STUDIES ON THE
PROKARYOTE - EUKARYOTE TRANSITION

by

Patrick J. Keeling

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia

August, 1996

ISBN 0-612-16040-8

Canada

Name [redacted]

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Biological Sciences / Biology / Molecular
SUBJECT TERM

0307
SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION
General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES
American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES
Pure Sciences
Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences
Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY
General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

The medium in which [the scientist] works does not lend itself to the delight of the listener's ear. When he designs his experiments or executes them with devoted attention to the details he may say to himself, "This is my composition: my pipette is my clarinet." And the orchestra may include instruments of the most subtle design. To others, however, his music is as silent as the music of the spheres. He may say to himself, "My story is an everlasting possession, not a prize composition which is heard and forgotten," but he fools only himself. The books of the great scientists are gathering dust on the shelves of learned libraries. And rightly so. The scientist addresses an infinitesimal audience of fellow composers. His message is not devoid of universality but its universality is disembodied and anonymous. While the artist's communication is linked forever with its original form, that of the scientist is modified, amplified, and fused with the ideas and results of others, and melts into the stream of knowledge and ideas which forms our culture. The scientist has in common with the artist only this: that he can find no better retreat from the world and also no stronger link with the world than his work.

Max Delbrück, Nobel Address

TABLE OF CONTENTS

Illustrations and Tables

Abstract

The transition between prokaryotic and eukaryotic cellular architecture has been examined with the aim of clarifying the nature of the ancestor of all extant eukaryotes. Specifically, the origin of DNA replication, endomembrane signalling and protein turnover processes were examined, along with structures such as the cytoskeleton and mitochondrion. One strategy used was to identify genes involved in these processes in the most ancient eukaryotic phyla, as the presence of these genes implies that the process or structure predates the divergence of all extant eukaryotes. Calmodulin, ubiquitin, E2 ubiquitin-conjugating enzyme and alpha-tubulin genes were isolated from a variety of taxa for this purpose, for the most part demonstrating that these genes predate extant eukaryotes. Another strategy was to identify an archaebacterial analogue of a eukaryotic process to determine the state of their common ancestor. An archaebacterial chromosomal replication origin was characterised to better define the DNA replication system in the ancestor of eukaryotes, but since no definite conclusions about this locus' activity could be made, such inferences about the ancestral state are not possible. Lastly, the origin of the mitochondrion was examined by identifying a gene, triosephosphate isomerase, which appears to be of mitochondrial origin, but whose product functions in the cytosol. The presence of this gene in deeply-branching amitochondrial protists suggests that these taxa may have had a mitochondrion which they secondarily lost. This would mean that the mitochondrion, contrary to the current conventional hypothesis, was also present in the ancestor of extant eukaryotes.

During the characterisation of alpha-tubulin genes, the phylum Diplomonadida was found to include several members which do not seem to use the universal genetic code. This result was followed up by providing definitive evidence that this alternate code is in current use in diplomonads, and the distribution of the code within this phylum was also examined.

Abbreviations

ACS	ARS-consensus sequence
Amp	ampicillin
ARS	autonomously replicating sequence
bp	base pair
cDNA	complementary DNA
CTAB	cetyltrimethylammonium bromide
EDTA	ethylene-diaminetetra-acetic acid
EF-1α	translation elongation factor-1, alpha subunit
HMG CoA	hydroxymethylglutaryl-coenzyme A
kb	kilobase pair
MTOC	microtubule organising centre
PAM	accepted point mutations
PCR	polymerase chain reaction
rRNA	ribosomal RNA
SDS	sodium dodecyl sulfate
TE	Tris (10 mM pH 8.0) EDTA (1 mM, pH 8.0)
Tet	tetracycline
tRNA	transfer RNA
UBC	ubiquitin-conjugating enzyme (gene)
UV	ultraviolet

Acknowledgments

Of all the pages of this thesis, none has required so much forethought or preparation as these acknowledgments. This is in part because so many people have contributed to my arrival at this point, and also because I know that this is the only section that anyone will ever read.

I must start by thanking Susan Koval and Bob Murray for giving me my first research job in London, and then inciting me to write to Ford for graduate work. Bob Murray said I would like the atmosphere here and he was right: Ford "runs" his lab in a peculiar way that lets us each explore different questions at our own risk. This requires a degree of independence that I doubt I would have had at this point if I worked elsewhere, and I think has done a great deal to shape the way I will continue to work by teaching me caution and skepticism. The members of the Abteilung Doolittle, both past and present, have also given me a lot both as a group and individuals. Steve, Leo, Analee, Cheryl, Arlin, Jim, Dave, Olof, Jeremy, Banoo, Sandie, Oisín, Margaret, Udeni, Mike, John, Claire and Naiomi have all been sources of day to day help and distraction, and I have learned something from each. I would like to mention in particular several people, beginning with Andrew Roger who had a strong influence on the direction of my thesis, has given me a great deal of advice, and through his encyclopedic knowledge of protists has allowed me to trick everyone into believing that I also know something about these organisms. I also owe a lot to Amanda Doherty-Kirby and Eve Teh who both worked with me on the calmodulin project as undergraduates. Their independence and skill rescued an interesting project which may otherwise have disappeared into the frost at the back of my freezer. There are also numerous people, here and elsewhere, who have helped me start to resolve a tougher lesson eluded to in the

words of Max Delbrück on the dedication page. These words may been seen as cynical if you wish, but the idea that all our contributions are anonymous is something I hadn't considered, and I think it helps to put some of the fallibility that goes with all our endeavors into perspective. In this vein I would like to thank Hans-Peter Klenk for the many pleasant and successful collaborations that we have undertaken.

Lastly, I want to thank Lisa who has been very patient with me for the last seven years.

Introduction

I. Early Taxonomy of Microorganisms

The animal-plant dichotomy that for centuries influenced philosophers' conception of life was fatally challenged by the introduction of the microscope by Anton van Leeuwenhoek in 1675. This technological step forever changed biology by introducing an entirely new world of living things which were neither animal nor plant. Early solutions to this problem considered various criteria to assign these single-celled organisms to Animalia or Plantae based on the presence or absence of photosynthesis and motility, but microorganisms continued to pose problems as they never really fit comfortably into either kingdom. Various new kingdoms were proposed by different systematists, the most influential being Owen's Protozoa, Hogg's Primigenum, and Haeckel's Protista. However, the confusion over the boundaries between kingdoms and the status of many taxa led to the persistence of an animal-plant dichotomy.

About this time another puzzle was developing which was to have even greater impact on taxonomy. Within Protista, Haeckel identified the phylum Moneres (later Monera) which contained cells without a recognisable nucleus: the bacteria and a collection of what turned out to be inaccurately identified organisms or things that were not really cells at all. Ferdinand Cohn recognised the important distinction between nucleate and anucleate cells and stressed the union of the bacteria with another anucleate cell type, the blue-green algae. This group Cohn called the Schizophyta ("fission plants") and placed it in Plantae, in part due to the photosynthetic nature of the blue-green algae, and also because both plants and bacteria had rigid cell walls. Haeckel recognised Cohn's united anucleate group and incorporated it into his system by making the Monera, a union of bacteria and blue-

green algae, but continued to classify Monera as a phylum of Protista (see Copeland, 1938; Whittaker, 1969; Cavalier-Smith, 1993, and references therein).

The importance of this new dichotomy between nucleate and anucleate was raised in the next century when Copeland proposed once again to eliminate the animal-plant dichotomy and replace it with a system consisting of Haeckel's Kingdoms Animalia, Plantae, and Protista, elevating Monera to regal status (Copeland, 1938). This became the basis for Whittaker's popular five-kingdom classification, which differed most notably by the designation of a kingdom Fungi distinct from Plantae (Whittaker, 1969).

Whittaker's five kingdom system provided an excellent framework, however, it was still rather vague in the transition between prokaryote and eukaryote (anucleate and nucleate), and in the position of protists as intermediates of this change. Moreover, the five kingdom system was soon challenged, as were earlier two-kingdom classifications, by a technological innovation.

II. The Problem of Monera and Phylogeny of Molecules

Despite repeated admissions that the greatest single division between taxa separated the prokaryotes from eukaryotes, the lack of definition within the prokaryotes was a persistent feature of these systems. Whittaker's summation in 1969 included only five phyla of Monera, merely half the number assigned to Protista, and a fraction of the number assigned to th other three kingdoms.

This paucity of higher order taxa within the bacteria was due to the lack of morphological characters with which to distinguish one bacterium from another. A few groups such as the cyanobacteria, the spirochetes, the myxobacteria, and the Gram-positive bacteria have determinative ultrastructural features; however, the vast number of bacteria known at the time had none that were apparent, and classifying these was a problem (Stanier and van Neil, 1962).

Cohn recognised the difficulty in finding a phylogenetic classification for bacteria, but, wanting a universal nomenclature, suggested in 1875 that bacteria be divided into "form genera". These divisions were originally based only on morphology, but were later expanded and multiplied by Migula in 1897 who also considered physiological characters (for references and varying interpretations on this period and later see Stanier and van Neil, 1941; van Neil, 1946; Woese, 1994). Form genera were supposedly intended to provide the bacteriologist with a means to identify bacteria, but were not intended to be phylogenetic divisions. Such divisions would require numerous assumptions about primitive versus derived states for which there was no evidence. Although this distinction between phylogenetic and determinative classification became muddied to many, the idea of a classification based on evolutionary relationship was fortunately still a goal of bacteriology for many others. In 1941 Stanier and van Neil pointed out that an imperfect natural system was better than any empirical one. However, van Neil also saw the problem clearly, and later proposed that until such a system existed, there should be a system of determinative keys for bacteriologists to identify and define bacteria separately from attempts to classify them on phylogenetic grounds (van Neil, 1946). It was understood that a natural system of bacterial classification was simply not obtainable with the tools of the day; he hoped for a technological innovation (which he supposed would come from microscopy) that would provide the necessary resolution.

In the end van Neil was correct, bacterial systematics was rejuvenated by a technological innovation, but he could not have foreseen the source of that innovation. In the mid-twentieth century, the role of informational macromolecules was rapidly unfolding, and led inevitably to the idea that genotype could be used directly to derive relationships among organisms by comparing the sequences of these molecules. This idea is generally credited to Zukerkandl and Pauling

(1965a,b), but in principle it was used before this. One of the first methods to indirectly measure genotype similarities was to compare the frequencies of conjugation and complementation between members of the enterobacteriaceae and Gram-positives (Luria and Burrous, 1957; Marmur *et al.*, 1962; Falkow, 1965). It was understood that the frequency of recombination observed between homologous loci was a function of their similarity at the nucleotide level, which in turn reflected relatedness. However, the various restriction-modification systems which were soon discovered and the extremely poor level of resolution limited the usefulness of this technique to closely related species (Falkow, 1965; Falkow and Formal, 1969). A more widely used and successful method that developed about the same time was the study of reassociation kinetics between nucleic acid chains from different bacteria: a technique also known to be affected by the degree of identity at the nucleotide level (McCarthy and Bolton, 1963). The sensitivity of hybridisation kinetics was not very promising either, at best only identifying groups of bacteria that we now recognise as very similar (Brenner *et al.*, 1969; Marmur *et al.*, 1962) Nevertheless, this pursuit was effective enough to identify the ribosomal RNA (rRNA) operons as a very highly conserved fragment of the genome and a good marker for inferring relationships between more distant bacterial groups (Doi and Igarashi, 1965), a notion that is still popular.

These approaches foundered not on account of the underlying concept of genotype comparison, but rather because of practical details. These problems were finally overcome, largely through the efforts of Carl Woese, by the technique of oligonucleotide cataloguing. This process involved digesting the small subunit rRNA, separating the fragments, and determining which organisms share short sequence patterns. The relationship between organisms could then be plotted as dendrograms to show phylogenetic groupings and evolutionary distance. Altogether these were a rough approximation of the sequence similarity between two

molecules, but they were far more accurate than hybridisation kinetics or recombination frequencies, and it was using this technology that Woese produced the biggest shake-up in bacterial taxonomy since Cohn created the taxon in 1875.

What Woese's group found was that a particular population of prokaryotes, methanogens, were only very distantly related to other bacteria (Fox *et al.*, 1977), and equally distant to the eukaryotes (Woese and Fox, 1977a). This unique position in the tree of life earned the methanogens, later with the extreme halophiles and thermoacidophlies, their own kingdom, the Archaebacteria.

The splitting of the prokaryotes into two distinct groups, the Archaebacteria and the newly named Eubacteria, created three Primary Kingdoms (or Urkingdoms, later called Domains). This split also revealed another problem that would be difficult to resolve with molecular phylogeny. It had previously been assumed that the eukaryotes arose from some group of prokaryotes (and there were lots of theories as to which prokaryote this was), but the presence of three kingdoms, all equally distant by small subunit rRNA phylogeny was not thought to be consistent with this idea as it would require that eukaryotic rRNA genes had evolved much more quickly since they diverged from the prokaryotes. To explain the small subunit tree, Woese proposed that all three groups arose nearly simultaneously from a common ancestor. This ancestor, he reasoned, was very primitive, lacking defined features of information flow and metabolism. He called the ancestor a Progenote to emphasise that the coupling of genotype to phenotype was not yet fully evolved at this time (Woese and Fox, 1977b).

The idea of the Progenote took a strong hold, and the word is still used today to denote the last common ancestor of all extant life despite the growing evidence that all three kingdoms share many of the central components of cellular life that were initially argued to be undeveloped in the Progenote. For instance all three Domains have DNA genomes, ribosomes, homologous DNA and RNA

polymerases, and many pathways of carbon metabolism that can be assumed to have been ancestral.

Even more detailed inferences about what the last common ancestor looked like depend on a more reliable rooting of the universal tree. It is impossible to root most molecular phylogenetic trees because to root a group one needs a sequence from a taxon outside that group, and no such sequence exists for a tree of all life. This theoretical barrier was overcome by two groups independently in 1989, by rooting the universal tree using ancient gene duplications. In theory, two products of a gene duplication event that took place before the divergence of the taxa in question can be used to reciprocally root one another. This strategywas applied to the universal tree by Iwabe and co-workers, who used translation elongation factors, ATPase subunits and tRNAs, and by Gogarten and his collaborators, who used ATPases to root the tree. Both analyses concluded that the root of the tree falls between the archaebacteria and the eubacteria, making the archaebacteria the sister group of the eukaryotes (Iwabe *et al.*, 1989; Gogarten *et al.*, 1989).

The ATPase data have since been somewhat obscured by the discovery of other ancient duplication products, but the elongation factor data have subsequently been confirmed (Baldauf *et al.*, 1996) and expanded to include a third family of proteins orthologous to bacterial translation initiation factor-2, which is also rooted at the same position (Keeling and Doolittle, 1995; Keeling *et al.*, 1996; Baldauf *et al.*, 1996). In addition, other duplicated genes, aminoacyl tRNA synthetases and the fused repeats of carbamoylphosphate synthetase , have also been applied to the question, and once again confirm the original rooting (Brown and Doolittle, 1995; Lawson *et al.*, 1996). A composite of the universal tree based on these data and numerous other molecular markers is shown in Figure I-1.

