation to form a dormant cyst is, in fact, a response to a hostile environment. We speculate that the Golgi apparatus is present throughout Giardia's life cycle but in functionally different forms. McCaffery and Gillin (1994) have also indicated Golgi-like membranes are present in both trophozoites and encysting cells. Encystation-specific vesicles (ESVs) have also been identified in encysting cells - ESVs measure more than 1 μm in diameter and are involved in cyst wall formation (Reiner et al., 1990; McCaffery and Gillin, 1994; McCaffery et al., 1994). How ESVs form is not known. McCaffery and Gillin (1994) have also seen structures resembling transitional elements, putative tubular-vesicular elements and putative transport vesicles. Further molecular marker characterization of Giardia endomembranes, which clearly have a complex organization, should prove interesting from both cell biological and evolutionary perspectives.

In conclusion, the evolution of ER is thought to have taken place in concert with the nucleus (Alberts et al., 1994; Gupta and Golding, 1996) and our findings of ER in the most primitive of known eukaryotes supports this central tenet in cell biology. The structural complexity of the endomembrane system, however, was not anticipated. This is at variance with the suggestion that the earliest eukaryotes possessed an elementary or diminutive endomembrane prototype that later evolved into endomembrane systems with greatly increased complexity.

This work was supported by a Medical Research Council of Canada grant to RSG.

REFERENCES

Adam, R. D. (1991). The biology of Giardia spp. *Microbiol. Rev.* **55**, 706-732.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994). *Molecular Biology of the Cell*, 3rd edn. Garland Publishing Co., New York.

Banting, G., Banting, J. and Lingelbach, K. (1995). A minimalist view of the secretory pathway in Plasmodium falciparum. *Trends Cell Biol.* **5**, 340-343.

Bockman, D. E. and Winborn, W. B. (1968). Electron microscopic localization of exogenous ferritin within vacuoles of Giardia muris. *J. Protozool.* **15**, 26-30.

Craig, E. A., Gambill, B. D. and Nelson, R. J. (1993). Heat shock proteins: molecular chaperones of protein biogenesis. *Microbiol. Rev.* **57**, 402-414.

Feely, D. E. and Dyer, J. K. (1987). Localization of acid phosphatase activity in Giardia muris trophozoites. *J. Protozool.* **34**, 80-83.

Feely, D. E., Holberton, D. V. and Erlandsen, S. L. (1990). In *The Biology of Giardia* (ed. E. A. Meyer), pp. 11-49. Elsevier Science Publishers, New York.

Friend, D. S. (1966). The fine structure of *Giardia muris*. *J. Cell Biol.* **29**, 317-331.

Gupta, R. S. and Dudani, A. K. (1987). Mitochondrial binding of a protein affected in a mutant resistant to the microtubule inhibitor podophyllotoxin. *Eur. J. Cell Biol.* **44**, 278-285.

Gupta, R. S., Aitken, K., Falah, M. and Singh, B. (1994). Cloning of *Giardia lamblia* HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proc. Nat. Acad. Sci. USA* **91**, 2895-2899.

Gupta, R. S. and Golding, G. B. (1996). The origin of the eukaryotic cell. *Trends Biochem. Sci.* **21**, 166-171.

Keister, D. B. (1983). Axenic culture of Giardia lamblia in TY1-S-33 medium supplemented with bile. *Trans. R. Soc. Trop. Med. Hyg.* **77**, 487-488.

Kessel, R. G. (1992). Annulate lamellae: a last frontier in cellular organelles. *Int. Rev. Cytol.* **133**, 43-120.

Lewis, M. J. and Pelham, H. R. B. (1992). Ligand-induced redistribution of a human receptor from the Golgi complex to the endoplasmic reticulum. *Cell* **68**, 353-364.

Lindmark, D. G. (1988). *Giardia lamblia*: localization of acid hydrolase activities in lysosome-like organelles of trophozoites. *Exp. Parasitol.* **65**, 141-147.

Lujan, H. D., Morotta, A., Mowatt, M. R., Sciaky, N., Lippincott-Schwartz, J. and Nash, T. E. (1995). Developmental induction of Golgi structure and function in the primitive eukaryote Giardia lamblia. *J. Biol. Chem.* **270**, 4612-4618.

McCaffery, J. M., Faubert, G. M. and Gillin, F. D. (1994). *Giardia lamblia*: Traffic of a trophozoite surface protein and a major cyst wall epitope during growth, encystation and antigenic switching. *Exp. Parasitol.* **79**, 236-249.

McCaffery, J. M. and Gillin, F. D. (1994). *Giardia lamblia*: ultrastructural basis of protein transport during growth and encystation. *Exp. Parasitol.* **79**, 220-235.

Meyer, E. A. (1994). Giardia as an organism. In *Giardia: from Molecules to Disease* (ed. R. C. A. Thompson, J. A. Reynoldson and A. J. Lynberg), pp. 3-13. CAB International, Wallingford, UK.

Porter, K. R. (1953). Observations on a submicroscopic basophilic component of the cytoplasm. *J. Exp. Med.* **97**, 727-750.

Preuss, D., Mulholland, J., Kaiser, C. A., Orlean, P., Albright, C., Rose, M. D., Robbins, P. W. and Botstein, D. (1991). Structure of the yeast endoplasmic reticulum: localization of ER proteins using immunofluorescence and immunoelectron microscopy. *Yeast* **7**, 891-911.

Rambourg, A. and Clermont, Y. (1990). Tridimensional electron microscopy: Structure of the Golgi apparatus. *Eur. J. Cell Biol.* **51**, 189-200.

Reiner, D. S., McCaffery, M. and Gillin, F. D. (1990). Sorting of cyst wall proteins to a regulated secretory pathway during differentiation of the primitive eukaryote, Giardia lamblia. *Eur. J. Cell Biol.* **53**, 142-153.

Sitia, R. and Meldolesi, J. (1992). Endoplasmic reticulum: a dynamic patchwork of specialized subregions. *Mol. Biol. Cell* **3**, 1067-1072.

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

Soltys, B. J. and Gupta, R. S. (1992). Interrelationships of endoplasmic reticulum, mitochondria, intermediate filaments and microtubules - a quadruple fluorescence labeling study. *Biochem. Cell Biol.* **70**, 1174-1186.

Soltys, B. J. and Gupta, R. S. (1994a). Immunoelectron microscopy of *Giardia lamblia* cytoskeleton using antibody to acetylated α-tubulin. *J. Euk. Microbiol.* **41**, 625-632.

Soltys, B. J. and Gupta, R. S. (1994b). Presence and cellular distribution of a 60 kDa protein related to mitochondrial hsp60 in Giardia lamblia. *J. Parasitol.* **80**, 580-590.

Soltys, B. J. and Gupta, R. S. (1996). Immunoelectron microscopic localization of the 60-kDa heat shock protein (hsp60) in mammalian cells. *Exp. Cell. Res.* **222**, 16-27.

Tai, J.-H., Ong, S.-J., Chang, S.-C., and Su, H. M. (1993). Giardia virus enter *Giardia lamblia* WB trophozoite via endocytosis. *Exp. Parasitol.* **76**, 165-174.

Terasaki, M., Song, J., Wong, J. R., Weiss, M. J. and Chen, L. B. (1984). Localization of endoplasmic reticulum in living and glutaraldehyde-fixed cells with fluorescent dyes. *Cell* **38**, 101-108.

Tokuyasu, K. T. (1983). Present state of immunocryoultramicrotomy. *J. Histochem. Cytochem.* **4**, 164-167.

Tokuyasu, K. T. (1986). Applications of cryoultramicrotomy to immunocytochemistry. *J. Microsc.* **143**, 139-149.

Wanker, E. E., Sun, Y., Savitz, A. J. and Meyer, D. I. (1995). Functional characterization of the 180-kD ribosome receptor in vivo. *J. Cell Biol.* **130**, 29-39.

Warren, C. and Wickner, W. (1996). Organelle inheritance. *Cell* **84**, 395-400.

Wright, R., Basson, M., D'Ari, L. and Rine, J. (1988). Increased amounts of HMG-CoA reductase induce 'karmellae': a proliferation of stacked membrane pairs surrounding the yeast nucleus. *J. Cell Biol.* **107**, 101-114.

(Received 21 February 1996 - Accepted 8 April 1996)

CHAPTER 5

Cloning of the HSP70 *(dnaK)* Genes from *Rhizobium meliloti* and *Pseudomonas cepacia:* Phylogenetic Analyses of Mitochondrial Origin Based on a Highly Conserved Protein Sequence.

In this chapter the evolution of mitochondria is investigated. Previous studies indicated that mitochondria originated from the α-proteobacterial subdivision of eubacteria. To study this issue based on HSP70 sequences, HSP70 genes were cloned and sequenced from *Rhizobium meliloti* and *Pseudomonas cepacia*, which are members of the α- and β-subdivisions of proteobacteria, respectively. Sequence comparison and phylogenetic analysis of these two sequences with other sequences from the data base including mitochondrial sequences has been carried out. My contribution in this chapter was the cloning and sequencing of HSP70 genes from those species including the growth of bacteria, isolation of genomic DNA, cloning the PCR products, and isolating and sequencing the genomic clones. In addition, I was also involved in the phylogenetic analysis of the data and in the preparation of the manuscript. The sequence data for *R. meliloti* and *P. cepacia* HSP70 genes have been deposited in the GenBank data base under the accession numbers L36602 and L36603, respectively.

JOURNAL OF BACTERIOLOGY, Dec. 1994, p. 7748–7753
0021-9193/94/$04.00+0

Vol. 176, No. 24

Cloning of the *hsp70* (*dnaK*) Genes from *Rhizobium meliloti* and *Pseudomonas cepacia*: Phylogenetic Analyses of Mitochondrial Origin Based on a Highly Conserved Protein Sequence

MIZIED FALAH AND RADHEY S. GUPTA*

Department of Biochemistry, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

Received 14 June 1994/Accepted 7 October 1994

The genes for *hsp70* (or *dnaK*) have been cloned and sequenced from *Rhizobium meliloti* and *Pseudomonas cepacia*, two bacterial species belonging to the α- and β-subdivisions of gram-negative proteobacteria, respectively. On the basis of global alignment of HSP70 proteins, several sequence signatures have been identified that are distinctive of mitochondrial homologs and gram-negative proteobacteria on the one hand and the chloroplasts and cyanobacteria on the other. Detailed phylogenetic analyses of HSP70 sequences from various eubacteria and eukaryotic organellar and cytosolic homologs support the inference regarding the origin of mitochondria from a member of the α-proteobacteria and of chloroplasts from cyanobacteria. The analysis presented here also suggests a monophyletic origin of the mitochondrial homologs.