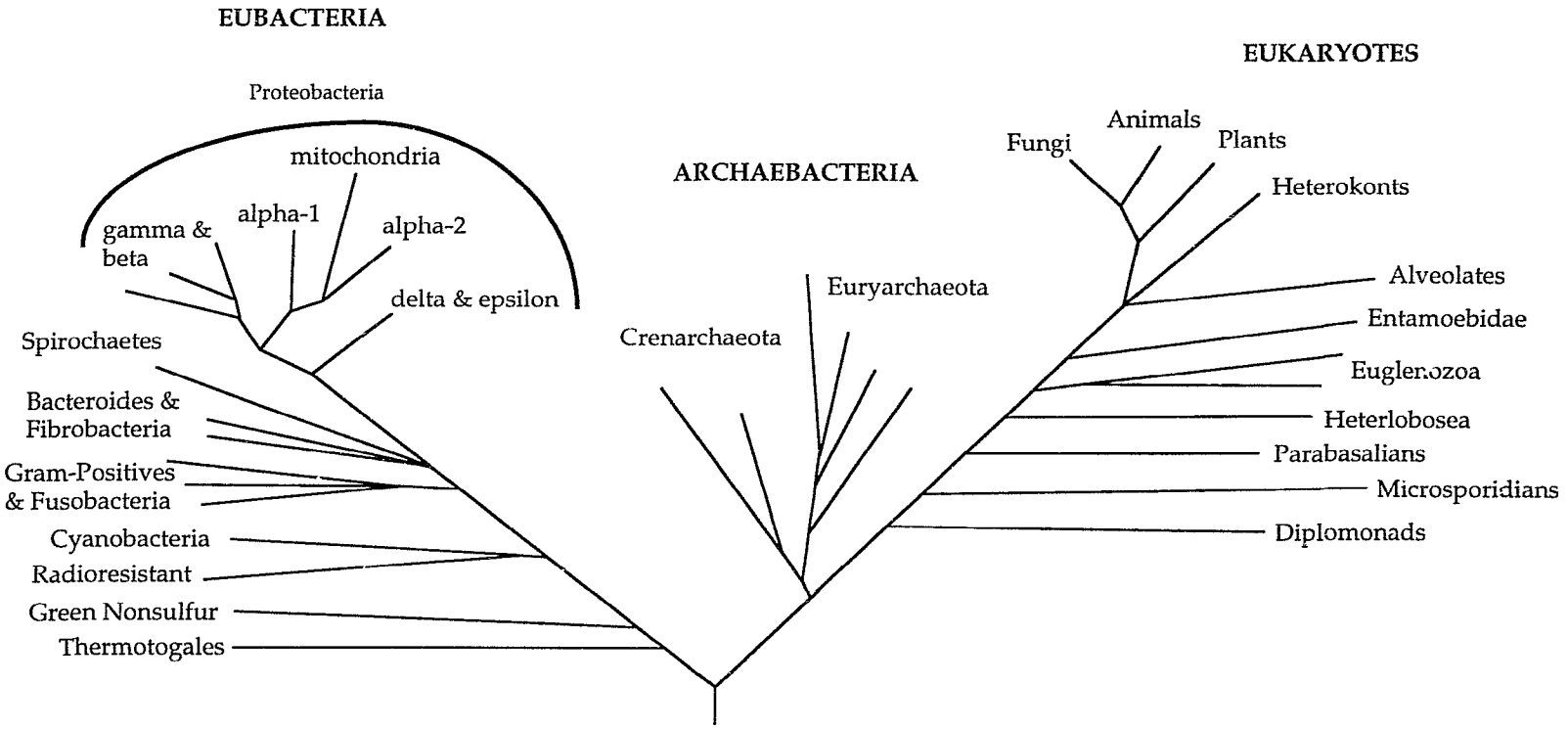


Figure I-1. Schematic universal phylogeny based on a composite of trees inferred from small subunit ribosomal RNA (Sogin, 1989; Cavalier-Smith, 1992; Olsen *et al.*, 1994; Ludwig and Schleifer, 1994; Van de Peer *et al.*, 1994), large subunit rRNA (Ludwig and Schleifer, 1994), GroEL (Viale *et al.*, 1994), RecA (Eisen, 1996), RNA polymerase (Klenk, 1995) and elongation factor-1 alpha (Hasegawa *et al.*, 1993; Hashimoto *et al.*, 1994; Delwiche *et al.*, 1995). The root is placed according to the conclusions of Iwabe *et al.* (1989) and other analyses as described in the text. The three domains, Eukaryotes, Archaebacteria and Eubacteria are distinguished as are major subdivisions within each group, with a special emphasis on taxa referred to in this study.

III. The Prokaryote - Eukaryote Transition.

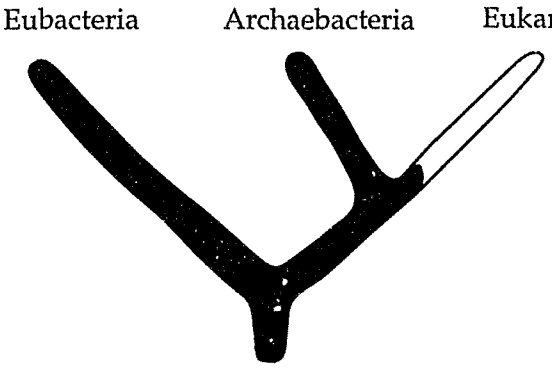
Part of the difficulty in understanding how eukaryotes could have evolved from prokaryotes is that they differ in so many ways that it is hard to decide on the order in which each of the many uniquely eukaryotic characteristics arose. There are a variety of theories which propose certain innovations to have precipitated the evolution of others. Stanier (1970) believed that endocytosis was the critical innovation, a view that was further developed by Cavalier-Smith (1991) who stressed the importance behind the advent of the cytoskeleton. Margulis (1981) has championed symbiogenesis, or the creation of novelty by symbiosis, and still other theories have postulated a complete fusion of two or more genomes leading to a chimeric eukaryote (Pühler *et al.*, 1989; Zillig, 1991; Sogin, 1991; Gupta and Golding, 1993). These are difficult to assess beyond simple plausibility without a better understanding of the nature of the transition: what the proto-eukaryote was like, and what it became.

The order of events underlying this transition is very difficult to resolve without two things: a model of the prokaryotic ancestor of eukaryotes and a reliable definition of the basal set of eukaryotic characteristics. Rooting the universal tree between archaebacteria and eubacteria helps fulfill the first of these goals by broadening the definition of universally ancestral features to include all those common to eubacteria and *either* archaebacteria *or* eukaryotes. Each of these characters is likely to have been present in the node that unites the archaebacteria and eukaryotes and this is the best current definition of the prokaryotic ancestor of eukaryotes. The rooting also allows the nature of this ancestor to be further narrowed by adding characteristics shared by eukaryotes and archaebacteria: these features are generally among those previously thought to be strictly "eukaryotic", but their presence in archaebacteria implies that they actually antedate the evolution of the nucleus. The second goal, a better definition of eukaryotes, is now possible

thanks to a growing body of molecular and morphological data on protists, which has identified several lineages that are likely the earliest among nucleated cells. By examining these lineages for characters generally assumed to be present in all eukaryotes, one can distinguish those that really were present in the ancestor of extant eukaryotes, and those which may have evolved later and are thus restricted to a subset of eukaryotes. These three means of deduction and some examples are detailed below (also see summary in Figure I-2)

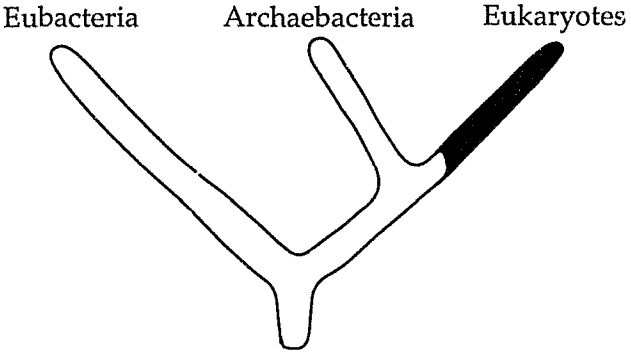
i. The ancestral state of living things. Woese's depiction of the last common ancestor as a progenote does not stand up to our current understanding of the distribution of cellular characteristics among living things. The abundance of molecular processes, metabolic pathways, and structural components which are characteristic ofall life testifies to the complexity of this ancestor, as each of these must have been present before all these life forms diverged. Different but homologous mechanisms for DNA synthesis, RNA synthesis, translation and central carbon metabolism have been found in all three Domains, arguing conclusively that the ancestor of these groups was a sophisticated cell likely similar in many ways to a modern eubacterium.

While these universally represented features can tell us a lot in general about the ancestral state of life, the most striking and detailed information can be found by comparing the archaebacteria and eubacteria, as these two groups span the root of the tree but share numerous homologous traits. Perhaps the best example of this is the likeness seen in gene structure. As in eubacteria, archaebacterial genes are organised into co-transcribed units, or operons, that are expressed as long, non-capped mRNAs with only short, bacteriail-like poly-A tails (Brown and Reeve, 1985; Brown *et al.*, 1989). It could be argued that these operons and operon clusters are the result of convergence upon the most efficient means of controlling



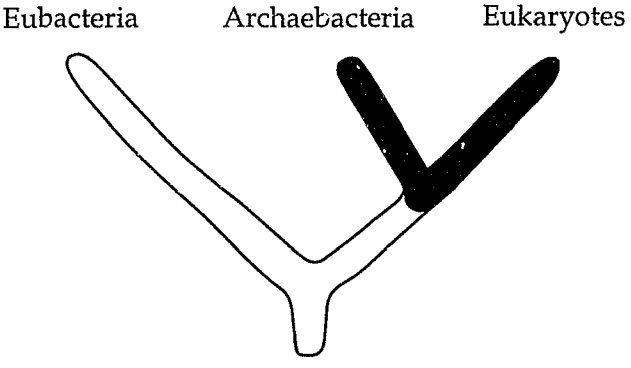
1. Characters present in Eubacteria and either Eukaryotes or Archaebacteria are considered ancestral to everything, and the ancestral state of the lineage that led to Eukaryotes. Any taxa without these traits is considered to have lost them.

Examples: operons, circular genomes, rotary motor flagella.



2. Characters restricted to the Eukaryotes are considered to have evolved after the divergence of the Archaeabcteria, but only if these characters are found to be present in the earliest Eukaryotes to have diverged.

Examples: linear genomes with true telomeres, cytoskeleton, nucleus, endomembrane system.



3. Characters shared between Eukaryotes and Archaebacteria are considered the ancestral state of the lineage leading to eukaryotes. These traits may have evolved earlier, and may have been lost or remain undetected in Eubacteria.

Examples: snRNP-based 16S rRNA processing, multi-subunit RNA polymerase and associated factors.

Figure I-2. Determining the ancestral state of characters by parsimony. Part 1 shows the ancestral state of a character found in archaebacteria and eubacteria. The same applies to characters found in eubacteria and eukaryotes (not shown). Part 2 is the inverse of 1, characters only present in eukaryotes. Part 3 shows the ancestral state of characters found in both archaebacteria and eukaryotes.

gene expression, yet several archaebacterial operons contain the same genes in the same order as their eubacterial homologues (reviewed in Ramírez *et al.*, 1993; Keeling *et al.*, 1994). In one case, the *Methanococcus vannielii* L1 operon, this similarity can be extended to the regulation of expression as it has been shown that the L1 protein in *Methanococcus* represses translation of its own mRNA by binding to a site resembling its binding site on the 23S rRNA in the same fashion as in *E. coli* (Hanner *et al.*, 1994).

In general, the structure of archaebacterial genomes also closely resembles that of eubacteria. They are similar in both size and structure, being small, compact, and circular. It is reasonable to suppose that this design is ancestral to all extant life, although we know too few of the details of genome construction to be certain that the trait could not have evolved twice independently. To describe the nature of the primitive genome more accurately, more comparative data are needed as well as a better understanding of archaebacterial genomes from a functional standpoint; for instance the mechanisms of replication initiation, termination, and chromosomal segregation need to be examined.

Locomotion is another important system of most cells and appears to be fundamentally homologous between eubacteria and archaebacteria, although this comparison is very complicated. Archaebacterial flagellins are more like members of the eubacterial type IV pilin-transport superfamily in both sequence and post-translational processing than they are like true eubacterial flagellins (Faguy *et al.*, 1994), but the rotary motor that drives them shares many physical characteristics with the eubacterial rotary motor, suggesting that the motor is a shared inherited character but in one group the structural part of the flagella has been substituted. The nature of the archaebacterial motor is a point of special interest as there has been a great deal of work on a sensory reception pathway in halophiles that governs the direction of the motor's rotation. In phototaxis, light-absorbing seven-helix

receptors analogous to eukaryotic opsins are coupled to a transducer homologous to those found in eubacterial chemotaxis pathways (reviewed in Spudich, 1993; Spudich, 1994). In eubacteria this transducer modulates the activity of a histidine kinase, CheA, which was recently characterised in *Halobacterium salinarium* where it also appears to play a general role in taxis (Rudolph and Oesterhelt, 1995). The presence of CheA implies that the switching mechanism in archaebacteria and eubacteria may also be of common origin, which would be interesting as the effects of switching are quite different. In many eubacteria the direction of motor rotation leads to either free swimming or random tumbling, while in archaebacteria switching merely changes the direction of swimming (Alam and Oesterhelt, 1984), a distinction that results in very different demands on the switching mechanism and motor.

A homologue of CheA has also been described in eukaryotes (Ota and Varshavsky, 1993), raising an important issue: it is unlikely that eukaryotes use CheA in exactly the same manner as prokaryotes since they do not have a homologous locomotory system. Nevertheless, CheA is part of the expanding number of molecular processes, protein families and individual proteins which are being found to be common to all life. Consider such recent findings as *inf*B orthologues in archaebacteria and eukaryotes (Keeling *et al.*, 1996), further evidence of bacteria polyadenylating mRNA (Cao and Sarkar, 1992; O'Hara *et al.*, 1995) and the growing number of claims for cross-domain homology based on secondary structures and weak sequence similarities. This latter kind of analysis has provided possible links between tubulin and FtsZ, and between actin, Hsp70 and FtsA (Bork *et al.*, 1992; Lutkenhaus, 1993; Sánchez *et al.*, 1994; Erickson, 1995; Erickson *et al.*, 1996), and is changing the way we think about the evolution of new processes. While cellular processes themselves differ between eukaryotes, eubacteria and archaebacteria, the components involved in carrying them out seldom

seem to have been purpose-built: Jacob's metaphor of evolution as tinkerer is as appropriate for molecules as for morphology (Jacob, 1977).

ii. "Eukaryotic" characteristics in archaebacteria. Even before Woese's definition of the archaebacteria, microbiologists had noticed certain molecular features that can be seen in retrospect to suggest a special relationship between archaebacteria and eukaryotes. Among the first of these were such things as the presence of N-linked glycoproteins, the lack of formyl-methionine, shared resistance or sensitivity to various antibiotics, and the presence of tRNA introns (White and Bayley, 1972; Mescher and Strominger, 1976; Zillig, 1987). These are the features that were first to arouse a great deal of excitement, and as more and more molecular mechanisms are studied in the archaebacteria, it is becoming apparent that some of these shared similarities run very deep. Detailed examples exist in the systems of rRNA processing (Duravic and Dennis, 1994, Potter *et al.*, 1995) and perhaps protein turnover (Wenzel and Baumeister, 1993; Wolf *et al.*, 1993), but the longest studied and most thoroughly understood is the similarity between eukaryotic and archaebacterial transcription.

This likeness was first recorded in the early eighties by Wolfram Zillig and colleagues, who had discovered archaebacterial DNA-dependent RNA polymerases to be of eukaryote-like complexity (Huet *et al.*, 1983). It has since become apparent that there are important mechanistic similarities underlying the multi-subunit polymerase that differentiate it from the eubacterial transcription apparatus. The eubacterial RNA polymerase has the ability to efficiently bind DNA; however, the holoenzyme includes a subunit, the sigma factor, which directs the polymerase specifically to promoters. Conversely, the eukaryotic enzyme cannot bind DNA, but rather recognises and binds a pre-initiation complex composed of transcription

factors which assemble independently at promoters. It is now known that archaebacterial promoters resemble their eukaryotic counterpart in both sequence motifs and relative position upstream of the start site (Reiter *et al.*, 1990; Hausner *et al.*, 1991). Moreover, the archaebacterial RNA polymerase cannot bind promoters efficiently (Hüdepohl *et al.*, 1990), but requires basal transcription factors to recognise the promoter and aid in polymerase binding (Thomm *et al.*, 1994; Wettach *et al.*, 1995). The recognition that at least two of these factors are homologous to eukaryotic transcription factors, TFIIB and TATA-binding protein (Ouzonis and Sander, 1992; Marsh *et al.*, 1994; Rowlands *et al.*, 1994; Hausner and Thomm, 1995), further underscores the high degree of detailed similarity between eukaryotic and archaebacterial transcription systems.

Another example is the process by which protein synthesis is initiated. Eukaryotic and eubacterial translation initiation involve analogous steps, but differ in how they are accomplished. In eubacteria three initiation factors direct most of the events while in eukaryotes the many common functions are carried out by a different set of factors which number into the dozens, only one of which is homologous (but not orthologous) to a eubacterial IF. The order of events and the underlying strategies also differ: in eukaryotes factors assemble the ribosome and initiator methionyl-tRNA around the 5' cap (reviewed in Merrick, 1992) while in eubacteria mRNA is bound to the free small subunit, guided by base pairing between the leader and the 16S rRNA, and then formyl-methionyl initiator tRNA is imported as part of a complex with IF-2 and GTP (reviewed in Kozack, 1983; McCarthy and Gualerzi, 1990).

At first glance, archaebacterial translation initiation appears to resemble that of eubacteria: there is no 5' cap on archaebacterial messages (Brown and Reeve, 1985), and sequences resembling Shine-Dalgarno sites are found both upstream (Brown *et al.*, 1989) or downstream (Dunn *et al.*, 1981) of the start codon of many

archaebacterial messages. However, although no archaebacterial translation initiation factor has been identified by its activity, there are now several examples of archaebacterial proteins homologous to eukaryotic translation initiation proteins, which may point to a deeper similarity between archaebacterial and eukaryotic initiation. The first of these to be discovered was the hypusine-containing protein homologous to eIF-5A (which is also distinguished by the presence of this modified amino acid) in *Sulfolobus* (Bartig *et al.*, 1992). This factor was thought to be involved in masking the charge of the unformylated initiator-methionine in eukaryotes (Merrick, 1992), an inference agrees nicely with the observation that archaebacteria also lack formyl-methionyl-tRN [illegible]White and Bayley, 1972). However, yeast cells depleted of eIF-5A continue [illegible]nthesise proteins at an only slightly decreased level, arguing that it is not a general translation factor at all (Kang and Hershey, 1994).

A less disputable example is found in an open reading frame upstream of the *Thermoplasma acidophilum* RNA polymerase operon which is closely related to eIF-1A (Keeling and Doolittle, 1995a), a factor which promotes dissociation of the ribosomal subunits (Thomas *et al.*, 1980). Another unidentified ORF in *Sulfolobus acidocaldarious* resembles two homologous subunits of eIF-2B (Keeling and Doolittle, 1995b), the guanine nucleotide exchange factor required by eIF-2 to recycle GTP (Bushman *et al.*, 1993). Unfortunately the putative presence of a functional analogue of eIF-2B (and by extension of eIF-2) in archaebacteria is significantly complicated by the fact that these eIF-2B subunits are part of a family of NDP-hexose phosphorylases and are closely related to a yeast protein, Psa1, that is thought to play a role in protein glycosylation (B. Benton and F. Cross, personal communication), another process shared between archaebacteria and eukaryotes (Mescher and Strominger, 1976).

A homologue of eubacterial initiation factor-2 has also been recognised in *Sulfolobus acidocaldarious* (Keeling *et al.*, 1996). However, this same study also identified a *eukaryotic* homologue of IF-2, which is even more akin to the *Sulfolobus* ORF. The role of this eukaryotic protein is not known (Sutrave *et al.*, 1994), but if it forms part of the translation initiation complex, then it has repeatedly escaped detection, which makes it difficult to decide just what the archaebacterial IF-2 homologue may be doing *in vivo*.

How can we reconcile this turmoil of conflicting information on translation initiation in archaebacteria? Given the lack of data, we probably shouldn't even try. The absence of a 5' cap and evidence for Shine Dalgarno-like base pairing rules out the use of a system entirely like eukaryotes, but the presence of non-formylated initiator methionine and putative eIFs hint that, like so many other molecular mechanisms in archaebacteria, translation initiation may in certain ways resemble that of eukaryotes.

iii. The Archezoa: basal eukaryotes. Even as similarities are being sought and found between prokaryotes and eukaryotes, there remain fantastic gaps between them in cellular architecture, in the way many proteins are used, and in the fashion by which many processes are carried out. These innovations define the eukaryotes, so it is important to be sure that such processes or cytological features truly are representative of all nucleated cells. In particular, there are numerous proteins, biochemical pathways, and cytological features that are widely thought of as being ancestral to all eukaryotes, but which have not been examined in protist lineages that diverged early in eukaryotic evolution. For example, although a few things may be inferred from what we know about archaebacteria, little is known about transcription, translation or protein turnover mechanisms in the earliest-diverging protists, and there is practically nothing known about their DNA

replication systems, cell cycle, recombination, chromatin structure or intracellular signaling. Without any data from these basal eukaryotic lineages, it is imprudent to assume that all eukaryotes are the same as animals, plants and fungi, just because they all have a nucleus.

This caution must be extended in particular to one group of eukaryotes, the Archezoa. This taxon, formally composed of the Metamonads, Microsporidia and Archamoebae (and formerly the Parabasalia), is a collection of predominantly parasitic amitochondrial protists which have consistently been shown through analyses of ultrastructural characteristics and molecular phylogeny to be the first lineages of eukaryotic cells to have diverged from the main eukaryotic line of descent (Cavalier-Smith, 1983; Vossbrinck et al., 1987; Cavalier-Smith, 1993; Leipe *et al.*, 1993). The combination of diverging first on molecular trees, and all being strictly anaerobic, amitochondrial cells led to the popular assumption that these taxa all diverged before the acquisition of the mitochondria, and have retained many other primitive features of the first nucleated cells (Cavalier-Smith, 1983).

The "primitive" nature of the archezoa, represented by this lack of mitochondria, in addition to the absence of peroxisomes, Golgi dictyostomes, spliceosomal introns, the possession of 70S ribosomes, and the fused 5.8S and 23S rRNA in Microsporidia, opens speculation as to what other "eukaryotic" characteristics they lack, and what "prokaryotic" characteristics will be found in archezoa. In fact, the more we look at archezoal genomes and their molecular biology, the more they appear to resemble other eukaryotes. A good example of this is seen in the chromosome structure of the diplomonad, *Giardia lamblia*. Archezoa, where known, have comparatively large genomes composed of multiple linear chromosomes, like other eukaryotes (Korman *et al.*, 1992). However, there is much more to the chromosome structure than simply being linear fragments; indeed at least three eubacterial taxa have independently developed linear chromosomes

(Crespi *et al.*, 1992; Chen *et al.*, 1993; Allardet-Servent *et al.*, 1993). In these bacteria, linearising a chromosome resulted in *ad hoc* adaptations to chromosome ends, but in *G. lamblia* the telomeres are very much the same as those of other eukaryotes (Le Blancq *et al.*, 1991a,b), suggesting that this is a homologous adaptation to a single event of chromosome linearisation which took place before the divergence of all known eukaryotes.

Similarly, the archezoa share a number of complex cytological characteristics with other eukaryotes. These include the cytoskeleton in which both actin and beta-tubulin have been identified, mitotic cell division, microtubule-based flagella and basal-bodies, and a complex endomembrane system (which includes the nucleus and endoplasmic reticulum).

Archezoa are also characteristically eukaryotic in their possession of a number of gene families that are either absent or represented by a single homologue in prokaryotes. For instance, although there are considerable similarities between archaebacteria and eukaryotic transcription systems, the eukaryotes have three homologous RNA polymerases (I, II, and III), whereas archaebacteria have only one. Evidence that this trait predates the divergence of all known eukaryotes comes from the demonstration that *G. lamblia* possess three RNA polymerases, two of which have been shown to fall securely into already defined classes (Lanzendörfer, 1992; Klenk *et al.*, 1995). Similarly, whereas eubacteria have three DNA polymerases of distinct families (A, B and C), eukaryotic DNA polymerases all descended from a single family B polymerase, but diverged into three subfamilies, alpha, delta, and epsilon (Braithwaite and Ito, 1993). Ongoing work by David Edgell is showing that this duplication also preceded the divergence of extant eukaryotes by identifying genes of all three eukaryotic subfamiles in either diplomonads or parabasalians (D. Edgell, personal communication).

IV. The Ancestral State of Eukaryotes.

The intent of the work presented in this thesis is to contribute to the process of ordering events in eukaryotic evolution by narrowing the relative date for the appearance of a number of proteins and systems that are supposedly common to all nucleated cells. The focus here will be events between the prokaryotic ancestor of eukaryotes, the divergence of extant eukaryotes and the subsequent divergence of eukaryotic cells. A summary of the work is outlined below, and the conclusions are shown in Figure I-3.

In Chapter 1 archaebacterial DNA replication initiation is examined to see if resemblances to that system in eubacteria or eukaryotes can be discerned. Characteristics shared with eubacterial replication initiation could be considered ancestral to all life, and where these differ in eukaryotes they must have changed after the divergence of archaebacteria. Chapters 2, 3, and 4 look at four highly conserved proteins which are generally thought to be present in all eukaryotes, calmodulin, ubiquitin, ubiquitin-conjugating enzyme E2, and alpha-tubulin. In each of these cases, homologues from early-diverging protist lineages were found, proving that orthologous genes existed before, or at the very least near the time that known eukaryotes diverged. Among the alpha-tubulins described in Chapter 4 is one from a diplomonad which contained two in-frame termination codons. This unexpected observation is followed up in Chapter 5 where evidence is presented that this organism, in addition to other diplomonad taxa, use a non-canonical genetic code. There is very little natural variation in the genetic code, and exceptions are generally thought to have arisen independently of one another. This appears to hold for the diplomonads as well, as the variant code is apparently restricted to one group of related organisms. Nevertheless, variants are always interesting, and this one yielded clues as to how the code actually evolves. Chapter 6 explores the possibility that genes for cytosolic proteins in eukaryotes may not really be part of

the nuclear heritage, but rather were inherited from the genome of the mitochondrial endosymbiont despite their complete lack of association with the modern organelle. One candidate protein, triosephosphate isomerase (TPI), was identified based on its unusual phylogeny, and examined further by sequencing TPI genes from several eubacteria including one of the closest relatives of the mitochondrial symbiont. The resulting phylogeny is in agreement with the notion that eukaryotic TPI is derived from the symbiont genome. This has interesting implications for the origin of the mitochondrion because one archezoan, *Giardia lamblia*, is known to possess a TPI gene similar to that of other eukaryotes. This suggests that *Giardia* may have also once possessed a mitochondrion, which in turn means that mitochondria are actually an ancestral feature of extant eukaryotes.

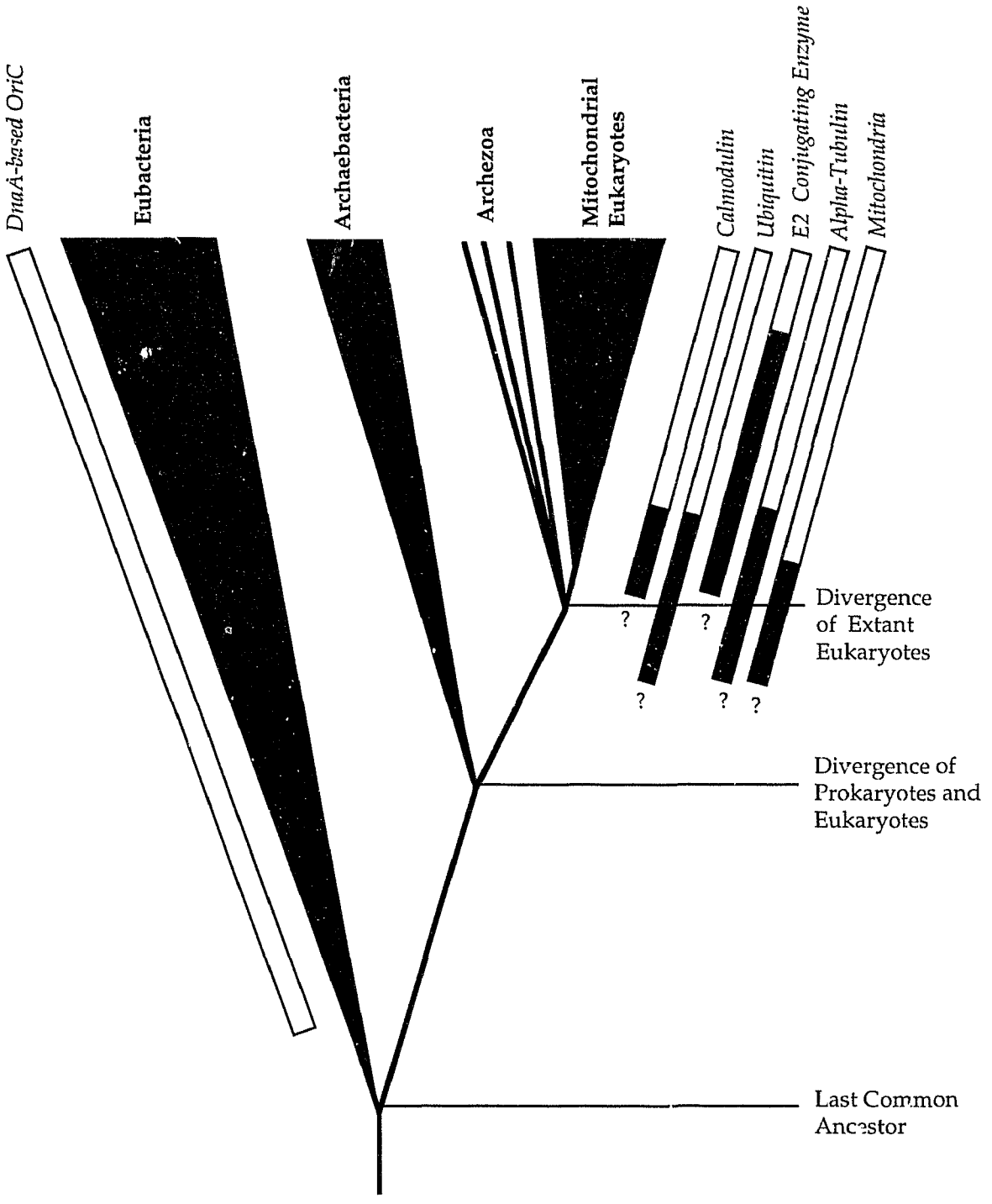


Figure I-3. Progress in developing a model of the ancestral state of extant eukaryotes. The tree shows three domains, and the diagonal bars represent the latest possible date of origin for six characteristics examined here. White portions of the bars represent the latest date of origin of that character before this work, and the black extensions represent the new time period resulting from this work. Question marks are meant to indicate that each of these characters may still have evolved much earlier than indicated.

Materials and Methods

I. Strains and Culture Conditions

Strains: *Escherichia coli* strain ED8767 [supE, supF, hsdS, lacY, recA56, (rk-mk-)] was used for maintaining and isolating *Hf. volcanii* and *Halobacterium* GRB cosmids. *E. coli* JM110 [rpsL, thr, leu, thi, lacY, galK, galT, ara, tonA, tsx, dam, dcm, supE44Δ(lac proAB) (F' tra Δ36, proAB, lacI^q ZΔM15), strR, Thr-, Leu-, Thi-] was used for shuttling and rescue of *Haloferax* plasmids. Strain XL-1BlueF' [recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, lac (F' proAB lacI^{q}ZΔM15 Tn10{tetR})] was used for library screening. General cloning and sequencing was performed using *E. coli* strain DH-5αF' [F' endA1, hsdR17, (rk-, mk+) supE44, thi-1, l-, recA1, gyrA96, relA1, Δ(lacZYA-argF)U169, F80dlacZΔM15]. All *E. coli* strains were taken from the collection of W.F. Doolittle. *Haloferax volcanii* strain WFD11 was used for all *Hf. volcanii* transformations and DNA isolations.

Some organisms were grown for the purpose of isolating genomic DNA and those that yielded results presented here are: *Rhizobium etli* CFN42, a gift from M.R. Esperanza at the Centro de Investigación Sobre Fijacion de Nitrogeno; *Agrobacterium radiobacter* K84 bv.2 a gift from C.R. Bell, Department of Biology, Acadia University; and *Hexamita* strains ATCC 50330 and ATCC 50380, both purchased from the American Type-Culture Collection.

DNA from other organisms was generously donated by individuals at Dalhousie and elsewhere. Those from which results were obtained that are presented here are as follows. *Francisella tularensis* LVS DNA was a gift from F. Nano, Department of Biochemistry & Microbiology, University of Victoria.

Helicobacter pylori 1107 DNA was a gift from A. Goodwin & P. Hoffman, Department of Microbiology & Immunology, Dalhousie University. *Chloroflexus aurentiacus* J-10-f1 DNA was a gift from J. Lopez & R.E. Blankenship, Department of Chemistry and Biochemistry, Arizona State University. *Rickettsia prowazekii* Madrid E DNA was a gift from H.H. Winkler, Department of Microbiology and Immunology, University of South Alabama College of Medicine. *Prochloron* sp. DNA was a gift from S. Douglas, Institute for Marine Biosciences, National Research Council, Halifax. *Hexamita inflata* AZ-4 & *Spironucleus muris* DNA were gifts from H. van Keulen, Department of Biology, Cleveland State University. *Trichomonas vaginalis* NIH-C1, *Tritrichomonas foetus* KV1, *Monocercomonas* sp. Ns-1PRR & *Trichomitus batrachorum* G11 DNA were gifts from M. Müller, Rockefeller University. *Naegleria fowleri* LEE DNA was a gift from N.R. Band, Department of Zoology, Michigan State University. *Encephalitozoon hellem* CDC:0291:V213 DNA was a gift from G.C. Clarke, School of Medical Parasitology, London. *Spraguea lophii* DNA was a gift from G. Hinkel, MBL, Wood's Hole. *Nosema lucustae* ATCC 30860, *Giardia lamblia* WB, & *Acrasis rosea* T-235 DNA were gifts from A.J. Roger, Department of Biochemistry, Dalhousie University. A guide to the taxonomy of these organisms is given in Appendix C.