The members of the 70-kDa heat shock family of proteins (HSP70) are ubiquitously present in all eukaryotic and prokaryotic organisms and constitute one of the most conserved classes of proteins found in all species (1, 11, 16). Because of their ubiquitous presence and high degree of sequence conservation, HSP70 homologs provide an attractive system for molecular evolutionary studies (1, 10, 11, 13). On the basis of morphological and biochemical similarities between eukaryotic cell organelles such as mitochondria and chloroplasts and certain groups of eubacteria, it has been proposed that the former arose by serial endosymbiotic capture of the latter species (see references 6, 8, 14, and 19). This view has received support from phylogenetic analysis of gene sequences. In the case of mitochondria, studies based on rRNA and cytochrome *c* sequences indicate that the endosymbionts that gave rise to mitochondria are (or were) related to the α-subdivision of proteobacteria, which have been referred to as purple bacteria in earlier studies (2, 6, 14, 18, 21). However, because of the large variation in the sizes of rRNAs in eukaryotic species (see reference 5) and in the cytochrome *c* sequences in prokaryotic species (see reference 2), it is necessary to confirm this inference using other molecules which show a higher degree of sequence and structural conservation. To examine this question, the cloning and sequencing of *hsp70* genes from *Rhizobium meliloti* and *Pseudomonas cepacia*, which are members of the α-subdivision and β-subdivision of the proteobacteria (20), were undertaken. Global comparisons and detailed phylogenetic analyses of the organellar, eubacterial, and eukaryotic HSP70 sequences presented here strongly support the endosymbiotic hypothesis for the origin of mitochondria from a member of the α-subdivision of proteobacteria and of chloroplasts from a cyanobacterial species.

Cloning of *hsp70* (*dnaK*) genes from *R. meliloti* and *P. cepacia*. The approach used for cloning the *hsp70* genes is described in earlier work (4, 12). When PCR was carried out with either *R. meliloti* or *P. cepacia* DNA using a set of degenerate oligonucleotide primers for two highly conserved regions in HSP70 homologs, strong and specific amplification of expected 0.65-kbp fragments was observed (results not shown; see references 4 and 12). The nucleotide and the deduced amino acid sequences of the amplified fragments were unique; however, they showed extensive similarity to other HSP70 sequences. To isolate the *hsp70* gene, the genomic DNA from *R. meliloti* or *P. cepacia* was digested with several restriction enzymes (viz., *Hin*dIII, *Eco*RI, *Bam*HI, and *Xho*I), blotted, and hybridized to the ^{32}P-labeled cloned PCR probes. In Southern blots of *R. meliloti*, the cloned probe hybridized to single unique fragments in the size range of 1.8 to 8.0 kbp (results not shown). For *Eco*RI-digested DNA, specific hybridization to a fragment of about 6.0 kbp was observed, and this fragment was subcloned in a plasmid vector as described previously (see references 4 and 12). The restriction map of the fragment indicated that the *hsp70* gene sequence was located at the 3′ end, and this region was sequenced in both directions. After the putative initiator ATG, the insert contained an open reading frame of 1,923 bp that corresponded to the HSP70 protein. In genomic Southern blots of *P. cepacia*, the cloned PCR probe hybridized to single unique fragments in the size range of 2.5 to 6.0 kbp (results not shown). In all cases, hybridization was seen to only one fragment, indicating that only a single *hsp70-dnaK* gene is present in this species. A 3.5-kbp *Eco*RI fragment that hybridized to the cloned probe was subcloned, and it was found to contain the entire *hsp70* gene sequence. The G+C contents of the *hsp70* gene for *R. meliloti* and *P. cepacia* were found to be 60.8 and 62.7%, respectively. In accordance with their high G+C contents, the codon usage in these species showed a strong preference for codons with either G or C in the third position (82.8 and 87.0% for *R. meliloti* and *P. cepacia hsp70*, respectively).

Sequence alignment and signature sequences. A pairwise similarity matrix of HSP70 sequences from organelles and representative prokaryotic and eukaryotic species is shown in Table 1. The minimum amino acid identity between any two HSP70 homologs is about 40%, showing its highly conserved nature. As seen from Table 1, the mitochondrial homologs from various species exhibited highest similarity to the homologs from the α-subdivision of proteobacteria (viz., *R. meliloti*, *Brucella abortus*, and *Caulobacter crescentus*). For the chloroplast homologs, maximum identity (71 to 73%) was observed for the cyanobacterial HSP70 sequence. These results support the concept that α-proteobacteria and cyanobac-

* Corresponding author. Phone: (905) 525-9140, ext. 22639. Fax: (905) 522-9033.

TABLE 1. Similarity matrix of HSP70 sequences

Species	Amino acid identity with other HSP70s (%)														
	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Rhizobium meliloti α (A)	74.3	82.4	67.2	66.0	60.2	60.4	**62.1**	**66.6**	**64.6**	56.2	57.1	59.0	64.5	47.3	48.7
Caulobacter crescentus α (B)		72.7	66.1	65.8	59.4	58.5	**59.4**	**62.6**	**63.7**	53.2	54.1	55.6	62.7	48.0	50.7
Brucella ovis α (C)			67.2	64.4	59.3	58.3	**59.0**	**63.7**	**62.2**	53.8	54.4	56.3	63.1	45.5	48.2
Escherichia coli γ (D)				72.7	59.4	56.1	57.1	59.7	59.4	52.7	54.2	55.3	58.3	46.7	44.0
Pseudomonas cepacia β (E)					58.6	58.2	54.3	57.7	58.0	52.7	52.8	53.4	59.2	45.4	47.5
Chlamydia trachomatis (F)						57.9	54.6	57.4	56.1	55.4	55.8	57.1	60.7	44.0	45.8
Synechocystis sp. (G)							52.5	56.3	53.5	**70.8**	**73.0**	**72.7**	61.3	47.6	49.4
Saccharomyces cerevisiae (m) (H)								58.9	63.8	51.1	54.2	53.5	55.0	45.4	48.0
Pea (m) (I)									59.4	51.4	54.1	53.9	57.4	47.9	46.9
Mouse (m) (J)										49.5	54.5	53.9	55.3	46.0	47.7
Pseudomonas lutherii (c) (K)											73.8	71.8	58.3	43.3	44.3
Cryptomonas phi (c) (L)												76.6	55.7	45.9	48.8
Porphyra umbilicalis (c) (M)													57.4	45.8	46.8
Bacillus subtilis (N)														45.8	47.6
Maize (O)															75.7
Human (P)															

[a] The percentages of identity between pairs of sequences were calculated by the program PALIGN from the PCGENE program package employing the structure-gene matrix and unit gap and open gap costs of 1 and 7, respectively. The letters m and c in parentheses refer to sequences from mitochondrial and chloroplast homologs, respectively.

teria are the closest relatives of mitochondria and chloroplasts, respectively (see references 6, 15, and 21).

A multiple alignment of HSP70 sequences from organelles and representative prokaryotic (including *R. meliloti* and *P. cepacia*) and eukaryotic species is shown in Fig. 1. The C-terminal 50 to 100 residues which show considerable variation, particularly among distant species (10, 12), are not shown in this alignment. As has been seen, all mitochondrial homologs contained a characteristic N-terminal targeting sequence which is required for mitochondrial import. In our earlier work, a number of sequence signatures were identified on the basis of global alignment of HSP70 sequences which provided clear distinction between eukaryotic cytosolic homologs and prokaryotic homologs and between the gram-positive bacteria and the gram-negative species (10, 11, 13). As seen from Fig. 1, all of the organellar HSP70s contained various sequence features that are characteristic of the gram-negative bacteria but none corresponding to eukaryotic cytosolic homologs, supporting the interpretation that they originated from eubacteria via endosymbiosis. From the sequence alignment presented in Fig. 1, we have also identified a number of signature sequences that are uniquely shared by mitochondrial HSP70s and the homologs from the proteobacteria (filled diamonds in Fig. 1) but that are not present in other eubacteria. Several of these changes were seen only in mitochondrial homologs and the members of the α-subdivision, indicating a closer relationship between these groups. A four-amino-acid insert (amino acids 210 to 213 in the *P. cepacia* sequence) which is present in HSP70s from the β- and γ-subdivisions of proteobacteria was not found in any of the homologs from mitochondria or in the α-subdivision members, indicating a closer relationship between the last two groups. The sequence alignment has also identified a number of sequence features that are uniquely shared by the chloroplast HSP70s and those from cyanobacteria (marked with open diamonds in Fig. 1), indicating a closer evolutionary relationship between these groups.

Phylogenetic analyses. Phylogenetic analysis was initially performed on a set of 25 HSP70 sequences, which included mitochondrial and chloroplast homologs as well as sequences from representative eubacterial, archaebacterial, and eukaryotic (cytosolic) species. For these purposes, only regions of HSP70 which could be aligned without ambiguity in all homologs (i.e., from amino acids 3 to 557 in the *R. meliloti* sequence) were considered. A neighbor-joining bootstrap consensus tree based on these sequences is shown in Fig. 2A. The tree shows a clear distinction (99% of the time) between archaebacteria and gram-positive bacteria on the one hand and the gram-negative bacteria and eukaryotic cytosolic and organellar homologs on the other. This relationship is the same as that noted in our recent studies (11, 13), and it is supported by the presence of common signature sequences (e.g., the presence or absence of a large insert in the N-terminal quadrant) that are unique to homologs from these groups. The neighbor-joining tree shows a strong affinity (in 100% of the bootstraps) between the chloroplasts and cyanobacterial HSP70 sequences, which supports the evolution of chloroplasts from a species related to this group of organisms. In the tree shown in Fig. 2A, members of the α-, β-, and γ-subdivisions of proteobacteria were often found to group together (84% of the time), and the mitochondrial homologs evolved from an ancestor common to this group. However, in 13% of the bootstraps (not shown), *P. cepacia* and *Escherichia coli* branched off earlier than either mitochondrial homologs or members of the α-proteobacteria, and in these cases a specific relationship of mitochondria to the α-proteobacteria was indicated. In contrast, in no case did mitochondrial homologs share a common ancestor with the β- and γ-subdivision members, with the exclusion of α-proteobacteria. The bootstrap data also show that the mitochondrial homologs grouped together 86% of the time, suggesting that they constitute a monophyletic group. Parsimony analysis of the sequences yielded a single most parsimonious tree requiring 3,512 amino acid substitutions (Fig. 2B). In this tree, *P. cepacia* and *E. coli* branched off earlier than the mitochondrial homologs or the members of the α-subdivision of purple bacteria, and a most recent common ancestor between the last two groups of sequences was indicated. In the parsimony tree, all mitochondrial homologs again formed a monophyletic group, and the branching order within this group was the same as that seen in the neighbor-joining tree. A specific relationship of chloroplast homologs to a