Culture Conditions: Media formulations are given in Appendix A. All *E. coli* growth was at 37°C. Liquid cultures of *E. coli* were grown in 2YT or LB. Colonies were isolated on 2YT plates and plaques on NZY agar with NZY top agar. Cosmids were maintained under 50 mg/ml kanamycin selection, and plasmids under 100 mg/ml ampicillin selection. Blue/white selection was induced by spreading 35 ul of 10% X-Gal on plates prior to plating cells. It was found that the addition of IPTG, a beta-galactosidase inducer, was not necessary with DH-5αF'. *Rhizobium etli* CFN42 was grown at 26°C in 5 ml aliquots of liquid Rhizobium X broth (ATCC

111) with aeration and maintained on 5 cm plates of Rhizobium X agar with twice the concentration of soil extract. *Agrobacterium radiobacter* K84 bv.2 was grown at 26°C in 5 ml aliquots of LB broth with aeration.

Halophilic archaebacteria were grown aerobically at 42°C in rich broth and on rich agar. Plasmids in *Hf. volcanii* were maintained under selection for mevinolin resistance at 50 mM.

Hexamita strains ATCC 50380 and ATCC 50330 were grown in Keister's Modified TYI-S-33 broth (ATCC 1404) at 15°C in filled 15 ml airtight tubes. Cultures were maintained in dark, microaerophilic conditions for approximately 10 days before harvesting. Maximum cell density was very low, so a large number of cultures were combined before nucleic acids were extracted.

II. General Molecular Techniques.

Enzymes, Plasmids and Reagents: Restriction enzymes, T4 DNA ligase, Klenow polymerase, calf intestinal phosphatase and deoxynucleotides were purchased from NEB, Gibco-BRL, or Boehringer-Mannheim. Taq polymerase was from Appligene or Gibco-BRL, and PFU polymerase from Stratagene. pBluescript SK+ (Stratagene) was used for subcloning and sequencing. PCR products were cloned into pCRII (Invitrogen) and either sequenced in pCRII, or subcloned into pBluescript. pLS47-4 was used to screen for *Hf. volcanii* autonomously replicating sequences.

All chemicals were reagent grade and were purchased from Sigma, BDH, Aldrich, or BioRad. Mevinolin (Lovastatin) was given by A. Alberts at Merck-Sharp and Dohme. Radionucleotides ^{32}P-dATP and ^{35}S-dATP were purchased from DuPont-NEN.

DNA Purification: Plasmid DNA was purified from *E. coli* by several techniques depending on the fate of the plasmid. For general plasmid recovery alkaline lysis followed by phenol extraction was used (Sambrook *et al.*, 1989). Preparation of templates for sequencing required somewhat cleaner DNA, and it was found that the fastest and most consistently adequate plasmid DNA was obtained using any one of a variety of commercially produced ion exchange columns, initially Magic MiniPreps (Stratagene) and thereafter Nucleospin or Nucleobond (Macherey-Nagel) for higher yields. Plasmids from *Hf. volcanii* can be isolated by any of these techniques, but alkaline lysis was used.

Genomic DNA was extracted from various cell types by slightly different protocols. Archaebacterial cells were lysed by resuspending cells in distilled water, and by adding 0.1 volumes of 10% Sarkosyl and 0.1 volumes 0.5 M EDTA. *R. etli* and *A. radiobacter* lysis was induced by incubation for up to 1 hour at 37°C in SET (150 mM NaCl, 100 mM EDTA, 60 mM TrisHCl pH 8.3) containing 50 ug/mL RNAse A, 10 mg/mL Lysozyme, 1% SDS, and 1 mg/ml Proteinase K. *Hexamita* cells were lysed in TE (10 mM TrisHCl, 1 mM EDTA, pH8.0) by adding 0.1 volumes of 10% Sarkosyl and 0.1 volumes 0.5 M EDTA. DNA was purified from these lysates by repeated phenol and phenol-chloroform-isoamyl alcohol (50:49:1) extractions and precipitated in 2 volumes of ethanol followed by repeated washings in 80% ethanol at room temperature. Further purification by extracting with CTAB was found to be helpful in some cases. CTAB extractions were performed on semi-purified DNA from *A. radiobacter*, *F. tularensis*, and *Hexamita* strains according to the procedure of Ausubet *et al.* (1995).

DNA was purified from agarose gels by crushing gel slices in an equal volume of TE and two volumes of phenol. The mixture was frozen at -70°C for 10 minutes, thawed, and centrifuged whereupon the aqueous phase was recovered and DNA precipitated by adding 0.1 volume of 3 M sodium acetate and 2 volumes of

ice-cold absolute ethanol followed by centrifugation. Where small quantities of DNA were involved, 1 ug of yeast tRNA was added to aid precipitation.

Isolation from polyacrylamide was occasionally necessary for extremely small bands (less than 100 bp), and this was done by a variation of the procedure of Maxam & Gilbert (1977). Gel slices were submerged in extraction buffer (0.5 M ammonium acetate, 0.01 M magnesium acetate, 0.1% SDS, 0.1 mM EDTA) at 37°C for at least 2 hours. The extraction buffer was then removed and DNA precipitated in 0.8 volumes isopropanol and washed twice in 80% ethanol at room temperature. Precipitation was aided by the addition of 50 ug linear acrylamide (from a 0.025% stock).

Transformations: *E. coli* was routinely transformed using electroporation. Electrocompetent cells of DH-5αF' or JM110 were prepared in advance and stored at -70°C. 1-2 µl of DNA to be transformed was mixed with 40 µl of electrocompetent cells in a chilled 1 mm path electroporation cuvette (BioRad). This was pulsed with 1.8 kV at 25 uFD and 200 Ohms and surviving cells recovered from the cuvette by washing with 0.5 ml liquid media. Cells were allowed to regenerate for 20-60 minutes with aeration at 37°C and then plated on the appropriate selective media.

Hf. volcanii was transformed according to the method of Cline and Doolittle (1989), with a few variations. Cells were resuspended in 0.5-1.0 ml regenerating salts following transformation, and were not washed but wereplated directly, generally without the use of top agar.

Polymerase Chain Reaction: Primers used in all PCR reactions are given in Appendix B. All solutions and plasticwear used for PCR were UV irradiated to destroy contaminating DNA and reactions were prepared under aseptic conditions with aerosol-free pipette tips. Unless otherwise noted, PCR reactions consist of 100 ul containing 1 x PCR buffer (10 mM TrisHCl, 50 mM KCl, 1.5 mM $MgCl_2$,

0.1% TritonX100, and 0.2 mg/ml BSA at pH9 at 25°C), 10 mM (each) dNTPs, 2 U Taq polymerase (and sometimes an additional 0.5 U Pfu polymerase), and 1 uM of each primer to which 50 to 200 ng template DNA was added.

The precise reaction conditions varied according to the stringency desired and the expected length of the product, but several were used repeatedly. In general, reactions began with a 2 minute denaturation step at 92°C, and proceed through 35 cycles consisting of 1 minute at 92°C, 1 minute at the annealing temperature(s), and between 15 seconds and 2 minutes at 72°C (the optimal extension temperature for Taq polymerase) all followed by a 5 minute polishing step at 72°C. Unless otherwise stated, PCR using degenerate primers was carried out with an extension time of 1 minute and annealing for the first 10 cycles at 35°C followed by 25 cycles at 45°C. Annealing temperatures or extension times for specific cases that vary from this are given in the Results section.

Screening Transformants: The normally laborious process of screening *E. coli* transformants was significantly improved by the use of PCR (for which I owe a great debt to S. Gupta). A volume of PCR mixture containing standard PCR buffer (Appligene), 5 mM dNTPs, 0.5 U/100 µl Taq polymerase, and 0.5 µM of each the forward and reverse M13 sequencing primers is divided into 10 ul aliquots. Colonies are then picked and dipped briefly into an aliquot and subsequently stabbed onto a plate for later use. The reaction consists of 25 cycles starting at 94°C for 60s, 57°C for 60s, and 72°C for a period depending upon the expected size of the insert, all followed by a 5 minute polishing phase at 72°C. The individual reactions may then be electrophoresed (see below) to determine the exact insert size.

Electrophoresis: Agarose gels of between 0.6% and 2.0% agarose in 1 x TAE Buffer (0.04 M Tris-acetate, 0.001 M EDTA, pH8.0) were routinely used in standard submerged electrophoresis tanks also containing 1 x TAE.

Small vertical acrylamide gels of 8.5 by 9.5 cm were used to resolve smaller DNA fragments. Gels generally were cast with 1 mm spacers between an alum and glass plate. Gels contained 8% acrylamide/bis-acrylamide (38:2), and 1 x TBE (0.09 M Tris-borate, 0.001 M EDTA, pH8.0) in a volume of 20 ml, and were polymerised by the addition of 100 ul 0.1% APS and 50 ul TEMED.

Sequencing gels were made and electorphoresed using 60 cm BioRad Sequigen Sequencing Cells with plates spaced 0.25 mm. Gel contained 50% Urea, 8% acrylamide/bis-acrylamide (38:2), and 1 x TBE and were run in 1 x TBE. Sequencing gels were not fixed, but dried immediately on either 3 mm Watman chromatography paper or regular scrap paper.

Sequencing: DNA sequence was determined on double-stranded templates by dideoxy chain termination reactions initially using Sequenase 2.0 (USB), and later a T7 polymerase from Promega. All sequences except where noted are double stranded, and PCR products were sequenced over at least two individual clones to exclude Taq errors (which were observed in the *Hexamita* 50380 alpha-tubulin, *Hexamita inflata* EF-1α and the *Trichomonas vaginalis* clamodulin inverse PCR products). The general strategy for sequencing clones was to make restriction subclones until the entire sequence was obtained on at least one strand, and to fill in second-strand sequence with primers where necessary. Several clones required unique strategies which are described in the Results section.

Southern Blots: Southern transfers to Genescreen Plus (Dupont-NEN) were carried out according to the manufacturer's directions. Probes were labeled using nick translation (USB) or random priming (USB) and blots were hybridised in a rotisserie oven (Hybaid, a brand I do not recommend) with tubes containing 10% dextran sulfate, 1 M NaCl, 1% SDS. Blots were washed twice for 10 minutes in 2x SSC (0.15 M NaCl, 0.015 M sodium citrate) at room temperature, twice for 30 minutes in 2x SSC and 1% SDS at the hybridisation temperature, and finally twice

for 30 minutes in 0.1% SSC at room temperature. For high stringency hybridisations were carried out at 65°C while lower stringency as attained by lowering the temperature to 55°C or 50°C.

Library Screening: Lambda-Zap libraries of *Trichomonas vaginalis* cDNA, genomic DNA cut with *Eco*RI, and genomic DNA cut with *Alu*I were generously provided by P. Johnson at UCLA. Titres were calculated, libraries screened, and plasmid products obtained according to the manufacturer's directions (Promega).

III. Sequence Analysis

Contig Assembly and Sequence Analysis: DNA sequences were read and proofread in SeqEdit and assembled into contiguous blocks with SeqMan, both components of the DNAStar package. Open reading frames were identified and restriction patterns analysed using DNA Strider 1.2. Sequences were compared to databases using the BLAST programs (Altschul *et al.*, 1990; Gish and States, 1993) via the NCBI mail server.

Sequences from current GenBank, EMBL, PIR, and SwissProt databases were accessed using TurboGopher, Network Entrez, the NCBI Retrieve mail server, or with the retrieve function in GCG using the VAX200 at Dalhousie University Computing. DNA and amino acid sequences were aligned using the PileUp program from the GCG package. Pattern and repeat searches, dot-plots, and sliding-window GC content calculations were also performed using the appropriate programs from the GCG package (Devereux *et al.*, 1984).

Phylogenetic Tree Construction: Phylogenetic trees based on amino acid or nucleic acid alignments were inferred using distance, parsimony or maximum likelihood methods depending on the question or limitations of the dataset. Corrected distance measures were calculated with the PAM 250 substitution matrix

using the PROTDIST program from the PHYLIP 3.5 or 3.57c packages (Felsenstein, 1993). Distance trees were constructed using neighbor-joining with the NEIGHBOR program from PHYLIP. Statistical support for distance trees was assessed by conducting between 100 and 500 bootstrap replicates generated by SEQBOOT, and trees were drawn using DRAWGRAM or DRAWTREE, also from the PHYLIP package. Unweighted parsimony trees were found by conducting 50 or 100 independent heuristic searches for the shortest tree with tree-bisection and reconnection using PAUP version 3.1.1 (Swofford, 1993). Bootstrap support was calculated using the same search strategy for 100 or 500 random replicates, each with a single sequence addition. Maximum likelihood trees were exhaustively searched using the PROTML program from the MOLPHY 2.2 package (Adachi & Hasagawa, 1992).

Additional statistical tests were performed to see if the differences observed between the tree topologies based on TPI and other markers are significant or the result of poor resolution in the TPI tree. Templeton tests based on parsimony were performed by calculating the number of steps at each position for alternative trees using PAUP. These values were used to calculate the standard deviation from the best tree and the percent confidence that each alternative topology is significantly worse than the shortest tree according to the equation described by Adachi & Hasegawa (1992). Templeton tests built into the PROTPARS program in the PHYLIP package and the PROTML program in the MOLPHY package were also used. These calculate the standard deviation for each alternative to the best topology and PROTML gives an estimate of the bootstrap percent for nodes that differ in the comparison.

Chapter I:

Searching for Autonomously Replicating Sequences in the Archaebacterium, *Haloferax volcanii*.

Introduction

The mechanisms by which eukaryotes and eubacteria replicate their genomes are sufficiently different to suggest that they may not be directly related. Both systems use the same general strategy: initiation proteins bind the chromosome and direct a local melting of DNA that allows the priming of bi-directional DNA synthesis. However, most of the binding sites and proteins involved in initiation are not detectably homologous between eubacteria and eukaryotes. These differences may be seen as a cause or the result of a general dissimilarity in the genome structure and replication strategy between the two. Eukaryotic genomes are generally much larger than prokaryotic genomes and are packaged into multiple, linear chromosomes, each requiring many replication initiation sites. The sequences required for replication initiation are well known in *Saccharomyces* (Marahrens and Stillman, 1992; Newlon and Theis, 1993) and other budding yeast (Matsuoka *et al.* 1993; Cregg *et al.* 1985; Herreros *et al.*, 1992), but even related fungi do not use the same core sequences as replication origins (Maundrell *et al.*, 1985; Maundrell *et al.*, 1988; Johnson and Barker, 1987; Sakai *et al.*, 1993). The growing pool of data from other eukaryotes argues that it may not be possible to clearly define a eukaryotic replication origin in terms of sequence motifs alone, but that a variety of conditions must be recognised together, many of which may be specific to a particular group of organisms (Brewer, 1994; Burhans and Huberman, 1994).

Eubacterial replication control is more uniform. Genomes are typically smaller than those of eukaryotes, and almost all are comprised of a single, circular chromosome which initiates replication at a single origin (*ori*C). The molecular

details of initiation have been worked out meticulously in both *Escherichia coli* (Bramhill and Kornberg, 1988; Funnell *et al.*, 1987) and *Bacillus subtilis* (Moriya *et al.*, 1988; Moriya *et al.*, 1994) and the basic architecture of the origin appears to be similar in a wide variety of phylogenetically diverse eubacteria (Ogasawara and Yoshikawa, 1992). In all cases where it is known, the DnaA protein is the principal initiator of replication, and with few exceptions origins are associated with a large suite of genes, the order of which is highly stable. Within some of the intergenic spaces of this cluster are arrays of DnaA binding sites, or dnaA-boxes, and while the distribution of dnaA-boxes in intergenic spacers varies between taxa, as does the order, spacing and number of repeats found within each dnaA-box region, the overall form is conserved (Ogasawara and Yoshikawa, 1992).

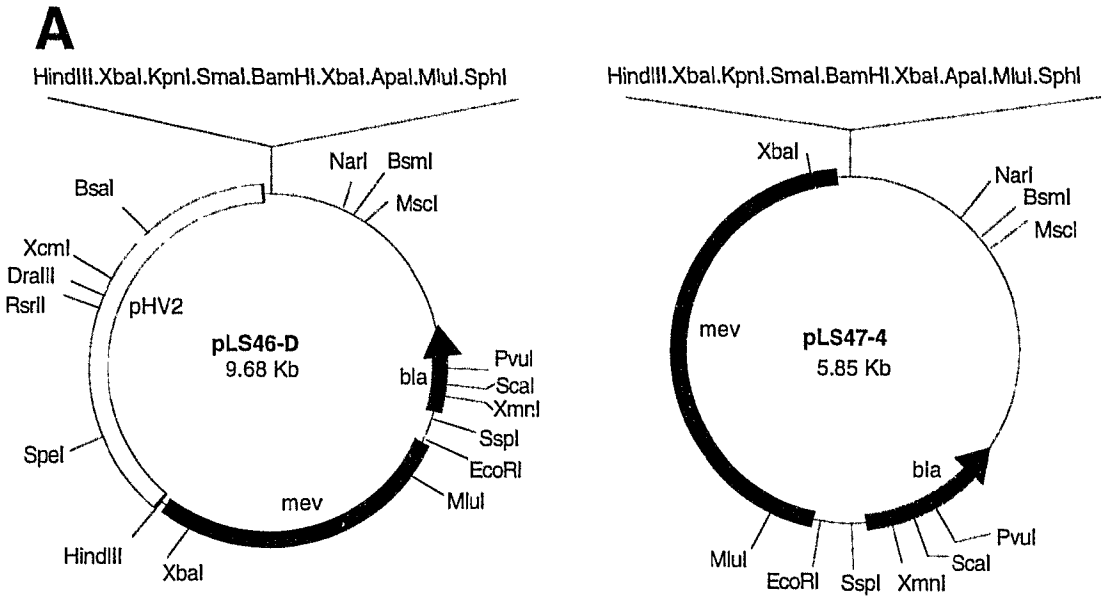
The relationship between the eukaryotic and eubacterial mechanisms of replication initiation and the nature of the system ancestral to both groups may be revealed by an analysis of archaebacterial replication. Practically nothing is known about how archaebacteria replicate their chromosomes, although genomic structural data might favor a eubacterial-like replication system in archaebacteria: physical maps for diverse archaebacteria all have a single circular chromosome, similar in size range to those found in eubacteria (reviewed in Keeling *et al.*, 1994). Moreover, two genes that almost always map close to the *ori*C in eubacteria, *gyr*B and *gyr*A, have been identified in the halophilic archaebacteria (Holmes and Dyall-Smith, 1991). If archaebacterial and eubacterial origins are indeed homologous, then one might expect the halophile replication origin to reside in this same genomic region. This was tested by screening fragments of the *Hf. volcanii* chromosome for the ability to confer transformational competence upon a plasmid that lacks a replicon. One such fragment was identified, sequenced and mapped to a position on the chromosome just 13 kb from the *gyr*BA cistron.

Results

A high-frequency transformation locus from *Hf. volcanii*. A genetic approach was taken to identify an archaebacterial replication origin by selecting for loci that impart a greater transformation frequency on a vector that otherwise lacks a replicon. This selection was performed using a new generation of selectable shuttle vectors developed by Leo Schalkwyk. These plasmids carry the mevinolin resistance determinant (hydroxymethylglutatyl CoA reductase or HMG CoA reductase) from *Haloarcula marismortui*, and are thus not subject to homologous recombination with the *Haloferax volcanii* chromosome (Figure 1-1A). The identity of this resistance determinant was confirmed by myself to be HMG CoA reductase by sequencing two small *Taq*I fragments, which show an average similarity to the *Hf. volcanii* HMG CoA reductase gene of 85% over a total of 62 amino acids (Figure 1-1B).

The pHV2 replicon of the shuttle vector pLS46-E was deleted, resulting in pLS47-4 (see Figure 1-1A), which should not be capable of replicating in halophiles. Indeed, transformation of WFD11 with pLS47-4 yields fewer than 0.008 mevinolin-resistant colonies per microgram of plasmid (no transformants were observed), suggesting that it cannot replicate or recombine with the chromosomal HMG CoA reductase gene of *Hf. volcanii*.

To screen for autonomously replicating sequences, total genomic WFD11 DNA was digested to completion using *Sau*3A and *Nla*III and shotgun cloned into the *Bam*HI or *Sph*I sites respectively of calf-intestine alkaline phosphatase (CIP)-treated pLS47-4. Transformation of WFD11 with these libraries produced 5-10 mevinolin resistant colonies per microgram of original genomic DNA (19 and 10 colonies for *Nla*III and *Sau*3A respectively), while pLS47-4 alone was never observed to produce any transformants. Many of these transformed colonies were unable to grow in liquid medium or reverted to mevinolin sensitivity and so were



B

Ha.marismortui	ATARVLKSGMTRA----------DVLADAAESTT
	ATARVLKSGMTRA L +AAE TT
Hf. volcanii	ATARVLKSGMTRAPVFRVADVAEAEALVSWTRDNFAALKEAAEETT
Ha.marismortui	SHGELQDVTPYVVGDSVFLRFSYDTKDAMGMNMATIATEAACDVV
	+HGEL DVTPYVVG+SV+LRF YDTKDAMGMNMATIATEA C VV
Hf. volcanii	NHGELLDVTPYVVGNSVYLRFRYDTKDAMGMNMATIATEAVCGVV

Figure 1-1. pLS46-E and pLS47-4. (A) Maps of pLS46-E and pLS47-4. *Ha. marismortui* mevinolin-resistance determinant (mev), halophile replicon (pHV2) and *E. coli* replicon and ampicillin resistance marker (bla). (B) Inferred amino acid sequence of two fragments of the *Ha. marismortui* mevinolin-resistance determinant aligned with positions 119- 209 of the HMG CoA reductase gene of *Hf. volcanii*. Dashes (-) indicate missing data in the space between the two fragments. Identity and similarity are indicated on the line between the two sequences.

discarded. In addition, plasmid DNA from transformants was never visible on ethidium bromide-stained agarose gels, suggesting that the copy number was extremely low. Total DNA from individual colonies was used to transform *E. coli* so that plasmid could be prepared in larger quantities. Most isolates failed repeatedly to transform *E. coli*, and were also discarded as having arisen either spontaneously, or by recombination. In other cases a small number of *E. coli* transformants was isolated, but plasmids born by them appeared to be highly unstable, as restriction analysis showed that they had undergone significant rearrangements. One plasmid that seemed better able to stably transform *E. coli* was chosen for further analysis. It was found to contain an insert consisting of two *Sau*3A fragments of 4.1 kb and 2.5 kb.

Convenient restriction sites for *Sau*3A and *Apa*I allowed overlapping subclones of the original 6.6 kb clone to be made in pLS47-4, and each was tested for its ability to transform WFD11 to mevinolin resistance. The 1.9 kb *Apa*I/*Sau*3A and 4.1 kb *Sau*3A fragments proved sufficient for transformation while the 4.7 kb *Sau*3A/*Apa*I fragment failed repeatedly to transform strain WFD11 (Figure 1-2). Under the same conditions the shuttle vector pLS46-E gave a transformation frequency of 3.8×10^4 colonies per microgram of plasmid.

Plasmids carrying the 1.9 kb, 4.1 kb and 6.6 kb fragments were cycled repeatedly through WFD11 and *E. coli* a total of three times. Each time plasmid DNA was isolated from *E. coli* and checked by restriction digestion to confirm its identity. The transformation frequency of *E. coli* was about 20 colonies per microgram of total DNA from *Haloferax* transformants, while the transformation frequency of the 6.6 kb clone and pLS47-4 purified from *E. coli* were 1×10^4 and 2×10^4 colonies to per microgram respectively, indicating that the copy number of the plasmids in *Haloferax* is indeed quite low.

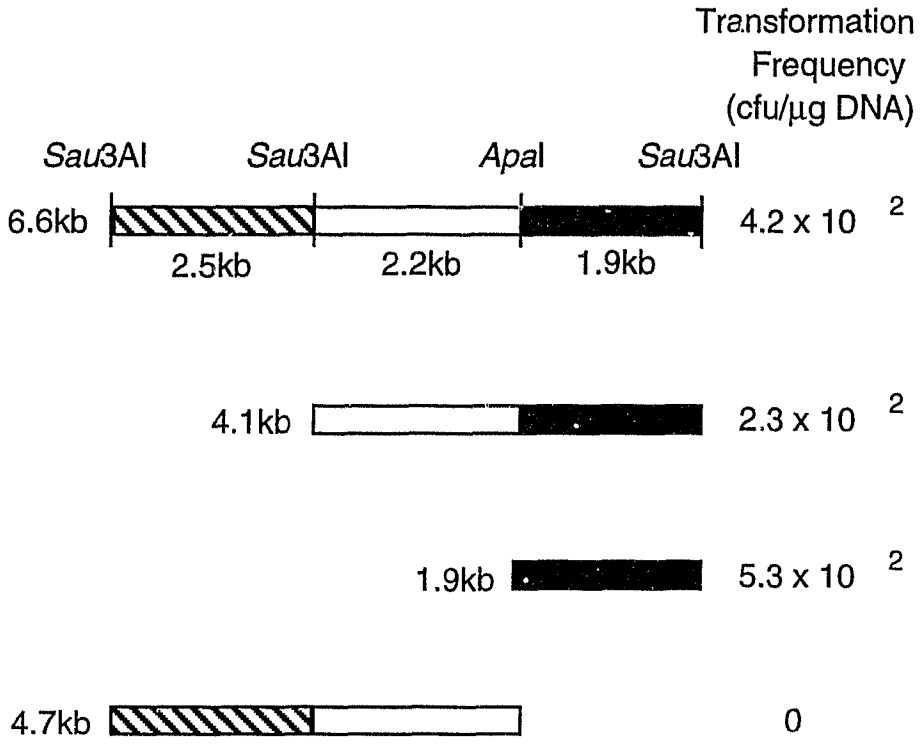


Figure 1-2. Activity of high-frequency transformation plasmids. Restriction fragments were deleted from 6.6 kb clone and used to transform WFD11 to mevinolin resistance in pLS47-4. The transformation frequency is shown to the right of each clone. From this it was concluded that the 1.9 kb *ApaI*/*Sau*3A fragment (shaded black) is all that is necessary for transformation. The corresponding frequencies of transformation for pLS46-E and pLS47-4 are 3.8 x 10^4 and less than 10^{-2} colonies per microgram of purified plasmid respectively.

Mapping and determination of genome order. The 6.6 kb clone was used as a probe against a cosmid library to identify its position in the genome (Charlebois *et al.*, 1991). Using this hybridisation data and a restriction map of the clone, two overlapping cosmids were identified, 547 and 5G7, to which the 6.6 kb clone was hybridised to show its exact position on the cosmid map, a locus henceforth named *hft* for high-frequency transformation. The cosmid blots show that the 2.5 kb *Sau*3A fragment is not contiguous with the 4.1 kb fragment and that only the latter maps to the overlap of cosmid 547 and 5G7 (Figure 1-3A). This, together with the failure of the 4.7 kb *Sau*3A/*Apa*I fragment to transform, led to the exclusion of the 2.5 kb fragment from all subsequent analysis.

The position of this locus on the map is of special interest as it is near the previously mapped *gyr*BA cistron. In eubacterial genomes the *gyr*B or *gyr*BA cistrons are near the origin and always transcribed away from it. To better understand the significance of *Hf. volcanii's* genomic organisation, *gyr*BA was remapped to determine its transcriptional orientation in the genome. Previously *gyr*BA had been mapped to the overlap between cosmids 547 and 516 (Charlebois *et al.*, 1991), but hybridisation of *Hf. volcanii* A2 *gyr*BA and *gyr*B individually to these cosmids showed that this position was incorrect. *Gyr*BA is actually more than 8 kb closer to *hft*, placing the two loci less than 13 kb from one another (Figure 1-3B) and in the same order, *hft*-*gyr*B-*gyr*A, in which eubacterial gyrase genes are found with respect to chromosomal replication origins (Figure 1-3CD; Figure 1-4).

Sequence of the *hft* locus. Initial attempts to sequence the 1.9 kb *Apa*I/*Sau*3A fragment by either double stranded or single stranded sequencing using various polymerases and protocols failed for reasons that are uncertain, but it became clear that the intact fragment could not be used as a template. *Taq*I and *Hin*PI libraries of the 1.9 kb clone were therefore constructed and random clones

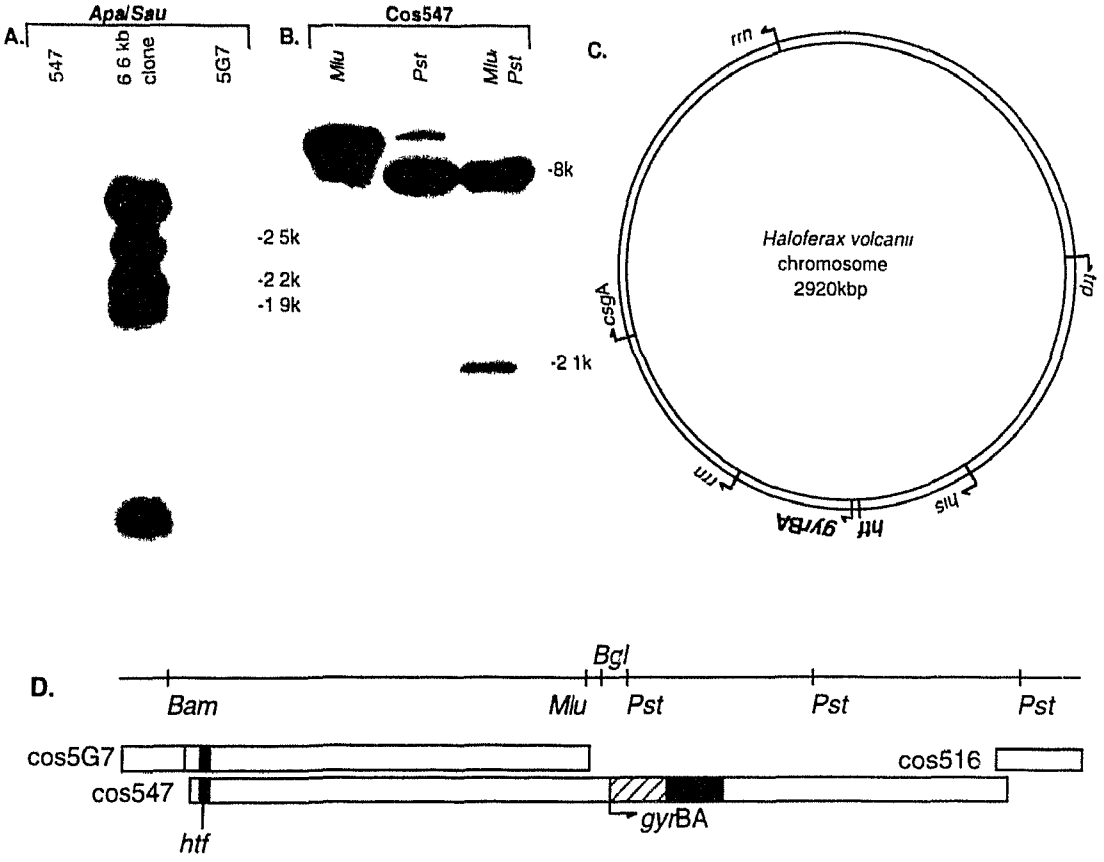


Figure 1-3. Mapping *hft* and genes for gyrase subunits. (A) Mapping *hft* on cosmids 547 and 5G7. Cosmids 547 and 5G7 and the 6.6 kb clone were digested with *Apa*I and *Sau*3A, and probed with the 6.6 kb fragment. The 2.5 kb *Sau*3A fragment does not appear in either cosmid, so is likely not contiguous with the rest of the clone. In lane 3 the 2.1 kb *Apa*I/*Sau*3A fragment is truncated at the *Mlu*I site corresponding to the end of cosmid 547. The inferred location at the end of cosmid 547 is confirmed by similar hybridisations using cosmids cut with *Bgl*II, *Pst*I, *Mlu*I, and *Bam*HI (data not shown). (B) Mapping the position and direction of *gyr*BA. Probing Southern blots of cosmid 547 with *gyr*BA or *gyr*B alone (the latter is shown) gives an identical pattern. The orientation must be B-A or the 13 kb *Pst*I fragment would not hybridise to both probes. This is confirmed by hybridisations with cosmids cut with *Bgl*II, *Pst*I, and *Bgl*II/*Pst*I digests (data not shown). Cosmid 516 did not hybridise to either probe (data not shown). (C) Chromosome of *Hf. volcanii* showing the direction of the transcription complex where known, the newly mapped *hft* and *gyr*BA are shown in bold. (D) Detailed map of the *hft* region showing some of the restriction sites relevant to mapping, and the positions of *gyr*BA and *hft* on cosmids 5G7 and 547. For scale, the distance from the end of *hft* to the start of *gyr*B is slightly less than 13 kb.