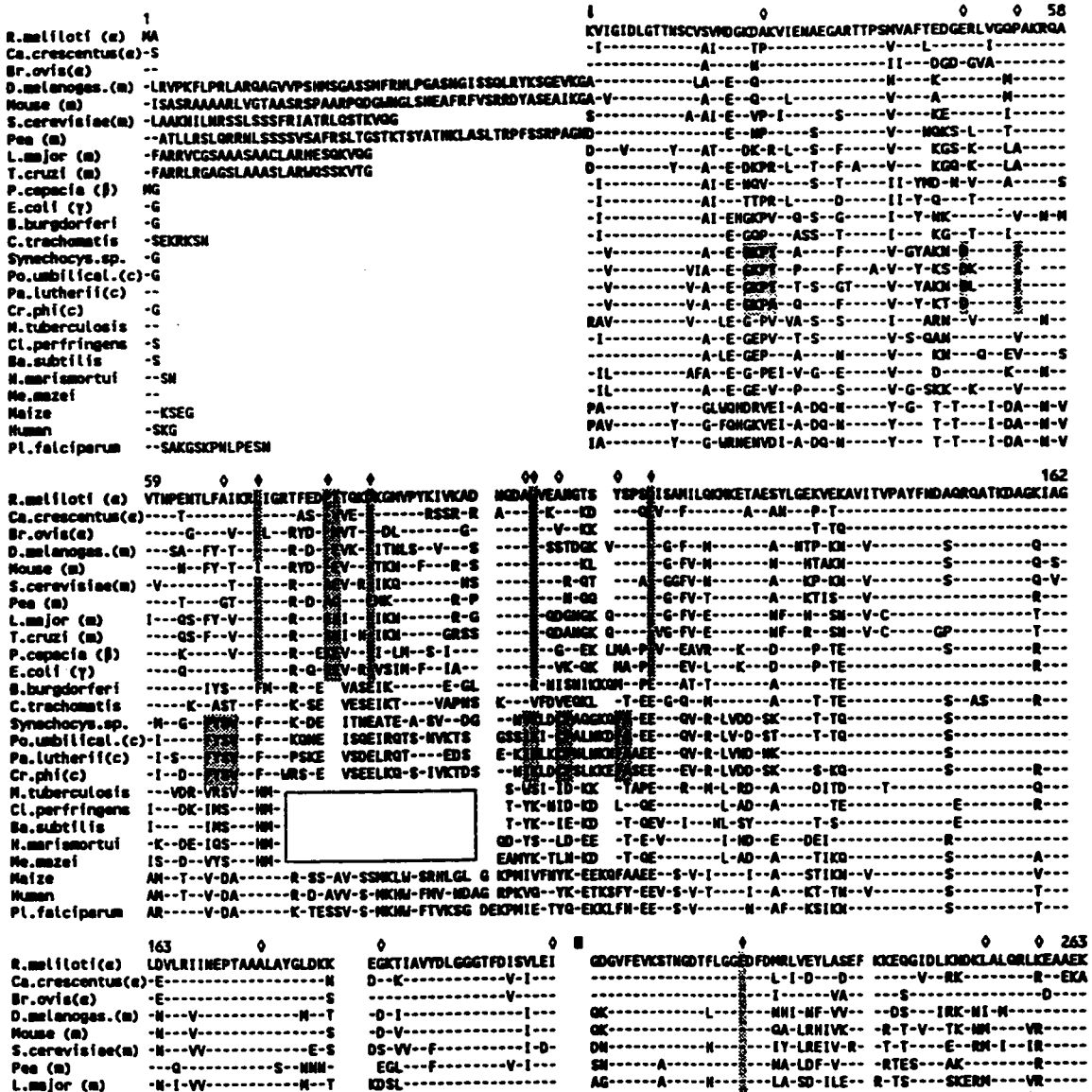
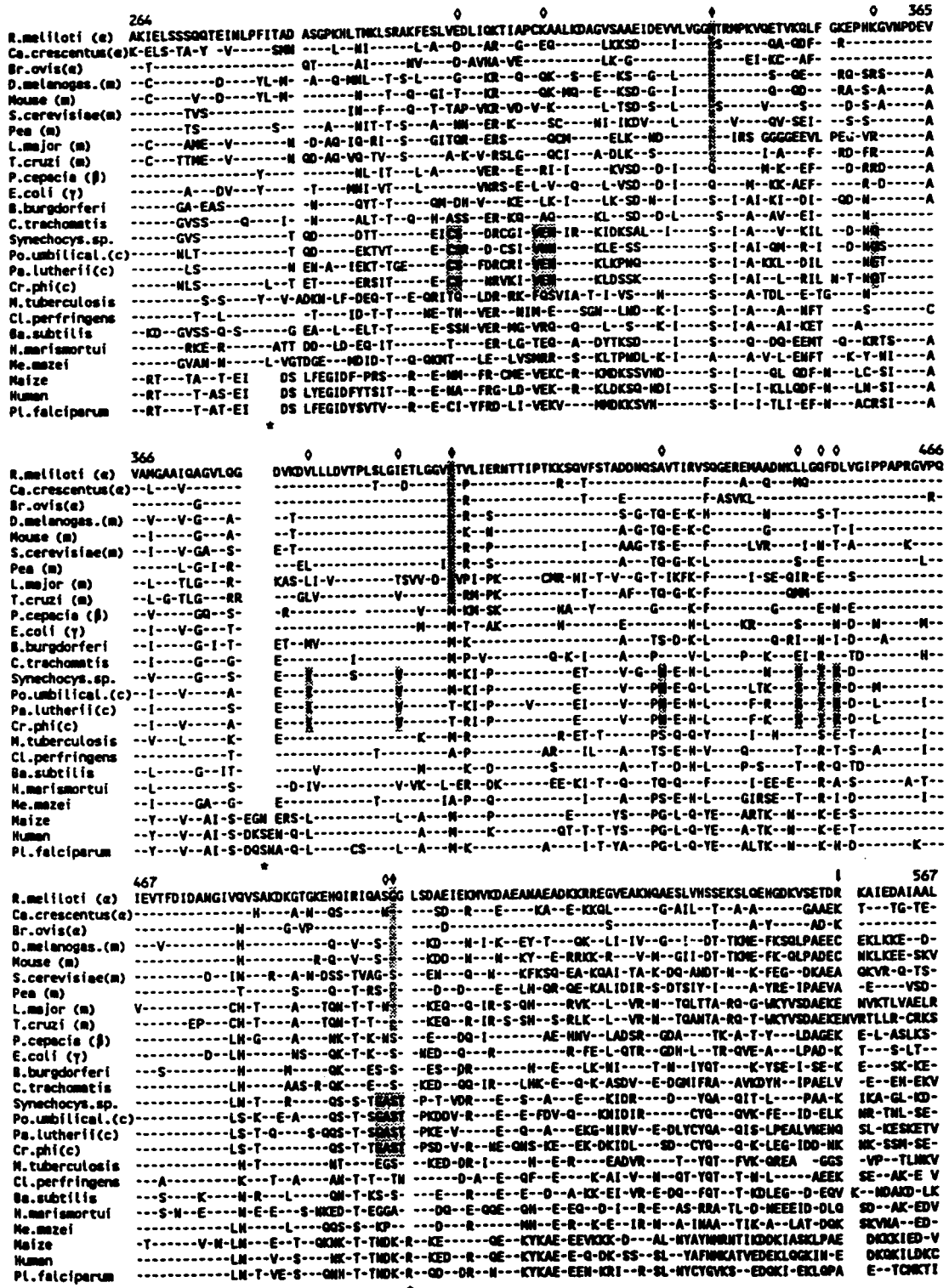


FIG. 1. Alignment of *R. meliloti* HSP70 sequence with other prokaryotic, organellar, and eukaryotic-cytosolic sequences. The C-terminal 63 residues in *R. meliloti* and the corresponding regions in other homologs which show considerable variation in distant homologs are not shown. The numbers at the top refer to positions in the *R. meliloti* sequence. The boxed area and the symbols describe different types of signature sequences. Boxed area, region in which a large insert is present in all HSP70s, except those from archaebacteria and gram-positive bacteria (11, 13); ★, signatures providing distinction between eukaryotic cytosolic and prokaryotic homologs (not all signatures of this kind are shown); ◆, signatures (shaded region shown below) that are common between mitochondrial homologs and those from purple eubacteria; ◇, signatures (boldface letters in the shaded region) that are common between chloroplast homologs and cyanobacteria; ■, the four-amino-acid insert that is specifically present in homologs from the β- and γ-subdivisions of proteobacteria; –, identity with the amino acid in the top line. The open spaces indicate gaps or deletions in the sequence. The species names are as follows: *B. burgdorferi*, *Borrelia burgdorferi*; *Ba. subtillis*, *Bacillus subtilis*; *Br. ovis*, *Brucella ovis*; *C. trachomatis*, *Chlamydia trachomatis*; *Ca. crescentus*, *Caulobacter crescentus*; *Cl. perfringens*, *Clostridium perfringens*; *Cr. phi*, *Cryptomonas phi*; *D. melanogaster*, *Drosophila melanogaster*; *H. marismortui*, *Halobacterium marismortui*; *L. major*, *Leishmania major*; *M. tuberculosis*,



Mycobacterium tuberculosis; *Me. mazei*, *Methanosarcina mazei*; *Pa. lutherii*, *Pavlova lutherii*; *Pl. falciparum*, *Plasmodium falciparum*; *Po. umbilicalis*, *Porphyra umbilicalis*; *T. cruzi*, *Trypanosoma cruzi*. m and c, mitochondrial and chloroplast homologs, respectively; α, β, and γ, denote the α-, β-, and γ- subdivisions of proteobacteria, respectively. The arrowheads mark the two ends of the sequence region which was employed in phylogenetic analysis.