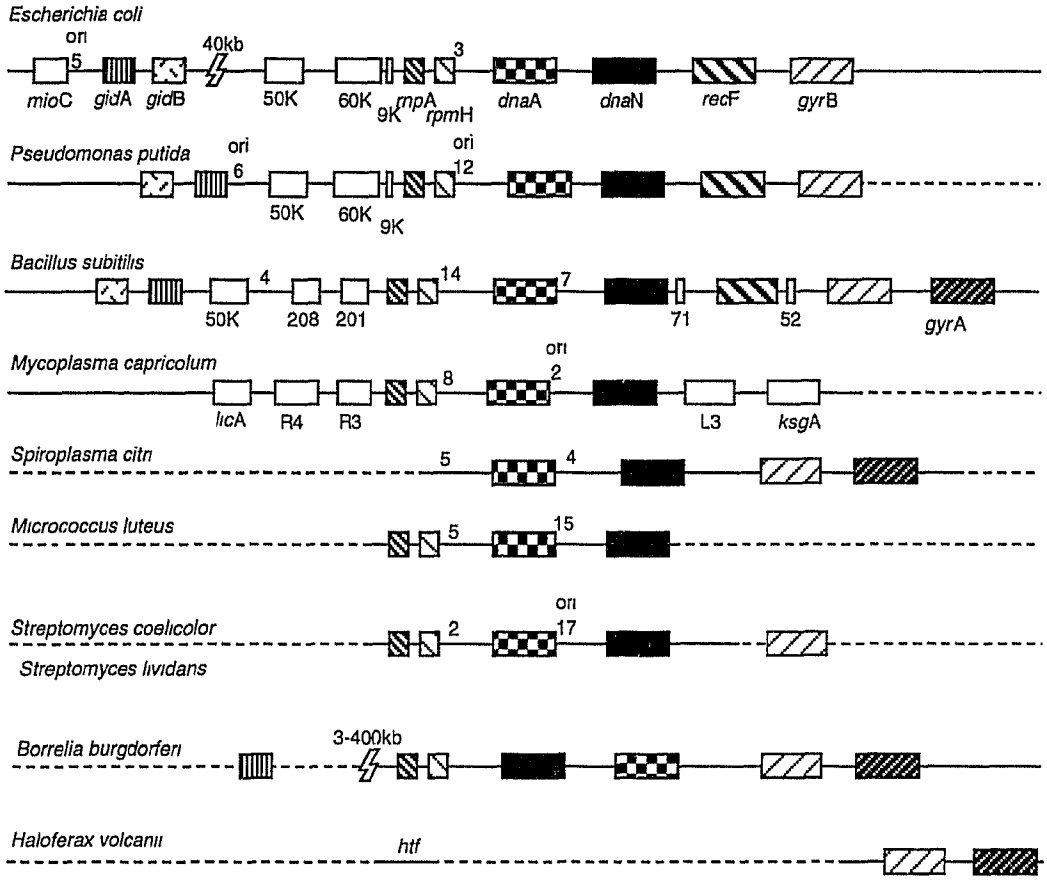


Figure 1-4: The *gyr*B region of eubacteria and *Hf. volcanii*. Numbers in intergenic regions indicate the presence of that many dnaA-boxes, allowing for a two nucleotide difference from the *E. coli* consensus. Unsequenced regions are dotted lines, and regions either known genetically to be autonomously replicating sequences, or which have been physically mapped as initiation sites are labeled "ori". The genes and spacers are not to scale. Data from other taxa have been omitted where they are exactly the same as a closely related taxon that is shown, or as in the case of *Caulobacter crescentus* and *Coxiella burnetii*, are anomalous cases that fall within a group showing this or a similar order. (Suhan *et al.* 1994; Marczynski and Shapiro 1992; Burland *et al.* 1993; Bramhill and Kornberg 1988; Yee and Smith 1990; Ogasawara *et al.* 1985; Moriya *et al.* 1992; Bailey and Bott 1994; Fujita *et al.* 1992; Miyata *et al.* 1993; Fujita *et al.* 1990; Zakrezewska-Czerwinska and Schrempf 1992; Calcutt and Schmidt 1992; Musialowski *et al.* 1994; Old *et al.* 1993; Holmes and Dyall-Smith 1991; Ye *et al.* GenBank accession Z1910)

sequenced. Initially, libraries were made from the entire fragment, then under-represented regions were targeted by constructing *Taq*I partial libraries from specific fragments known from the restriction map to correspond to the regions of inadequate representation. In this way, the entire 1.9 kb was sequenced with over four-fold redundancy (Figure 1-5).

Nucleotide and conceptual translations were compared to existing databases using BLASTN and BLASTX programs, and all uninterrupted reading frames were individually assessed by BLASTP searches (Gish and States, 1993; Altschul *et al.*, 1990). In no case was a significant match found, and no reading frames could be fit to alignments of *dna*A, *dna*N, *rpm*H, or *rnp*A, genes associated with eubacterial replication origins.

There are numerous repeated sequences throughout the clone, but the high GC-content (67% overall, a general feature of the *Hf. volcanii* genome) makes it difficult to assess the importance of these. However, this high GC-content does make the appearance of an AT-rich region containing several long runs of AT-pairs from position 670 to 800 more interesting. AT-rich regions are characteristic of many origin sequences (Yoshikawa and Ogasawara, 1991; Zakrzewska-Czerwinska and Schrempf, 1992), and are not expected to be frequent in a non-coding sequence of *Hf. volcanii*.

Scanning specifically for the eubacterial DnaA-box consensus sequence, TTAT(A/C)CA(A/C)A, revealed three possible binding sites, each of which contains two mismatches. In eubacterial origins, DnaA-boxes do vary from the consensus (Fujita *et al.*, 1990; Fujita *et al.*, 1992), but in this case all putative binding sites coincided with regions of relatively high AT sequence, and since the binding sites themselves are also AT-rich, the significance of these matches is questionable. Similarly, three *Saccharomyces* ARS Consensus Sequences (ACS) based on the sequence (A/T)TTTAT(A/G)TTT(A/T) were detected, but once again,

```
   1 GGGCCCCCTC AAAATCGGCG AGCAGACGAA CCGCGGCGAC CGAGATTCAC TGCCAGTTGG
  61 ACGTGGGACA GGGCGGCTAC CAGATTCCGA ACAACCCCGA CACCATCGAG TTCCTCGAAC
 121 ACGACATCGA CTTCGTCATG TGCGTCGAGA CCGGCGGGAT GCGCGACCGA CTCGTCGAAA
 181 ACGGCTTCGA CGACGACTAC AACGCGCTCG TCGTCCACCT CGGCGGCCAG CGGCGCGCGC
 241 CACCCGGCGT ATCACCAAGC GCCTGCACGA CGAACTCGAC CTGCCGGTGT GGTCTTCACC
 301 GACGGCGACC CGTGGTCCTA CCGCATCTTC GGCTCGGTCG CCTACGGCTC TATCAAATCC
 361 GCGCACCTCT CGGAGTACCT CGCCACACCC GACGCGAAGT TCGTCGGCAT CCAGCCGCAG
 421 GACATCGTTG ACTACGACCT CCCGACCGAC CCGCTCGCGA CTCCGACATC AACGCGCTCC
 481 AGTCCGAACT GGAGGACCCG CGGTTCATGG GCGACTACTG GACCGAGCAG ATAGAGCTCC
 541 AACTCGACAT CGGCAAGAAG GCAGAACAGC AGGCGCTTGC CTCCCGCGGT CTCGACTTCG
 601 TGACCGACGA GTACCTGCCG ACGCGCCTCG ACGAGATGGG TATCATCTAA CCCCGCTCAC
 661 TCCCGTCGCC GTTTTTCTAC CTCGTTTCGG TCAGATTGAA CAGCGCCACG ATAGTGAACC
 721 CGCCGTAGAA GCCGAGGGTG TGAATCGCCA GCTCGGTCGC GAACCCCGCG GTGGTTTGGA
 781 TGGGTTTGAT GTGAAAAAAC AGCGCGTCGG TCGTCAGGAC GACGAGCGTG GCGGCGAGAC
 841 AGATGAAGAG CCGACGAACA CTGAAAAGCG GGCTCGACCG GAGGGCGTCG CGCATGGCGG
 901 TTCGACGGTC GCCCGCTGTA TGAGCCTGTC GCCGCGCGGA GGGTTCGGAT TCGGTTCGGT
 961 GGCGTCGCGG CTCGCGGTTC GCGCTTACGG ATACAGCCCG TCGCCGCGGA ATCCGTCGAA
1021 ATAGCGTCTG ACCACGCCGT TCAGTGTTCG CGGCTCGCTC GTGCGAGGTC GTGACCCGCC
1081 CCGGAGAACA GCGCGAGGTC GCGTCCGCGA CGCCGCCTTT GAGTTCTCGA ATCCGCGGTT
1141 CGGGGAACAG TCGGTCTGCC TTCCCGGCGG CGACGAGCGT CGATGCGTCG ATGTCACCGA
1201 GTATCTCTCG GGAGTCGTGT TCGAGACAGG CCGTACAGGA GACGACCGCG TCGGCGGGGA
1261 CGGCGGGCCG GAAGTCGACG ACCCGGCCGG CCGCCTCGAT GAGCGCCGGG GGACCGTCGC
1321 TTCGAGGCCG GTCGCGGACT CCCGCTCCGG TCCGCGACGA CCTCGGCCCA CCGGCTCTTG
1381 CCCGCCAGCG AGCGCCAGCG CGTCACGACG TTCTCGCCGT GGCCGCCGAG TCGCGTCCCG
1441 GC GCGACGA CCGCCAGCGA GTCCACGTAG TGCCGTAGTC GGCGGCGAGG TACTGGGCGA
1501 C  .GCCGCC CATCGAGACG CCGATAACGT CGGCCGGCCA GAGGTCCTGT TCGTCGATGA
1561 CCCCGGCGTA GCCCGCGGCC ATGTCGCGGG TGGTCGAGCC GACCGGGAGG TGTCGCGAGC
1621 GGCCGACCAC CCACACGTCG CGGTCGTCGA ACTCGCGGAA CAGCGGAGCG CCGCGCAGTC
1681 CCGCCGTTCG GGTCGATGCG CTGGAAGGCG TCGGAGAGGC CGGGAAGCAC CACGAGCGGG
1741 TCGGCGTCGG CGTCACCGAA GCGGTAGTAG GCCGCCGGCC GCCGAGCATG CCGTAGTCGA
1800 GGTCCATACG CGAACCGACG GCCCGACGCG GCAAAATCGT TCGGTCCATT CGCCAGTCCG
1861 TCTGGTTGTC TGTTTGTCTG CCCGGCTGTC GGTCCCGCCG TCGCCGCCTC AGACCTGATC
```

Figure 1-5. Sequence of the 1.9 kb fragment conferring high-efficiency transformation. Numbering is from the *Apa*I site to the *Sau*3A1 site. The unusually AT-rich region runs from about nucleotide 670 to about 800.

each of these contain two mismatches. The ACS is also AT-rich and since these putative sites tend to fall close to, or overlap with the putative DnaA-boxes, it is likely that both motifs are artifacts arising from the local AT content.

Replication of *hft*-plasmids *in vivo*. To determine the state of the *hft*-plasmid *in vivo*, total DNA from three isolates of WFD11 transformed with plasmid bearing the 1.9 kb fragment was digested and probed with pLS47-4. The hybridisation pattern is consistent with the presence of greater than one genome equivalent of the plasmid integrated at the *hft* locus, with no evidence of any free plasmid (Figure 1-6). The plasmids appear to be integrated as tandem, head-to-tail multimers of more than three repeats, implying that the *Ha. marismortui* marker does not confer resistance at a single copy per cell in *Hf. volcanii*. If there is any free plasmid in *Haloferax*, it is extremely rare and may also be distributed between several different conformations, each of which would migrate differently during electrophoresis rendering them difficult to detect by Southern blot hybridisation.

Discussion

A locus from the chromosome of the archaebacterium *Haloferax volcanii* ,which confers a high-frequency of transformation on an otherwise non-viable plasmid was isolated in an attempt to define an archaebacterial replication origin. The sequence of this locus has none of the conclusive hallmarks of other replication origins, but it does have numerous repeats, AT-rich regions, and perhaps most significantly it apparently contains no coding region over the entire 1,920 bp that has been sequenced. Moreover, like eubacterial *ori*C loci, which are almost always found within 10 kb of *gyr*BA (Ogasawara and Yoshikawa, 1992), *hft* is less than 13 kb upstream of the *Hf. volcanii gyr*BA cistron. Since the *gyr*BA genes are the only members of the suite of eubacterial genes surrounding the origin to have recognised

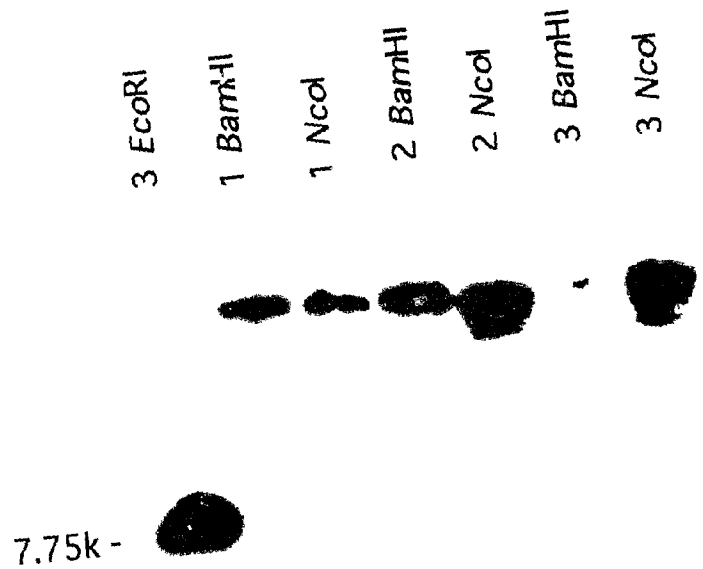


Figure 1-6. *hft* plasmid integration into the chromosome. Southern blot of total DNA from three WFD11 colonies transformed with the 1.9 kb *hft* clone in pLS47-4, probed with the same 1.9 kb clone. Lane one is transformant 3 digested with *Eco*RI, which cuts pLS47-4 opposite the polylinker. Lanes 2-7 are transformants 1 through 3 cut with *Bam*H1 and *Nco*I, enzymes that do not cut the plasmid or insert. The large bands that appear in lanes 2-7 are most easily seen as the result of the plasmid integrating into the chromosome, and tandemly multiplying. By this reasoning these large fragments are tandem multimers of the plasmid flanked by genomic DNA at the site of integration. Free plasmid would be uncut, and would likely appear in several conformations, but no other bands are detectable. In lane 1 the two large but faint bands visible above the 7.7 kb plasmid band may correspond to the two genomic fragments flanking the insertion, each of which would be fused to fragments of the first or last repeated plasmid.

archaebacterial homologues, the presence of a locus with these properties in this region of the chromosome is intriguing.

The inability of the locus immediately adjacent to *hft* to act similarly reveals that the observed activity is not a general characteristic of any fragment of the *Haloferax* genome, and argues that *hft* activity does impart greater plasmid viability in some way. Moreover, since *hft*-bearing plasmids could be passed through *Haloferax* and *E. coli* repeatedly, this effect is either due to free replication as plasmids, or integration and excision from the chromosome at an unusually high level. The demonstration that log phase transformants carry the plasmid integrated into the chromosome supports the latter, but it may not be so simple. *Ori*C plasmids in eubacterial systems have similar properties; they are all remarkably unstable, most are present at less than a single copy per chromosome, and are quickly lost, often regardless of selective pressure (Yee and Smith, 1990; Zakrzewska-Czerwinska and Schrempf, 1992; Zyskind *et al.*, 1983; Marczynski and Shapiro, 1992; O'Neill and Bender, 1988). One of the best studied examples is *Bacillus subtilis* where plasmids bearing chromosomal origins were also shown to be typically unstable (in freshly transformed colonies only a small fraction of the cells contain plasmids). More interesting still, is that in *Bacillus*, plasmids that are not lost are found integrated into the chromosome at a frequency of 100% while under selection (Moriya *et al.*, 1992). If this were taking place in *Haloferax*, the long doubling time would impede the isolation of transformants prior to integration, and might result in the observations described here. Moreover, the identification of a *Bacillus ori*C plasmid initially failed because few markers were able to confer a selectable phenotype at the extremely low copy number supported by the origin. Chromosomal integration and tandem amplification, as shown in Figure 1-6, may be the only way for transformants to survive the selection conditions imposed here.

Altogether, the *hft* fragment of the *Hf. volcanii* genome has many of the characteristics of an autonomously replicating sequence, but cannot readily be distinguished from a highly recombinogenic locus. Comparative studies in other archaebacteria, and a broader survey of conserved gene order between this region of the genome in archaebacteria and eubacteria are both necessary before any conclusions can be drawn.

Chapter II:

Calmodulin

Introduction

Calmodulin is a member of the EF-hand family, proteins defined by the presence of one or more short folds that bind calcium with a very high specificity (Strynadka and James, 1989; Heizmann and Hunziker, 1991). EF-hand proteins are found in both prokaryotes and eukaryotes (Swan *et al.*, 1987), but calmodulin has only been directly identified in eukaryotes, where it is responsible for regulating a number of physiological functions by activating other proteins in response to changes in the local concentration of calcium ions (for review see Klee *et al.*, 1980; Means and Dedman, 1980). Four calcium ions are bound, each by one of four EF-hands that are situated as two pairs of opposing folds joined by a short helix, resulting in a dumbell-like configuration. Of the four domains, pairwise similarity reveals that each of the two opposing pairs is most similar to the EF-hand in the same position on the other end of the dumbell. This suggests that the four-fold calmodulin protein evolved by two tandem duplications of an EF-hand motif.

There is biochemical evidence for the presence of calmodulin in the deep-branching diplomonad, *Giardia lamblia*, where a protein with many similar characteristics has been observed (Munoz *et al.*, 1987), but no sequence reported. A protein with traits resembling calmodulin has also been reported in the archaebacterium *Halobacterium halobium* (Rothärmel and Wagner, 1995). This raises the possibility that calmodulin may be even older than eukaryotes, but once again no sequence is known.

Results

Isolation of calmodulin genes. Genomic DNA from *Hexamita inflata*, *Trichomonas vaginalis*, *Naegleria fowleri*, and *Acrasis rosea* were used as templates in PCR amplification reactions with primers specific for all known calmodulin sequences (CAM-1 and CAM-2). In each case a single product of the expected size was isolated and three individual clones sequenced. All but that of *H. inflata* were shown to encode open reading frames with an extremely high sequence similarity to calmodulin (on the amino acid level 70%, 81% and 82.5% respectively identical to human calmodulin; Figure 2-1). In contrast, the *H. inflata* product was not recognisable.

Of the taxa for which calmodulin has been described, *T. vaginalis* is thought to be the first to have diverged from other eukaryotes (Gunderson *et al.*, 1995), so it was chosen for more detailed characterisation. The *T. vaginalis* PCR fragment was used as a probe against a Southern blot of genomic DNA from *T. vaginalis* (Figure 2-2). In DNA cut with *Eco*RI, *Hin*dIII or *Sau*3A, a single band was detected by the probe, suggesting that it recognises a single-copy locus (this is also supported by the agreement between this pattern and the map of the genomic clone described below). The largest of these fragments, a 1.25 kb *Hin*dIII fragment, was sought by inverse PCR amplification using primers based on the known sequence of the small fragment (I-CAM-1 and I-CAM-2), and circularised *Hin*dIII-digestion products of *T. vaginalis* genomic DNA as a template. A product of the expected size was isolated, cloned, and three individual copies sequenced.

The calmodulin gene was found to lie at one extreme end of the fragment, truncated at the amino terminus by a *Hin*dIII site that corresponds to codon 15 of the majority of known calmodulin genes (see Figure 2-1). The 3' end of the gene is marked by a termination codon at exactly the same position as that of most known calmodulin homologues, and the sequence is extremely conserved throughout the

```
Acrasis          ---------- ---------- ---------- ---------- ---------- ----------
N.fowleri        ---------- ---------- ---------- ---------- ---------- ----------
Trichomonas      ---------- ---------- -AFNIFDKDG DGRITAKELG TVMRSLGQNP SEAELQDMIN
N.gruberi        MSREAISNNE LTEEQIAEFK EAFSLFDKDG DGTITTSELG TVMRSLGQNP TEAELHDMIN
Solanum          ......MAEQ LTEEQIAEFK EAFSLFDKDG DGCITTKELG TVMRSLGQNP TEAELQDMIS
Oryza            ......MADQ LTDDQIAEFK EAFSLFDKDG DGCITTKELG TVMRSLGQNP TEAELQDMIN
T.cruzi          ......MADQ LSNEQISEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Euglena          ......MAEA LTHEQIAEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Plasmodium       ......MADK LTEEQISEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Stylonychia      ......MADN LTEEQIAEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Tetrahymena      ......MADQ LTEEQIAEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Drosophila       ......MADQ LTEEQIAEFK EAFSLFDKDG DGTITTKELG TVMRSLGQNP TEAELQDMIN
Homo             ......MADQ LTEEQVTEFK EAFSLFDKDG DGCITTRELG TVMRSLGQNP TEAELRDMMS
Aspergillus      ......MADS LTEEQVSEYK EAFSLFDKDG DGQITTKELG TVMRSLGQNP SESELQDMIN
Neurospora       ......MADS LTEEQVSEFK EAFSLFDKDG DGQITTKELG TVMLSLGQNP SESELQDMIN
Saccharomyces    ......MSSN LTEEQIAEFK EAFALFDKDN NGSISSSELA TVMRSLGLSP SEAEVNDLMN

Acrasis          --DADGNGTI DFPEFLTLMA RKMKDTDTEE EIRDAFKVFD KDGNGLISAA ELRHVMTNLG
N.fowleri        --DADGNGTI DFTEFLTMMA KKMKDTDNEE EIKEAFKVFD KDGNGFISAQ ELRHVMCNLG
Trichomonas      EIDLDGNGTI EFDEFLYMMN RQMKEGDTEE EIKDAFRVFD KDGDGKITAA ELAHIMKNLG
N.gruberi        EVDADGNGTI DFTEFLTMMA KKMKDTDNEE EIKEAFKVFD KDGNGFISAQ ELRHVMCNLG
Solanum          EADADQNGTI DFPEFLNLMA RKMKDTDSEE ELKEAFKVFD KDQNGFISAA ELRHVMTNLG
Oryza            EVDADGNGTI DFPEFLNLMA RKMKDTDSEE ELKEAFRVFD KDQNGFISAA ELRHVMTNLG
T.cruzi          EVDQDGSGTI DFPEFLTLMA RKMQDSDSEE EIKEAFRVFD KDGNGFISAA ELRHVMTNLG
Euglena          EVDQDGSGTI DFPEFLTMMS RKMHDTDTEE EIKEAFRVFD KDGNGFISAA ELRHVMTNLG
Plasmodium       EIDTDGNGTI DFPEFLTLMA RKLKDTDTEE ELIEAFRVFD RDGDGYISAD ELRHVMTNLG
Stylonychia      EVDADGNGTI DFPEFLSLMA RKMKDTDTEE ELVEAFKVFD RDGNGLISAA ELRHVMTNLG
Tetrahymena      EVDADGNGTI DFPEFLSLMA RKMKDTDTEE ELIEAFKVFD RDDNDLISAA ELRHVMTNLG
Drosophila       EVDADGNGTI DFPEFLTMMA RKMKDTDSEE EIREAFRVFD KDGNGFISAA ELRHVMTNLG
Homo             EIDRDGNGTV DFPEFLGMMA RKMKDTDNEE EIREAFRVFD KDGNGFVSAA ELRHVMTRLG
Aspergillus      EVDADNNGTI DFPEFLTMMA RKMKDTDSEE EIREAFKVFD RDNNGFISAA ELRHVMTSIG
Neurospora       EVDADNNGTI DFPEFLTMMA RKMKDTDSEE EIREAFKVFD RDNNGFISAA ELRHVMTSIG
Saccharomyces    EIDVDGNHQI EFSEFLALMS RQLKSNDSEQ ELLEAFKVFD KNGDGLISAA ELKHVLTSIG

Acrasis          EKLTD----- ---------- ---------- ------
N.fowleri        EKLTD----- ---------- ---------- ------
Trichomonas      EPLTQEEVDE MIAQADTNKD GIIDYGEFVH LMLTS*
N.gruberi        EKLTDEEVDE MIREADIDGD NQINYTEFVK MMMQK*
Solanum          EKLTDEEVDE MIREADIDGD GQVNYEEFVR MMLAK*
Oryza            EKLTDEEVDE MIREADVDGD GQINYEEFVK VMMAK*
T.cruzi          EKLTDEEVDE MIREADVDGD CQINYEEFVK MMMSK*
Euglena          EKLTDEEVDE MIREADVDGD CQINYEEFVK MMMSK*
Plasmodium       EKLTNEEVDE MIREADIDGD GQINYEEFVK MMIAK*
Stylonychia      EKLTDEEVDE MIREADVDGD GHINYEEFVR MMMAK*
Tetrahymena      EKLTDEEVDE MIREADIDGD GHINYEEFVR MMMAK*
Drosophila       EKLTDEEVDE MIREADIDGD GQVNYEEFVT MMTSK*
Homo             EKLSDEEVDE MIRAADTDGD GQVNYEEFVR V VSK*
Aspergillus      EKLTDDEVDE MIREADQDGD GRIDYNEFVQ LMMQK*
Neurospora       EKLTDDEVDE MIREADQDGD GRIDYNEFVQ LMMQK*
Saccharomyces    EKLTDAEVDD MLREVS.DGS GEINIQQFAA LLS.K*
```

Figure 2-1. Calmodulin sequences from *Trichomonas vaginalis, Acrasis rosea* and *Naegleria fowleri* aligned with those of some representative eukaryotes. Length heterogeneity at the amino terminus is indicated by spaces, gaps within the alignment by dots (.), missing data by dashes (-), and termination codons as asterisks (*).

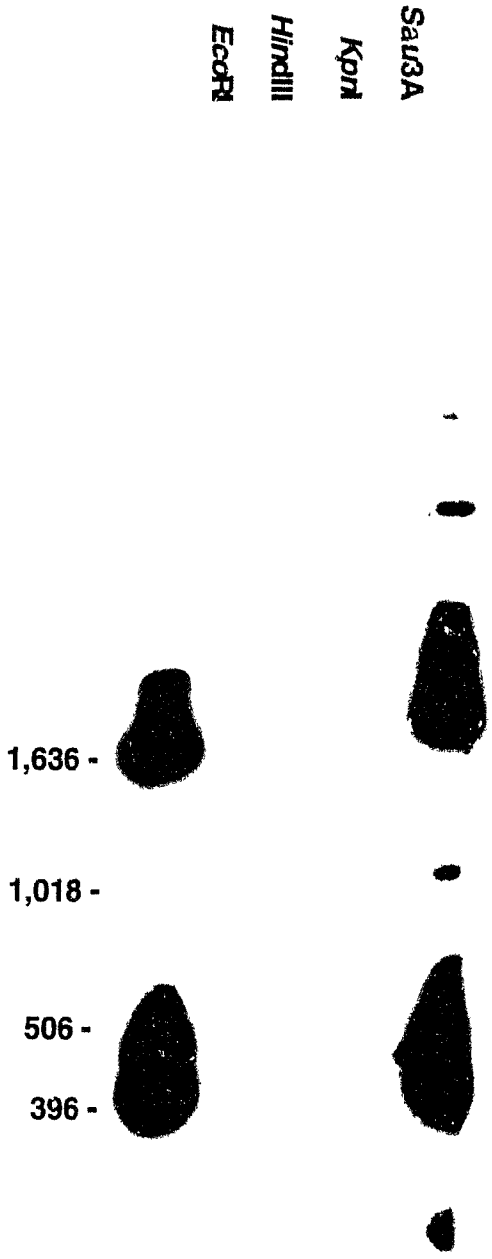


Figure 2-2. Southern blot of *T. vaginalis* genomic DNA probed with calmodulin PCR fragment. The single band in the *Hin*dIII lane corresponds to a 1.250 kb fragment which was subsequently isolated by inverse PCR and sequenced. The small *Eco*RI and *Sau*3A fragments which hybridized to the calmodulin probe were calculated to be congruent with the restriction map of the 1,250 bp *Hin*dIII clone. The *Kpn*I lane yielded only a very large, faintly-hybridizing fragment.

length of the gene (over 134 amino acid positions, the *T. vaginalis* calmodulin is identical to the human gene at 89 positions).

Discussion

Another among the list of characteristics that seem to define the eukaryotes is the use of calmodulin as a receptor of intracellular calcium. In an attempt to clarify whether calmodulin was present in the ancestor of extant eukaryotes, a fragment of the calmodulin gene from *Trichomonas vaginalis* (a Parabasalian), *Naegleria fowleri*, and *Acrasis rosea* (two Heterlobosea) has been sequenced and found to be extremely similar to homologues from other eukaryotes.

There is biochemical evidence for the presence of calmodulin in another protist taxon that perhaps diverged even earlier than either Parabasalia or Heterolobosea, the Diplomonad *Giardia lamblia.* Here a protein with many similar characteristics has been recorded (Munoz *et al.*, 1987), but since no sequence is known, it cannot conclusively be called calmodulin. Interestingly there has also been a report of a protein with traits resembling calmodulin in the archaebacterium *Halobacterium halobuim* (Rothärmel & Wagner, 1995). Again no sequence has been identified, but in light of the relatively close relationship between archaebacteria and eukaryotes, this may be an indication that the use of calmodulin predates eukaryotes. On the other hand there is no evidence for the existence of a eubacterial calmodulin and no such sequence appears in the genomes of either *Haemophilus influenzae*, or *Mycoplasma genitalium,* the only two eubacterial genomes that have been fully sequenced (Fleischman *et al.*, 1995; Fraser *et al.*, 1995).

Chapter III:
Ubiquitin
and E2 Ubiquitin-Conjugating Enzyme

Introduction

Ubiquitin is a small, highly conserved protein which is conjugated to other proteins. It predominantly serves as a signal for the degradation of misfolded or short-lived proteins, but ubiquitin-conjugation also plays a role in chromatin structure, DNA repair, cell-cycle control, membrane translocation, and a host of other cellular activities (Goldknopf and Busch, 1977; Jentsch *et al.*, 1987; Goebl *et al.*, 1988; Davie and Murphy, 1990; Holloway *et al.*, 1993; Sommer and Jentsch, 1993). In the conjugation pathway (reviewed in Jentsch, 1992), ubiquitin is first converted to an adenylated intermediate by E1 ubiquitin-activating enzyme, which proceeds to covalently bind the ubiquitin molecule through a thioester linkage. The ubiquitin moiety is subsequently transferred to E2 ubiquitin-conjugating enzyme by transesterification. This enzyme may then catalyse the formation of an isopeptide bond between ubiquitin and the target protein, in some cases through an E3-ubiquitin thioester intermediate (Scheffner *et al.*, 1995). These enzymes have only been characterised in animals, plants, and fungi where there is generally a single E1 and E3, but numerous families of E2 conjugating enzymes. The substrate choice for ubiquitination is to some extent specified by the different physical characteristics and activities of the E2 involved in ubiquitination, so the presence of distinct families is an important indicator of the activities of the pathway.

Ubiquitin itself was originally named for its presence in all cell types (Schlesinger and Goldstein, 1975; Goldstein *et al.*, 1975). Ironically the evidence originally presented for bacterial ubiquitin was questionable, and for some time it was believed that there was no bacterial ubiquitin system (see Zwickl *et al.*, 1990).

However, the discovery of a ubiquitin-mediated proteolytic pathway involving the 20S proteasome in the archaebacterium, *Thermoplasma acidophilum*, indicates that this function, at least, predates eukaryotes (Wenzel and Baumeister, 1993; Wolf *et al.*, 1993). Moreover, evidence has emerged for the presence of proteasome and ubiquitin-like polypeptides in a eubacterium as well (Lupas *et al.*, 1994; Durner and Börger, 1995; Rohrwild *et al.*, 1996). Nevertheless, this remains something of a puzzle as no prokaryotic gene encoding ubiquitin has been identified, leading to questions about the provenance of the peptide sequences. In support of this possibility is the absence of ubiqutin genes in the genomes of *Haemophilus influenzae* and *Mycobacterium genitalium* (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995), the only two complete eubacterial genomes that are currently public.

Ubiquitin genes are found in three major forms: as isolated open reading frames, as fusions with genes for small ribosomal proteins, or as polymers of head-to-tail ubiquitin-coding sequences which are co-translated and converted to monomers by proteolysis. In most eukaryotes there are many ubiquitin genes of one sort or another (the exception being *Giardia*; Krebber *et al.*, 1994). Since ubiquitin repeats are highly conserved, relatively short (228 nucleotides), often found in tandem, and have been sequenced from many taxa, these genes are an ideal model in which to study concerted evolution (Sharp and Li, 1985; Tan *et al.*, 1994).

Results

***Trichomonas vaginalis* and *Giardia lamblia* ubiquitin genes.** Genomic DNA from a number of protists, archaebacteria and eubacteria were used as templates in PCR reactions using primers UB-A and UB-3, with an annealing temperature of 55° and an extension time of 1 minute. Products corresponding to the predicted size of a single ubiquitin-coding region were observed only for

Giardia lamblia and *Trichomonas vaginalis*. These bands were isolated, cloned and sequenced, revealing that they were indeed ubiquitin genes. At the same time, a *G. lamblia* ubiquitin gene with the identical sequence was deposited in GenBank, and subsequently reported in Krebber *et al.* (1994) and was therefore no longer pursued.

The *T. vaginalis* product was hybridised to a Southern blot of genomic *T. vaginalis* DNA to confirm its provenance, and seven independent copies were sequenced. The nucleotide sequence of those clones varies at 35 out of 121 positions, resulting in three distinct polypeptide sequences. The possibility that these sequences are part of a polyubiquitin was addressed by amplifying under the same conditions with a set of primers, RUB-1 and RUB-2, designed to detect fused repeats. Products of the expected size were isolated and sequenced. The sequence confirmed the presence of at least one polyubiquitin locus as each contained the 3' end of a ubiquitin gene fused to the 5' end of a downstream gene.