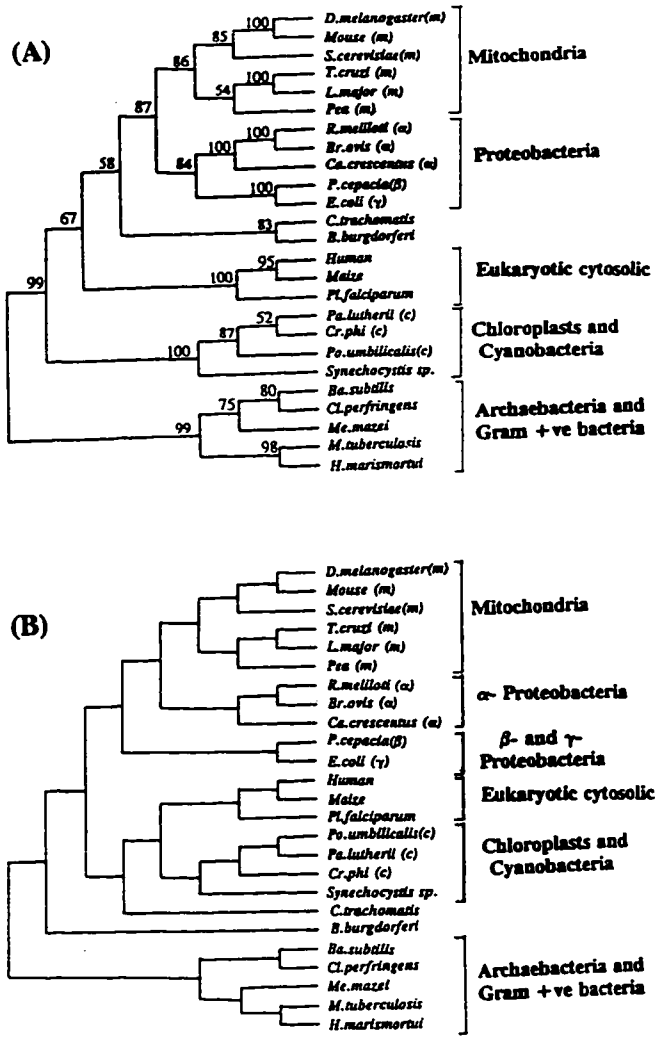


FIG. 2. Phylogenetic trees based on HSP70 sequences. (A) A consensus neighbor-joining tree based on 25 sequences obtained after 200 bootstraps. The numbers at the forks indicate the percentages of times that the species, which are to the right of the forks, grouped together from 200 trees. (B) Parsimony tree of the same species. The sequence alignment shown in Fig. 1 (between the arrowheads) was employed for these analyses. The species names are the same as those indicated in the legend to Fig. 1. The trees were constructed using the programs BOOT, NEIGHBOR (17), CONSENSE, and PROTPARS from the PHYLIP version 3.5 program package (3).

cyanobacterium HSP70 was again supported by the parsimony tree.

To further investigate the question of whether mitochondrial homologs are monophyletic or polyphyletic in origin, phylogenetic analysis was carried out on a set of 29 HSP70 sequences which included 9 mitochondrial homologs, 9 sequences from eubacteria, 11 eukaryotic cytosolic homologs, and 2 archaebacterial sequences. Gray et al. (6, 7) have previously suggested that if the mitochondrial homologs are of monophyletic origin, then the tree topology based on mitochondrial sequences should be similar to that based on nuclear-coded cytosolic homologs which evolved independently of these. A neighbor-joining bootstrap consensus tree based on these sequences is shown in Fig. 3. The branch lengths in this tree are proportional to the distances between the species. This

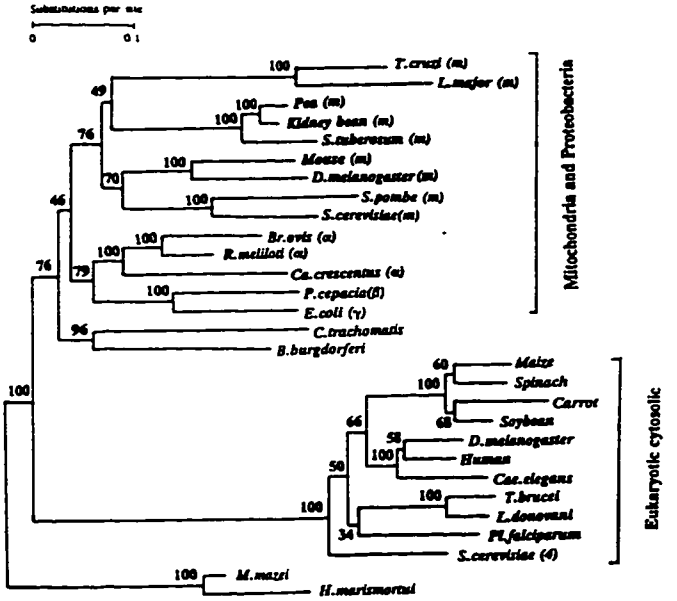


FIG. 3. A neighbor-joining bootstrap consensus tree of mitochondrial and cytosolic HSP70 homologs obtained after 100 bootstraps. The tree was constructed using the programs BOOT, NEIGHBOR, CONSENSE, and PROTDIST from the PHYLIP 3.5 package (3). The horizontal branch lengths indicate the distances between the species. The scale bar denotes 0.1 substitution per amino acid. *Cal. elegans*, *Caenorhabditis elegans*. The number 4 in *S. cerevisiae* refers to the gene *SSA4*. The other species names and abbreviations are the same as those given in the legend to Fig. 1.

tree has been rooted using the HSP70 homologs from the archaebacterial species, which are indicated to be an ancient group on the basis of earlier work (11, 12). As can be seen, all of the cytosolic homologs formed a monophyletic cluster with strong affinity for each other (100 of 100), and these were separated from the eubacterial and organellar HSP70s by a long branch length. The various mitochondrial homologs also formed a monophyletic cluster with reasonable affinity (76 of 100) and shared a most recent eubacterial ancestor with the proteobacteria. Within the mitochondrial cluster, strong affinities were observed for the species constituting the animal, plant, fungal, and protist groups (100 of 100). However, the orders of branching of some of these groups, particularly fungi and protists, were found not to be reliable. (In Fig. 3, 49% of the time the protist species grouped with the plants; however, in the bootstrap sets not shown, 26% of the time they showed closer association with the animal and fungi species and 6% of the time they branched more deeply than either animals, plants, or fungi.) These results are comparable to the results obtained for the nuclear-cytosolic HSP70 sequences. Similarly to the mitochondrial homologs, strong affinities of the species within the plant and animal groups were observed. However, the relative branching orders of the species are once again not reliable because of uncertainties in the positions of fungus and protist species. Amongst the bootstrap sets not included in the consensus tree (Fig. 3), the yeast *Saccharomyces cerevisiae* was found to branch with the animal and fungal species 25% of the time, and 19 times of 100 it branched with the protist species. These observations indicate that except for the uncertain positions of fungal and protist species, the tree topologies in the two branches of the tree (viz., mitochondrial and nucleocytosolic) are generally similar.

Thus, sequence features and the phylogenetic analysis based on the highly conserved HSP70 family of protein sequences are generally in good agreement with the inferences reached in earlier studies regarding the endosymbiotic origin of mitochondria and chloroplasts from the α-proteobacterial and cyanobacterial groups of species (2, 6, 14, 15, 18, 21). The xenogenous origin of organellar homologs from the gram-negative eubacterial species is clearly evident from the presence in all organellar HSP70 sequences of a number of signature sequences that are characteristics of this group of eubacteria. At the same time, the organellar HSP70s lack various signature sequences that are common in all of the eukaryotic cytosolic homologs. A specific relationship between the mitochondrial HSP70s and those from the α-proteobacteria is suggested by a number of observations, including (i) the observation of highest sequence identity between these two groups of homologs, (ii) the shared presence in these sequences of a number of unique signatures that are not found in other homologs, and (iii) in phylogenetic trees based on HSP70 sequences, a shared most recent common ancestor between the branch leading to mitochondria and either the gram-negative proteobacteria (neighbor-joining tree) or more specifically the α-subdivision members of the proteobacteria (parsimony tree). However, a closer relationship between mitochondrial HSP70s and those from α-proteobacteria is suggested by the fact that both of these groups lacked a four-amino-acid insertion that is present in the β- and γ-subdivision members. The origin of mitochondria from a member of the α-subdivision of proteobacteria and of chloroplasts from a cyanobacterium species is also supported by recent phylogenetic studies based on the HSP60 (GroEL) and HSP10 (GroES) families of protein sequences (9).

All of the mitochondrial homologs which were considered in the present studies showed a monophyletic branching in both the neighbor-joining and the parsimony trees. The observed bootstrap scores for this node in the neighbor-joining trees (76 and 87) indicates that this inference was reasonably supported. In an earlier study, which was based on mitochondrial small subunit rRNA sequences, Gray et al. (8) suggested a polyphyletic origin of mitochondria to account for the observations that the green alga *Chlamydomonas reinhardtii* did not branch with the higher plants and that an anomalous very early branching of the higher plant sequences was seen. However, as noted by Gray (5), since the plant mitochondria are known to readily incorporate and harbor a variety of promiscuous DNAs, including both nuclear and chloroplast rRNA sequences, the anomalous branching of higher plant sequences is phylogenetically not very reliable. In view of this reported anomaly (7), it would be of interest to determine the branching position of mitochondrial HSP70 from *C. reinhardtii* to see whether it branches with or separately from the mitochondrial homologs from higher plant species.

The sequence data for *R. meliloti* and *P. cepacia hsp70* genes have been deposited in the GenBank database under the accession numbers L36602 and L36603, respectively.

We acknowledge the assistance of B. Singh for help with the PCR experiment and Enzo Ruscitti for computing the pairwise similarity scores. We thank K. B. Freeman and G. B. Golding for helpful comments on the manuscript and Barbara Sweet for secretarial work. G. B. Golding is also thanked for access to his computer for phylogenetic analyses.

This work was supported by a research grant from the Medical Research Council of Canada to R.S.G.

REFERENCES

1. **Boorstein, W. R., T. Zeigelhoffer, and E. A. Craig.** 1994. Molecular evolution of the HSP70 multigene family. J. Mol. Evol. **38**:1–17.
2. **Dickerson, R. E.** 1980. Evolution and gene transfer in purple photosynthetic bacteria. Nature (London) **283**:210–212.
3. **Felsenstein, J.** 1993. PHYLIP manual, version 3.5. Herbarium. University of Washington, Seattle, Wash.
4. **Galley, K. A., B. Singh, and R. S. Gupta.** 1992. Cloning of HSP70 gene from *Clostridium perfringens* using a general polymerase chain reaction based approach. Biochim. Biophys. Acta **1130**:203–208.
5. **Gray, M. W.** 1988. Organelle origins and ribosomal RNA. Biochem. Cell Biol. **66**:325–348.
6. **Gray, M. W.** 1992. The endosymbiont hypothesis revisited. Int. Rev. Cytol. **141**:233–357.
7. **Gray, M. W., R. Cedergren, Y. Abel, and D. Sankoff.** 1989. On the evolutionary origin of plant mitochondria and its genome. Proc. Natl. Acad. Sci. USA **86**:2267–2271.
8. **Gray, M. W., and W. F. Doolittle.** 1982. Has the endosymbiont hypothesis been proven? Microbiol. Rev. **46**:1–42.
9. **Gupta, R. S.** Evolution of the chaperonin families (HSP60, HSP10 and TCP-1) of proteins and the origin of eukaryotic cells. Mol. Microbiol., in press.
10. **Gupta, R. S., K. Aitken, M. Falah, and B. Singh.** 1994. Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. Proc. Natl. Acad. Sci. USA **91**:2895–2899.
11. **Gupta, R. S., and G. B. Golding.** 1993. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria and eukaryotes. J. Mol. Evol. **37**:573–582.
12. **Gupta, R. S., and B. Singh.** 1992. Cloning of the HSP70 gene from *Halobacterium marismortui*: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model for the evolution of the HSP70 gene. J. Bacteriol. **174**:4594–4605.
13. **Gupta, R. S., and B. Singh.** Evolutionary analysis of HSP70 protein sequences and a chimeric model for the origin of eukaryotic cell nucleus. Curr. Biol., in press.
14. **Margulis, L.** 1981. Symbiosis in cell evolution. W. H. Freeman & Co., San Francisco.
15. **Morden, C. W., C. F. Delwiche, M. Kuhsel, and J. D. Palmer.** 1992. Gene phylogenies and the endosymbiotic origin of plastids. BioSystems **28**:75–90.
16. **Morimoto, R. I., A. Tissières, and C. Georgopoulos (ed.).** 1994. Biology of the heat shock proteins and molecular chaperones, p. 1–610. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
17. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method of reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.
18. **Schwartz, R. M., and M. O. Dayhoff.** 1978. Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. A perspective is derived from protein and nucleic acid sequence data. Science **199**:395–403.
19. **Taylor, F. J. R.** 1987. An overview of the status of evolutionary cell symbiosis theories. Ann. N. Y. Acad. Sci. **503**:1–16.
20. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51**:221–271.
21. **Yang, D., Y., Oyaizu, H. Oyaizu, G. J. Olsen, and C. R. Woese.** 1985. Mitochondrial origins. Proc. Natl. Acad. Sci. USA **82**:4443–4447.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

The results based on HSP70 sequence analysis presented in this thesis provide strong evidence for and novel insights into the evolutionary relationships of prokaryotes, eukaryotes and organelles. Based on the HSP70 sequence, we have determined the placement of *Mycoplasma capricolum* in the phylogeny. Our study strongly suggested that this bacterium, although it lacks a cell wall, is evolutionarily related to the gram-positive group of bacteria. The sequence alignment of HSP70 homologs from various species has revealed signature sequences that are characteristic of gram-positive bacteria and archaebacteria on one end, and other signature sequences that are characteristic of gram-negative bacteria and eukaryotes (including organelles) on the other. Similar to archaebacteria and gram-positive bacteria, mycoplasma species lack the 23-27 amino acid insert that is found in HSP70 homologs from gram-negative bacteria and eukaryotes. The absence of this insert in homologs from mycoplasma strongly indicates that this species is evolutionarily related to gram-positive bacteria and archaebacteria. Gram-positive bacteria and archaebacteria in HSP70 trees always branch within each other in one clade. The HSP70-based phylogeny has shown that mycoplasma species branch within this clade and are closely related to the low-G+C-content gram-positive group of bacteria.

We have also added further evidence to the idea of the endosymbiotic origin of mitochondria from gram-negative bacteria. Pairwise sequence comparisons show that the various mitochondrial

HSP70 homologs exhibited higher similarity to the homologs from α-proteobacteria as compared to either β- and γ-proteobacteria or other groups of eubacteria. These results strongly indicate the evolution of mitochondria from α-proteobacteria. This inference has been further supported by the multiple alignment of HSP70 sequences from organelles and various representatives of prokaryotes and eukaryotes. The mitochondrial HSP70 homologs possess all sequence features characteristic of only the gram-negative bacteria. This provides strong evidence for the xenogenous origin of mitochondrial homologs from the gram-negative group of bacteria. Some unique sequence signatures are found only in HSP70 homologs from mitochondria and α-proteobacteria, but not found in other bacteria (namely β-and γ-proteobacteria), indicating a close evolutionary relationship between mitochondria and α-proteobacteria. Furthermore, the parsimony tree of HSP70 has shown that the mitochondrial homologs of HSP70 branch together in a distinct monophyletic group and they share a common ancestor only with homologs from α-proteobacteria; homologs from β- and γ-purple bacteria appear to have separated earlier. The above observations together constitute strong evidence supporting the view that the endosymbiont that gave rise to mitochondria was a member of the α-subdivision of gram-negative proteobacteria. Based on our analysis, all mitochondrial species branch in a monophyletic group indicating that the endosymbiotic event that gave rise to mitochondria likely occurred only once.

Giardia lamblia, based on our analysis has been found to be the earliest branching species of eukaryotes and its HSP70 genes for the eukaryotic cytosolic and ER homologs are paralogous and arose by gene duplication very early in the evolution of eukaryotic cells. Alignment of these sequences with other HSP70 counterparts, has revealed that the *G. lamblia* sequences conform to the sequence signatures found in the cytosolic and ER HSP70 homologs. The sequence signatures

shared between eukaryotic and gram-negative bacterial homologs of HSP70, were also found in *G. lamblia* homologs. Most prominent is the 23-27 amino acid insert in the N-terminal quadrant of the HSP70 homologs, which is present exclusively in eukaryotic homologs and those from gram-negative bacteria. The phylogenetic analyses show that ER-HSP70 of *G. lamblia* is the deepest eukaryotic homolog in the HSP70-based phylogeny and formed the root of the eukaryotic tree. The cytosolic HSP70 homolog formed the deepest branch of all cytosolic homologs. The placement of the ER and cytosolic homologs in the phylogeny of HSP70 indicates that *G. lamblia* has branched very early in the evolution of eukaryotes. Because eukaryotic cytosolic/ER HSP70 homologs, including those of *G. lamblia*, share sequence signatures (such as the 23-27 amino acid insert) only with gram-negative bacteria and because the eukaryotic cytosolic/ER homologs form a monophyletic clade branching with gram-negative bacteria, it is proposed that eukaryotic cytosolic/ER genes of HSP70 are derived from a gram-negative bacterial ancestor. However, it is well known that some genes in eukaryotes are closely related to their counterparts in gram-negative bacteria because they were gained in the endosymbiotic events that gave rise to mitochondria and chloroplasts. The close relationship of eukaryotic cytosolic/ER homologs of HSP70 to those of gram-negative bacteria is not due to the same endosymbiotic events, because *G. lamblia* is believed to have branched off from the eukaryotic lineage prior to the endosymbiotic event that gave rise to mitochondria. Thus, the possibility that the eukaryotic cytosolic/ER homologs of HSP70 could be derived from the gram-negative ancestor that gave rise to mitochondria is unlikely. These results thus greatly strengthen the conclusion that all eukaryotic cytosolic HSP70 sequences, including those from *G. lamblia*, have been derived from a gram-negative bacterium.

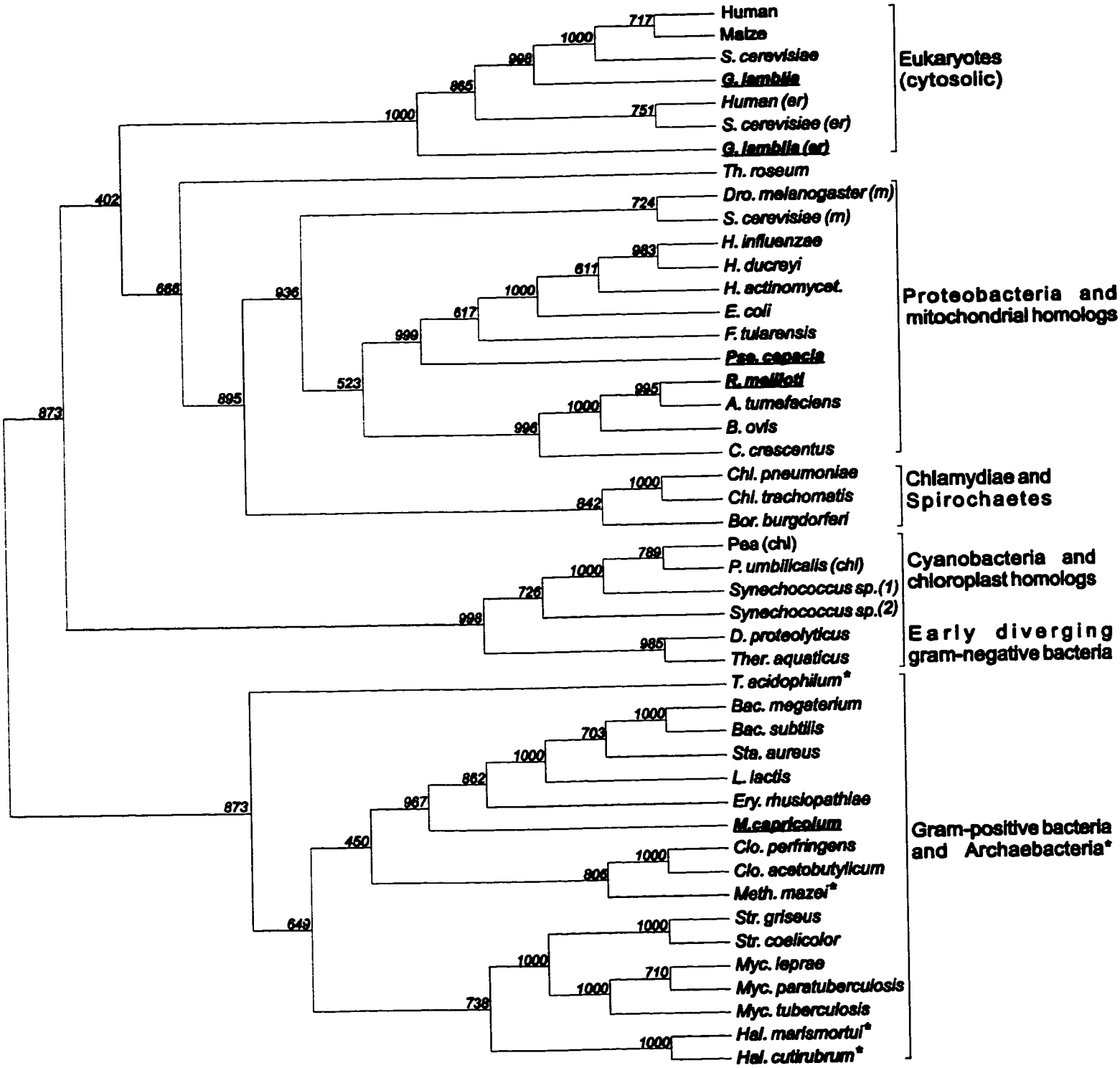
Our identification of ER in *G. lamblia* also has evolutionary implications for the origin of

eukaryotic cells. We have identified ER membranes at the ultrastructural level in *G. lamblia* cells using cryoelectron microscopy techniques and antibodies to ER-HSP70. In higher eukaryotes the ER is contiguous with the nucleus; as well, the ER and nucleus have similar composition (Alberts *et al.*, 1994). In view of these results and because of the existence of ER-HSP70 and ER compartment in the earliest diverging eukaryote, *G. lamblia*, we propose that the evolution of ER and the nucleus took place in concert. Our conclusion is consistent with the current dogma in eukaryotic evolution which predicts that the nucleus and the ER co-evolved early in the evolution of eukaryotic cells (Alberts *et al.*, 1994; Gupta and Golding, 1996), and therefore all eukaryotic cells should possess both organelles.

The branching pattern of the species that have been studied in this thesis is shown in Figure 3. The evolution of *M. capricolum*, *G. lamblia* and mitochondria is consistent with other studies based on rRNA and protein sequences (Gray, 1992; Hashimoto and Hasegawa, 1996; Maniloff, 1992; Sogin *et al.*, 1989; Woese *et al.*, 1980; Yang *et al.*, 1985). The fact that many different protein/nucleic acid sequences give similar evolutionary branching patterns of the above species strongly indicates that the observed evolutionary relationship is reliable and that the HSP70-based phylogeny is likely to represent the true phylogeny. However, with regard to the evolution of the eukaryotic cells, the HSP70-based phylogeny is at considerable variance with phylogenies based on other protein sequences (Doolittle and Brown, 1994; Gogarten *et al.*, 1989; Iwabe *et al.*, 1989). While the HSP70-based phylogeny postulates that eukaryotes have evolved from the gram-negative group of bacteria, the phylogeny for proteins of duplicated genes of elongation factors and ATPase subunits as well as other proteins, points to the evolution of eukaryotes from an archaebacterial ancestor. Based on HSP70 studies, the archaebacteria species do not branch with eukaryotic species

Figure 3. HSP70-based phylogeny.

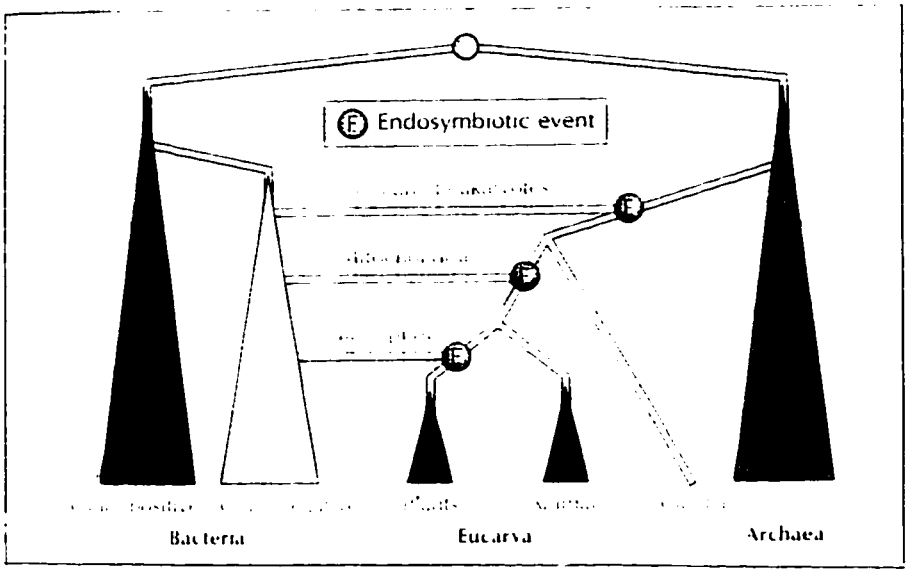
A consensus neighbor-joining tree obtained after 1,000 bootstraps. The numbers at the forks indicate the number of times the species to the right of the fork grouped together in the bootstrap trees. The tree is unrooted and has been arbitrarily divided at the approximate midpoint. The archaebacterial species are indicated (asterisks). The underlined species were studied in this thesis. This phylogeny was reproduced from Gupta *et al.*, 1997.


Human
Maize
S. cerevisiae
G. lamblia
Human (er)
S. cerevisiae (er)
G. lamblia (er)
Th. roseum
Dro. melanogaster (m)
S. cerevisiae (m)
H. influenzae
H. ducreyi
H. actinomycet.
E. coli
F. tularensis
Pse. cepacia
R. melliotl
A. tumefaciens
B. ovis
C. crescentus
Chl. pneumoniae
Chl. trachomatis
Bor. burgdorferi
Pea (chl)
P. umbilicalis (chl)
Synechococcus sp.(1)
Synechococcus sp.(2)
D. proteolyticus
Ther. aquaticus
T. acidophilum*
Bac. megaterium
Bac. subtilis
Sta. aureus
L. lactis
Ery. rhusiopathiae
M.capricolum
Clo. perfringens
Clo. acetobutylicum
Meth. mazei*
Str. griseus
Str. coelicolor
Myc. leprae
Myc. paratuberculosis
Myc. tuberculosis
Hal. marismortui*
Hal. cutirubrum*
717
1000
998
865
751
1000
402
724
983
611
1000
617
999
523
995
1000
996
895
873
686
936
1000
842
789
1000
726
998
985
703
1000
862
967
873
450
1000
806
649
1000
710
1000
738
1000
Eukaryotes (cytosolic)
Proteobacteria and mitochondrial homologs
Chlamydiae and Spirochaetes
Cyanobacteria and chloroplast homologs
Early diverging gram-negative bacteria
Gram-positive bacteria and Archaebacteria*

nor have they sequence signatures specific to eukaryotic homologs. To determine if HSP70-based phylogeny is also supported by other gene sequences, Golding and Gupta (1995) have carried out a detailed phylogenetic analysis of all protein sequences in the data bases for which information was available from the main groups of organisms (namely, archaebacteria, gram-negative and gram-positive bacteria and eukaryotes). Their study involved a total of 24 different proteins. It was found that a group of nine proteins including the elongation factors and ATPase subunits supported an evolutionary relationship between eukaryotes and archaebacteria. However, another group of seven proteins including HSP70, supported an evolutionary relationship between eukaryotes and gram-negative bacteria. The remaining eight proteins do not support or refute either of the relationships. It was also noted in this study that none of the protein sequences analysed indicated a close evolutionary relationship between eukaryotes and gram-positive bacteria. Thus, the eukaryotic species are either closely related to archaebacterial or to gram-negative eubacterial species. To explain these apparently contradictory phylogenies, Gupta and colleagues have proposed a chimeric model (Figure 4) for the origin of eukaryotic cells (Gupta and Golding, 1996; Gupta and Singh, 1994). They postulate that a primary fusion event between an archaebacterium and a gram-negative bacterium gave rise to the ancestral eukaryotic cell. Archaebacteria and gram-positive bacteria are ancient lineages and related to the universal ancestor of all organisms. The gram-negative bacteria are derived from the gram-positive group of bacteria. Subsequently, a fusion event took place between a member of the gram-negative group of bacteria and an archaebacterium of the eocyte group (Gupta and Golding, 1996; Gupta and Singh, 1994). After the fusion event an assortment or selection of genes took place such that while some genes were retained from the archaebacterial parent, others were derived from the gram-negative bacterium. The ancestral eukaryotic cell that

Figure 4. Evolutionary relationships within the prokaryotes and the origin of the eukaryotic cell nucleus.

This model proposed by Gupta and colleagues (Gupta and Golding 1993; Gupta and Singh 1994) and redrawn later by Irwin (1994) accounts for phylogenies based on different gene sequences. From HSP70 sequence features, the root of the universal tree is placed in the lineage from which archaebacteria and gram-positive bacteria evolved (Gupta and Golding 1993). Other prokaryotes evolved from these primary lineages. At a later time, a fusion event between an eocyte species and gram-negative bacterium gave rise to a pro-eukaryotic cell from which all contemporary eukaryotes evolved.


Endosymbiotic event
Bacteria
Eucarya
Archaea

emerged was *G. lamblia*-like in the sense that it lacked mitochondria and other organelles. At a later stage, endosymbiotic capture of an α-proteobacterium and cyanobacterium resulted in the origin of mitochondria and chloroplasts, respectively.

Thus, the chimeric model explains the fact that different cytosolic genes from eukaryotic species, exhibit a greater similarity to homologs from either archaebacteria or gram-negative bacteria. Furthermore, it readily explains certain characteristics of the eukaryotic cells that are either similar to archaebacteria or eubacteria (Woese, 1981; Zillig, 1991). As the difference between the structural organization of the prokaryotic and eukaryotic cell types is enormous and as all eukaryotic species, including the most primitive eukaryote, *G. lamblia*, contain a nucleus and ER, the transition from prokaryotes to eukaryotes which has no known intermediates appears to have happened by a sudden change explained in the model as a fusion event between two different prokaryotes.

On the basis of HSP70 and other protein sequences, archaebacteria and gram-positive bacteria are evolutionarily related. According to the chimeric model both these prokaryotic groups are ancient and are proposed to have evolved first before other organisms. If this is the case, how can we explain the fact that other protein sequences (e.g. elongation factors) indicate that these groups are evolutionarily distant. Gupta and Singh (1994) postulated that during early evolution with the rapid changes in environment (Kasting, 1993), the archaebacterial and gram-positive eubacterial genes have experienced different pressure of functional constraints and therefore some genes diverged slowly (e.g. HSP70) others (e.g. elongation factors) diverged faster. Genes which slowly diverged (e.g. HSP70, glutamate dehydrogenase and glutamine synthetase) show a close relationship between archaebacteria and gram-positive bacteria (Benachenhou-Lahfa *et al.*, 1993; Brown *et al.*, 1994; Gupta and Golding, 1993).

The chimeric model proposes that both archaebacteria and gram-positive bacteria are related to the common (universal) ancestor (Figure 4), and that archaebacteria and gram-positive bacteria are not monophyletic groups (Gupta and Singh, 1994). It is not known at present whether the universal ancestor was related to an archaebacterium, a gram-positive bacterium or a different kind of cell. Further genome exploration should be useful for understanding the evolutionary relationship between archaebacteria and gram-positive bacteria and their relationship to the common ancestor. Studies on cellular characteristics and molecular data, including genome sequences of species from gram-positive bacteria and archaebacteria as well as from other prokaryotes should also provide important insights as to which of these prokaryotes are the most ancient. The ancient nature of archaebacteria and gram-positive bacteria proposed by the model is consistent with the fact that most prokaryotes that grow and survive in thermophilic and anaerobic conditions which are similar to those of the early earth, belong to either archaebacteria or gram-positive bacteria (Joklik *et al.*, 1992; Pace, 1991; Schlegel and Jannasch, 1992; Woese, 1987; Woese *et al.*, 1990).

Another important point concerning the evolution of eukaryotic cells needs special attention. Based on phylogenies of different protein and nucleic acid sequences (e.g. EF-1α/Tu, EF-2/G, rRNA and others), all eukaryotic species form a monophyletic group branching from a single common ancestor. The monophyletic branching of eukaryotes is also strongly supported by phylogenetic analysis based on HSP70 sequences. Furthermore, the presence in all eukaryotic HSP70 homologs, including those of *G. lamblia*, of a number of sequence signatures that are not found in any other prokaryotic or orgenellar homologs, provide evidence that all extant eukaryotic species ranging from protists to fungi, plants and animals have been derived from a single ancestral cell. This would suggest that the fusion event that gave rise to all contemporary eukaryotes was unique (Gupta, 1997;

Gupta and Golding, 1996). It would be of great interest to experimentally recreate such a fusion event. However, the possibility that such a fusion will be successful appears remote because the prokaryotes that gave rise to the ancestral eukaryotic cell about 1.8 to 2.0 billion years ago may no longer exist and the specific conditions that led to this fusion are also not known. However, further studies on the evolution of prokaryotes may identify the archaebacterium and the gram-negative bacterium involved in this fusion. This is a great challenge to us and also emphasises the need and importance of studying prokaryotic evolution. Gupta and Golding (1996) have noted that the prokaryotic partners involved in the fusion event are most likely an archaebacterium from the eocyte group and a gram-negative bacterium. The inference that the archaebacterial partner involved in the fusion event was an eocyte or *Crenarchaeotal* archaebacterium is based on Rivera and Lake's results with the EF-1α and EF-2 sequences (Rivera and Lake, 1992). At present, an HSP70 gene sequence from the eocyte group of archaebacteria is not available. My previous attempts to clone the HSP70 gene from an eocyte species (namely *Sulfolobus solfataricus*) have not been successful. However, further studies in this regard are currently being carried out in the laboratory.

The HSP70 sequence from an eocyte will also provide insights into the monophyly/polyphyly of archaebacteria. As indicated previously, Woese and others argue that all archaebacteria (i.e. methanogenes, halophiles, and sulfo-thermophilic "eocytes") form a monophyletic cluster or domain separate from the eubacterial and eukaryotic domains (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989; Puhler *et al.*, 1989; Woese, 1987; Woese *et al.*, 1990). Although this view has gained wide acceptance, and is further endorsed by studies on biochemical features of archaebacteria (Zillig, 1991), Lake (1991) has contended that archaebacteria are polyphyletic. The studies by Lake and colleagues indicate that sulfo-thermophiles (eocytes) are more closely related to eukaryotes than to

other archaebacterial groups, whereas halophiles and methanogens are more closely related to eubacteria than to eukaryotes (Lake, 1991; Rivera and Lake, 1992).

In view of the above, two models have been proposed concerning the evolutionary relationship of archaebacteria to eukaryotes. In the archaebacterial tree favored by Woese and others (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989; Puhler *et al.*, 1989; Woese, 1987; Woese *et al.*, 1990), eukaryotes share a common ancestor with all of the archaebacteria (methanogenes, halophiles, and sulfo-thermophiles). However, in the eocyte tree supported by Lake's work (Lake, 1991; Rivera and Lake, 1992), eukaryotes share a common ancestor only with the sulfo-thermophiles (eocytes). The archaebacterial tree is not supported by HSP70-based phylogeny and the eocyte tree can not be proved due to lack of HSP70 sequences from eocytes. To settle this issue and test the hypothesis that one of the fusion partners was an "eocyte", the HSP70 sequence from this group of archaebacteria should be determined and analysed.

References

Acca M., M. Bocchetta, E. Ceccarelli, R. Creti, K. O. Stetter, and P. Cammarano. (1994). Updating mass and composition of archael and bacterial ribosomes: Archae-like features of ribosomes from the deep branching bacterium *Aquifex pyrophilus*. *System Appl. Microbiol.* 16:629-637.

Adam R. D. (1991). The biology of Giardia spp. *Microbiol. Rev.* 55:706-732.

Alberts B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. (1994). Molecular Biology of the Cell. Garland Publishing Inc., New York.

Benachenhou-Lahfa N., P. Forterre, and B. Labedan. (1993). Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. *J. Mol. Evol.* 36:335-346.

Boorstein W. R., T. Ziegelhoffer, and E. A. Craig. (1994). Molecular evolution of the HSP70 multigene family. *J. Mol. Evol.* 38:1-17.

Brennicke A., L. Grohmann, R. Hiesel, V. Knoop, and W. Schuster. (1993). The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. *FEBS* 325:140-145.

Brooks B. W., R. G. E. Murray, J. L. Johnson, E. Stackebrandt, C. R. Woese, and G. E. Fox. (1980). Red-pigmented micrococci: a basis for taxonomy. *Int. J. Sys. Bacteriol.* 30:627-646.

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

Brown J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle. (1994). Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. *J. Mol. Evol.* 38:566-576.

Bult C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, J. C. Venter, et al. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus janaschii* . *Science* 273:1058-1073.

Cain A. J. (1962). The evolution of taxonomic principles. *Symposium of the Society for General Microbiology* 12:1-3.

Cavalier-Smith T. (1987). The origin of eukaryotic and archaebacterial cells. *Ann. N. Y. Acad. Sci.* 503:17-54.

Cavalier-Smith T., (1992), p. 64-87. Origin of secondary metabolism. *In* Secondary metabolites: their function and evolution. Ciba Foundation symposium 171, Wiley, Chichester.

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

Craig E., P. J. Kang, and W. Boorstein. (1990). A review of the role of 70 kDa heat shock proteins in protein translocation across membranes. *Antonie van Leeuwenhoek* 58:137-146.

Craig E. A., B. D. Gambill, and R. J. Nelson. (1993). Heat shock proteins: molecular chaperones of protein biogenesis. *Microbiol. Rev.* 57:402-414.

Dayhoff M. O., L. T. Hunt, P. J. McLaughlin, and D. D. Jones (1972). Gene duplication in evolution: The globins. p. 17-30. *In* Vol. 5, M. O. Dayhoff (ed.). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Spring, MD.

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

De Rosa M., A. Gambacorta, and A. Gliozzi. (1986). Structure, biosynthesis, and physiochemical properties of archaebacterial lipids. *Microbiol. Rev.* 50:70-80.

Dickerson R. E. (1980). Evolution and gene transfer in purple photosynthetic bacteria. *Nature* 283:210-212.

Doi M., M. Wachi, F. Ishino, S. Tomioka, M. Ito, Y. Sakagami, A. Suzuki, and M. Matsuhashi. (1988). Determination of the DNA sequence of the *mre*B gene and of the gene products of the *mre* region that functions in the formation of the rod shape of *Escherichia coli* cells. *J. Bacteriol.* 170:4619-4624.

Doolittle W. F. and J. R. Brown. (1994). Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA* 91:6721-6728.

Dworkin, M. (1992). Prokaryotic diversity, p. 48-74. *In* A. Balows, H. G. Truper, D. Martin, W. Harder, and K.-H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, New York.

Edmonds C. G., P. F. Crain, R. Gupta, T. Hashizume, C. H. Hocart, J. A. Kowalak, S. C. Pomerantz,

K. O. Stetter, and J. A. McCloskey. (1991). Post-transcriptional Modification of tRNA in thermophilic *Archae* (Archaebacteria). J. Bacteriol. 173:3138-3148.

Eisen J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J. Mol. Evol.* 41:1105-1123.

Felsenstein J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Rev. Genet.* 22:521-565.

Flaherty K. M., C. DeLuca-Flaherty, and D. B. McKay. (1990). Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein [see comments]. *Nature* 346:623-628.

Fox G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese. (1980). The phylogeny of prokaryotes. *Science* 209:457-463.

Gogarten J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, and et al. (1989). Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* 86:6661-6665.

Golding G. B. and R. S. Gupta. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *J. Mol. Evol.* 12:1-6.

Gray M. W. (1988). Organelle origins and ribosomal RNA. *Biochem. Cell Biol.* 66:325-348.

Gray M. W. (1992). The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* 141:233-357.

Gray M. W., R. Cedergren, Y. Abel, and D. Sankoff. (1989). On the evolutionary origin of plant mitochondria and its genome. *Proc. Natl. Acad. Sci. USA* 86:2267-2271.

Gray M. W. and W. F. Doolittle. (1982). Has the endosymbiont hypothesis been proven? *Microbiol. Rev.* 46:1-42.

Gupta R. S. (1995). Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. *Mol. Microbiol.* 15:1-11.

Gupta, R. S. (1996). Evolutionary relationships of chaperonin, p. 27-64. In R. J. Ellis (ed.), The Chaperonins. Academic Press, N.Y.

Gupta R. S. (1997). Protein phylogenies and signature sequences: evolutionary relationships within prokaryotes and between prokaryotes and eukaryotes. *Antonie Van Leeuwenhoek*, in press.

Gupta R. S., K. Aitken, M. Falah, and B. Singh. (1994). Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA* 91:2895-2899.

Gupta R. S., K. Bustard, M. Falah, and D. Singh. (1997). Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. *J. Bacteriol.* 179:345-357.

Gupta R. S. and G. B. Golding. (1993). Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* 37:573-582.

Gupta R. S. and G. B. Golding. (1996). The origin of the eukaryotic cell. *TIBS* 21:166-171.

Gupta R. S., G. B. Golding, and B. Singh. (1994). HSP70 phylogeny and the relationship between archaebacteria, eubacteria, and eukaryotes. *J. Mol. Evol.* 39:537-540.

Gupta R. S. and B. Singh. (1992). Cloning of the HSP70 gene from *Halobacterium marismortui*: relatedness of archaebacterial HSP70 to its eubacterial homologs and a model for the evolution of the HSP70 gene. *J. Bacteriol.* 174:4594-4605.

Gupta R. S. and B. Singh. (1994). Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. *Curr. Biol.* 4:1104-1114.

Hartl F. U. (1996). Molecular chaperones in cellular protein folding. *Nature* 381:571-580.

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

Hasegawa M., T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata. (1993). Early branchings in the evolution of eukaryotes: ancient divergence of entamoeba that lacks mitochondria revealed by protein sequence data. *J. Mol. Evol.* 36:380-388.

Hashimoto T. and M. Hasegawa. (1996). Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1alpha/Tu and 2/G. *Adv. Biophys.* 32:73-120.

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

Irwin D. M. (1994). Molecular evolution. Who are the parents of eukaryotes? *Curr. Biol.* 4:1115-1117.

Iwabe N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. (1989). Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* 86:9355-9359.

John P. (1987). *Paracoccus* as a free-living mitochondrion. *Ann. N. Y. Acad. Sci.* 503:140-150.

Joklik W. K., H. P. Willett, D. B. Amos, and C. M. Wilfert. (1992). *Zinsser Microbiology*, 20th ed. Appleton & Lange. Norwalk, Conneticut/San Mateo, California.

Jones W. J., D. P. Nagle,Jr., and W. B. Whitman. (1987). Methanogens and the diversity of archaebacteria. *Microbiol. Rev.* 51:135-177.

Kandler, O., and Konig, H. (1985). Cell envelope of archaebacteria, p. 413-457. *In* C. R. Woese and R.S. Wolfe (eds), The bacteria: a treatise on structure and function, vol. 8, Archaebacteria. Academic Press, London.

Kasting J. F. (1993). Earth's Early Atmosphere. *Science* 259:920-926.

Knoll A. H. (1992). The early evolution of eukaryotes: a geological perspective. *Science* 256:622-627.

Lake J. A. (1988). Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* 331:184-186.

Lake J. A. (1989). Origin of the eukaryotic nucleus: eukaryotes and eocytes are genotypically related. *Can. J. Microbiol.* 35:109-118.

Lake J. A. (1991). Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. *Trends Biochem. Sci.* 16:46-50.

Lane D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. USA* 82:6955-6959.

Langworthy T. A. (1985). Lipid of archaebacteria, p. 459-498. In C. R. Woese and R. S. Wolfe (ed.). The bacteria, vol. 8. Archaebacteria. Academic Press. Inc., New York.

Leipe D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. (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.

Li, W.H. and Graur, D., (1991). In: *Fundamentals of Molecular Evolution*, Sunderland, Massachusetts:Sinauer Associates.

Maniloff, J., (1992). Phylogeny of *Mycoplasmas*., p. 549-559. In: *Mycoplasmas:molecular biology and pathogenesis*, Manillof, J., McElhaney, R.N., Finch, L.R., and Basemann, J.B. American Society for Microbiology, Washington, D.C.

Margulis, L. (1970). In: *Origin of eukaryotic cells*,Anonymous New Haven:Yale University Press.

Martin W., H. Brinkmann, C. Savonna, and R. Cerff. (1993). Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc. Natl. Acad. Sci. USA* 90:8692-8696.

Meyer T. E., M. A. Cusanovich, and M. D. Kamen. (1986). Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proc. Natl. Acad. Sci. U S A* 83:217-220.

Moulder J. W. (1985). Comparative biology of intracellular parasitism. *Microbiol. Rev.* 49:298-337.

Murray, R. G. E. (1962). Fine structure and taxonomy of bacteria, p. 119-144. *In* G. C. Ainsworth and P. H. A. Sneath (ed.) Microbial classification. The Society for General Microbiology Symposium 12.

Murray, R. G. E. (1974). A place for bacteria in living world, p. 4-9. *In* R. E. Buchanan, N. E. Gibbons (ed.). Bergey's manual of determinative bacteriology, 18th edition. Williams & Wilkins, Baltimore.

Murray, R. G. E. (1992). The family *Deinococcaceae*, p. 3732-3744. *In* A. Balows, H. G. Truper, D. Martin, W. Harder, and K.-H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, N.Y.

Nei, M. (1987). In: *Molecular Evolutionary Genetics*. Nei, M. Columbia University Press, New York.

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

Olsen G. J., C. R. Woese, and R. Overbeek. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* 176:1-6.

Pace N. R. (1991). Origin of life--facing up to the physical setting. *Cell* 65:531-533.

Priest, F., and Austin, B., (1993). Modern bacterial taxonomy, 2nd ed. Chapman &Hall, London.

Puhler G., H. Leffers, F. Gropp, P. Palm, H. P. Klenk, F. Lottspeich, R. A. Garrett, and W. Zillig. (1989). Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci. USA* 86:4569-4573.

Rivera M. C. and J. A. Lake. (1992). Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* 257:74-76.

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

Sandman K., J. A. Krzycki, B. Dobrinski, R. Lurz, and J. N. Reeve. (1990). HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones. *Proc. Natl. Acad. Sci. U S A* 87:5788-5791.

Schlegel H. G., and H. W. Jannasch. (1992). Prokaryotes and their habitats, p. 75-125. *In* A. Balows, H. G. Truper, D. Martin, W. Harder, and K.-H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, New York.

Schnabel R., W. Zillig, and H. Schnabel. (1982). Component E of the DNA-dependent RNA polymerase of the archaebacterium *Thermoplasma acidophilum* is required for the transcription of native DNA. *Eur. J. Biochem.* 129:473-477.

Schwartz R. M., and M. O. Dayhoff. (1978). Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science* 199:395-403.

Searcy G. D. (1975). Histone-like protein in the prokaryote *Thermoplasma acidophilum*. *Biochim. Biophys. Acta.* 395:535-547.

Searcy G. D. (1987). Phylogenetic and Phenotypic relationships between the eukaryotic nucleocytoplasm and thermophilic archaebacteria. *Ann. N. Y. Acad. Sci.* 503:168-179.

Smith M. W., D. F. Feng, and R. F. Doolittle. (1992). Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem. Sci.* 17:489-493.

Sneath, P.H.A., Mair, N.S., Sharpe, M.E., and Holt, J.G. (1986). In: *Bergey's manual of systematic bacteriology*. The Williams and Wilkins Co., Baltimore

Sogin M. L. (1991). Early evolution and the origin of eukaryotes. *Curr. Opin. Genet. Dev.* 1:457-463.

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

Sogin M. L., H. G. Morrison, G. Hinkle, and J. D. Silberman. (1996). Ancestral relationships of the major eukaryotic lineages. *Microbiologia* 12:17-28.

Stanier R. Y. (1970). Some aspects of the biology of cells and the possible evolutionary significance, p. 1-38. *In* H. P. Charles and B. C. G. Knight (ed.) Organization and control in prokaryotic and eukaryotic cells. The Society for General Microbiology Symposium 20.

Steel M. A., P. J. Lockhart, and D. Penny. (1993). Confidence in evolutionary trees from biological sequence data. *Nature* 364:440-442.

Takahata N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957-966.

Taylor F. J. (1987). An overview of the status of evolutionary cell symbiosis theories. *Ann. N. Y. Acad. Sci.* 503:1-16.

Tiboni O., P. Cammarano, and A. M. Sanangelantoni. (1993). Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum Pyrococcus woesei: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. *J. Bacteriol.* 175:2961-2969.

Van de Peer Y., J.-M. Neefs, P. De Rijk, P. De Vos, and R. De Wachter. (1994). About the order of divergence of the major bacterial taxa during evolution. *Syst. Appl. Microbiol.* 17:32-38.

Van den Eynde H., Y. Van de Peer, J. Perry, and R. De Wachter. (1990). 5S rRNA sequences of representatives of the genera Chlorobium, Prosthecochloris, Thermomicrobium, Cytophaga, Flavobacterium, Flexibacter and Saprospira and a discussion of the evolution of eubacteria in general. *J. Gen. Microbiol.* 136:11-18.

Viale A. M. and A. K. Arakaki. (1994). The chaperone connection to the origin of the eukaryotic organelles. *FEBS* 341:146-151.

Viale A. M., A. K. Arakaki, F. C. Soncini, and R. G. Ferreyra. (1994). Evolutionary relationships among eubacterial groups as inferred from GroEL (chaperonin) sequence comparisons. *Int. J. Syst. Bacteriol.* 44:527-533.

Welch W. J. (1992). Mammalian stress response: cell physiology, structure/function of stress proteins, and implication for medicine and disease. *Physiol. Rev.* 72:1063-1081.

Welch W. J. (1993). How cells respond to stress. *Sci. Am.* 268:56-64.

Wittmann-Liebold B., A. Kopke, E. Ardent, W. Kromer, T. Hatakeyama, and H-G Wittmann. (1990). *In* p. 598-616. The Ribosome: Structure, Function and Evolution. H. We, American Society for Microbiology, Washington DC.

Woese C. R. (1981). Archaebacteria. *Sci. Am.* 98-122.

Woese C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51:221-271.

Woese, C.R. (1991). The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. *In* Evolution at the molecular level. Selander, R.K., Clark, A.G., and Whittmay, T.S. Mass.Sinauer Associates Inc., Publishers, Sunderland.

Woese, C.R. (1992). Prokaryote systematics: the evolution of a science. pp. 1-8. *In* The Prokaryotes, 2nd ed., Springer-Verglag, New York.

Woese C. R. (1994). Microbiology in transition. *Proc. Natl. Acad. Sci. USA* 91:1601-1603.

Woese C. R. (1996). Phylogenetic trees: Whither microbiology? *Curr. Biol.* 6:1060-1063.

Woese C. R., O. Kandler, and M. L. Wheelis. (1990). Towards a natural system of organisms: proposal for the domains *Archaea, Bacteria,* and *Eucarya. Proc. Natl. Acad. Sci. USA* 87:4576-4579.

Woese C. R., J. Maniloff, and L. B. Zablen. (1980). Phylogenetic analysis of the *mycoplasmas. Proc. Natl. Acad. Sci. USA* 77:494-498.

Yang D., Y. Oyaizu, H. Oyaizu, G. J. Olsen, and C. R. Woese. (1985). Mitochondrial origins. *Proc. Natl. Acad. Sci. USA* 82:4443-4447.

Zhu X., X. Zhao, W. F. Burkholder, A. Gragerov, C. M. Ogata, M. E. Gottesman, and W. A. Hendrickson. (1996). Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* 272:1606-1614.

Zillig W., P. Palm, W-D Reiter, F. Gropp, G. Puhler, H-P Klenk. (1988). Comparative evaluation of gene expression in archaebacteria. *Eur. J. Biochem.* 173:473-482.

Zillig W., R. Schnabel, and K. O. Stetter. (1985). Archaebacteria and the origin of the eukaryotic cytoplasm. *Curr. Top. Microbiol. Immunol.* 114:1-18.

Zillig W. (1991). Comparative biochemistry of Archaea and Bacteria. *Curr. Opin. Genet. Dev.* 1:544-551.

Zillig, W., Palm, P., Klenk, H.-P., Langer, D., Hudepohl, U., Hain, J., Lanzendorfer, M. and Holz, I. (1993), p. 367-391. *In* The biochemistry of *archae* (Archaebacteria), eds. Kales, M., Kushner, D.J. and Matheson, A. T. Elsevier, Amstrdam.

Zuckerkandl E. and L. Pauling. (1965). Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357-366.

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6″

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