To obtain a minimal estimate of the number of repeats in the locus, the amplification reaction buffer was optimised for amplification of products composed of greater than a single unit. As indicated above, the standard reaction conditions result in a single, monomer sized product, but it was found that by increasing the amount of Tris buffer in the reaction, larger multimers could be preferentially amplified. Figure 3-1 shows that these conditions result in products ranging in size from a single ubiquitin unit to at least six head to tail repeats. This not only confirms the presence of a polyubiquitin, but also gives a lower limit to its size. Several isolates from the band corresponding to a di-ubiquitin product were cloned and sequenced, and these were found to be of the expected structure.

In total, 14 amplification products of three different forms were sequenced. Of these 7 are monomeric (1A-G), four are dimeric (2A-D) and three correspond to junctions (jA-C). Of these sequences, several were represented more than once, or

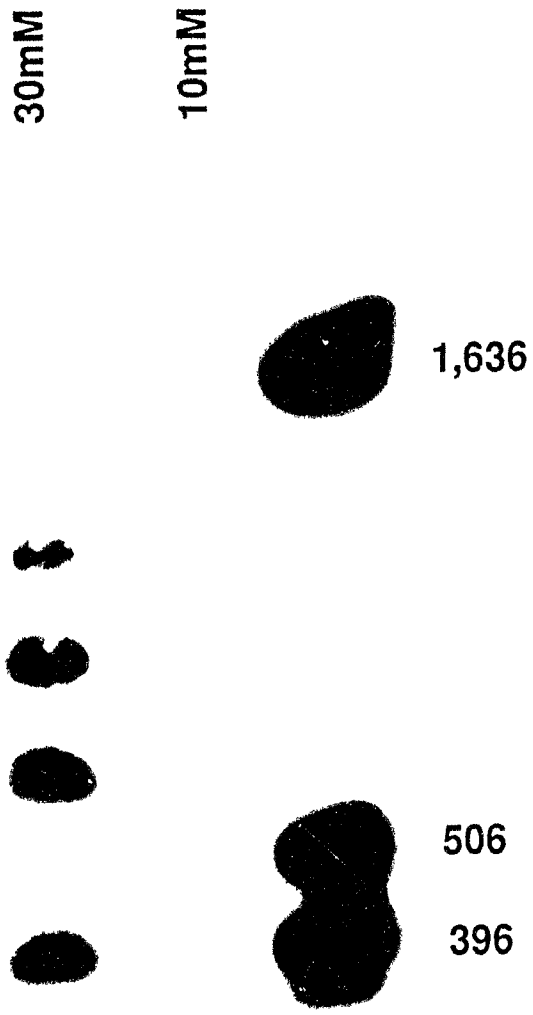


Figure 3-1. Amplification of polyubiquitin. Southern blot of PCR reactions hybridised to a cloned monomeric ubiquitin gene from *T. vaginalis*. This shows the effects of the altered reaction buffer compared to the standard reaction buffer (lanes 1 and 2 respectively). The reaction carried out with 30 mM Tris (see materials and methods) yields products ranging from a single monomer (173 nucleotides) to at least six tandem repeats (1,313 nucleotides), whereas the reaction performed in the standard 10 mM Tris yields only a single product corresponding to a single ubiquitin gene.

were subsequently found to exactly match part of the cDNA (in summary, 1D and 1B were identical; jB and jC were identical; 1E, 1F and 1G were identical; 2A, 2B, and jA were identical, and 2C and 2D were identical to one another and to the c9/c10 cDNA repeats). These are not included in the analysis.

Isolation and Sequencing of cDNA. To characterise a larger number of intact repeats in their natural order, a *T. vaginalis* cDNA library was screened with the previously sequenced amplification products. This resulted in the isolation of a single cDNA clone that was found to contain an insert of about 2.5 kb.

A 2.5 kb mRNA could potentially contain ten full ubiquitin repeats. Sequencing a long series of highly conserved repeats poses an interesting technical problem as the lack of heterogeneity between repeats precludes the use of primers or conven nal restriction subcloning. The high degree of homogeneity was in this case favorable, however, as it allowed a set of terminal deletions to be easily constructed based on the observation that each monomer contained a *Bgl*II restriction site immediately adjacent to the first methionine codon. 5' terminal monomers were deleted by incubating 1 microgram of cDNA clone with 15 units of *Bam*HI for 1.5h, followed by 5 units of *Bgl*II for 20 min, at which time half the reaction was removed and stopped by the addition of 10 mM EDTA. The remainder was allowed to digest for an additional 20 minutes and then the two halves were pooled once again. This digest was then electrophoresed overnight in 0.7% agarose and individual bands isolated. The purified fragments were composed of linear pBluescript fused to a known number of repeats, all flanked by compatible *Bam*HI and *Bgl*II overhangs. These were circularised overnight in a dilute ligation and used to transform *E. coli*. 3' terminal monomers were deleted in much the same way, except that the cDNA was first treated with *Bgl*II, then overhangs were filled in by incubation at 37°C for five minutes with Klenow and 10 mM deoxynucleotides in

the digestion reaction buffer. The DNA was then ethanol precipitated, resuspended and digested overnight with *Hinc*II, which recognises a site on the opposite side of the insert from the *Bam*HI site. Individual deletions were isolated and ligated as described above.

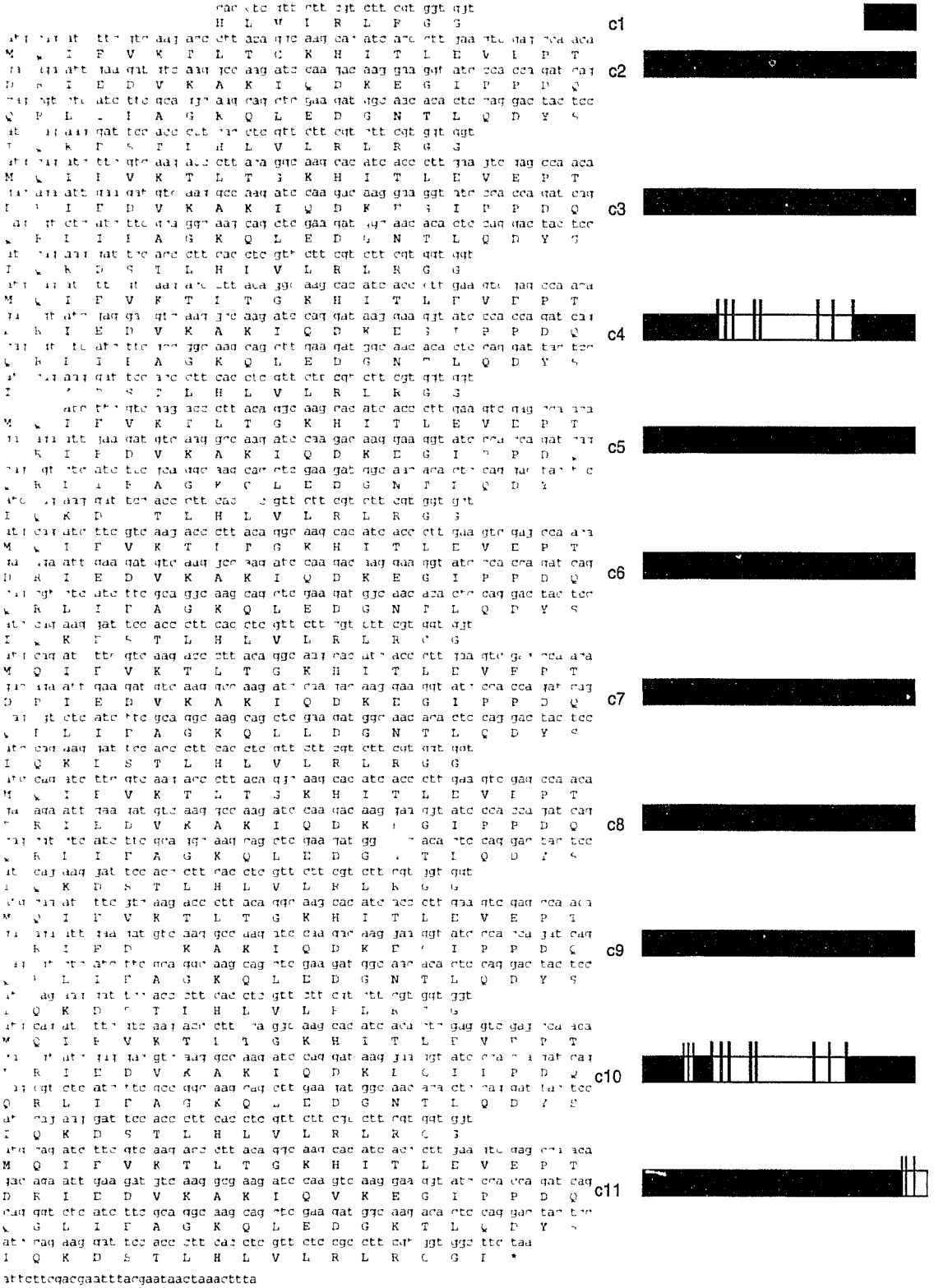
Sequencing in this manner revealed that the 5' end of the cDNA was truncated 39 bp upstream of the terminal glycine of a coding unit, which is followed by ten intact units, the last of which ends in an extra phenylalanine residue, a stop codon and 34 bp of untranslated sequence (Figure 3-2). The actual length of the gene is unknown, but has to be eleven units or greater. This is comparatively large, but by no means the largest reported polyubiquitin allele (Wong *et al.*, 1992).

The unique feature of this polyubiquitin is the high degree of conservation between coding units: of the ten complete units, seven have precisely the same nucleotide sequence, and two of the other three share a block containing ten substitutions from this sequence.

Concerted evolution of ubiquitin repeats. To analyse this rather stark case of concerted evolution, pairwise distances between coding units were calculated and phylogenetic trees constructed based on unweighted parsimony. The data set used contained only the *Trichomonas* sequences that are unique or present in the cDNA; this was thought to best represent the range of sequences present in the genome by avoiding over-representation of preferentially amplified PCR products, while at the same time considering the repetitive nature of the cDNA. The final set of data considered is shown in Figure 3-3 where the repeated cDNA sequence is used as a standard to which the other sequences are aligned.

Ubiquitin is not a good marker for phylogeny as it is too short, and too highly conserved. However, phylogenetic tree construction may still be useful to show that the members of a repeat family are all more closely related to one another

Figure 3-2. Sequence and schematic diagram of the cDNA clone showing identical repeats and the position of substitutions that vary from this sequence. On the left, the cDNA is composed of a 5' truncated repeat, ten intact repeats and a short 3' untranslated region. On the right, the repeats shown in gray are all absolutely identical, deviations from this sequence in the other three are shown as thin vertical marks to denote a substitution unique to that repeat, and a heavy vertical mark within a white box to denote a set of variant positions which are shared between two of the repeats.



attcttcgacgaatttacgaataactaaacttta

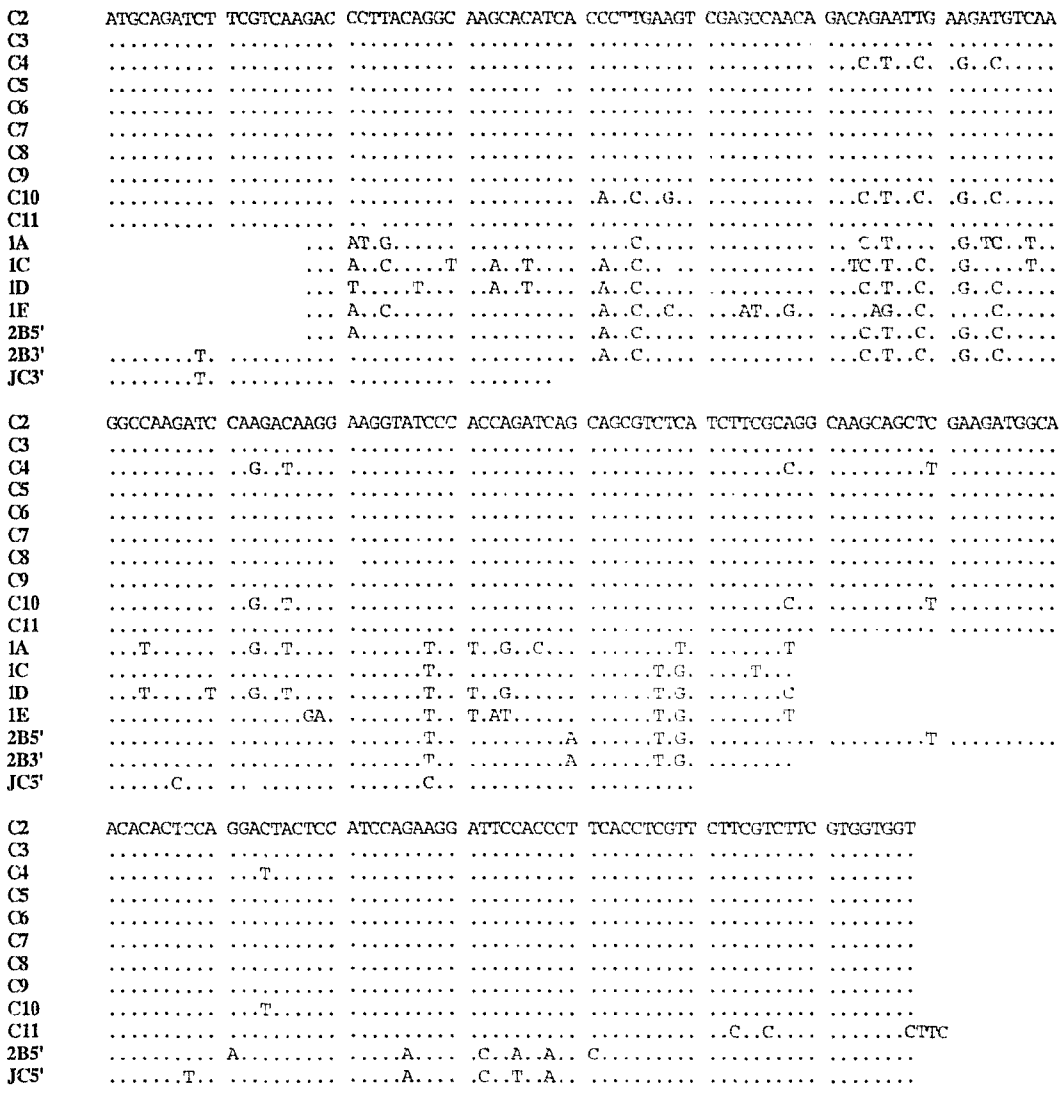


Figure 3-3. Nucleotide sequences of ubiquitin genes from *Trichomonas vaginalis*. Only unique sequences, and those of the cDNA are shown.

than to any other known sequence. This is seen to be the case in Figure 3-4, where all the nucleotide sequences from a particular taxon group together to the exclusion of all other sequences, a sign that they are evolving together. It is noteworthy that *T. vaginalis* units 1A and 1E, which are the two that differ at the amino acid level, are found to branch well within the *Trichomonas* cluster, and that the substitutions that unite cDNA units c4 and c10 are also common to most of the amplification products.

Table 3-1 Substitutions Between Repeats in Polyubiquitin Loci

Species	Mean	Corrected Mean	Range	Corrected Range
Zea maize	23.5	0.103	8 to 33	0.035 to 0.145
Phytophthora infestans	16.4	0.072	5 to 24	0.022 to 0.105
Geodia cydonium	3.8	0.017	1 to 7	0.004 to 0.031
Cricetulus griseus	3.7	0.016	0 to 5	0 to 0.022
Bombyx mori	4.2	0.018	1 to 6	0.004 to 0.026
Bos taurus	32.3	0.142	30 to 35	0.132 to 0.154
Euplotes eurystor. is	16.3	0.071	10 to 20	0.044 to 0.088
Tetrahymena pyriformis	63.4	0.278	36 to 95	0.158 to 0.417
Trypanosoma cruzi	7.3	0.032	7 to 8	0.031 to 0.035
Trichomonas vaginalis	**20.5**	0.090	0 to **45**	0 to 0.198

Table 3-1 addr .es the range of variability found between repeats in a number of species. Similar tables can be found elsewhere (Sharp and Li, 1987; Tan *et al.*, 1993), so the concentration here is on more recent data, while trying to give a good representation of the range of homogeneity. The mean and upper range of pairwise distances found in *T. vaginalis* are higher than those of most other taxa, while not the highest, arguing that there is a relatively high degree of variability in *Trichomonas,* despite the block of extremely homogeneous repeats in the cDNA.

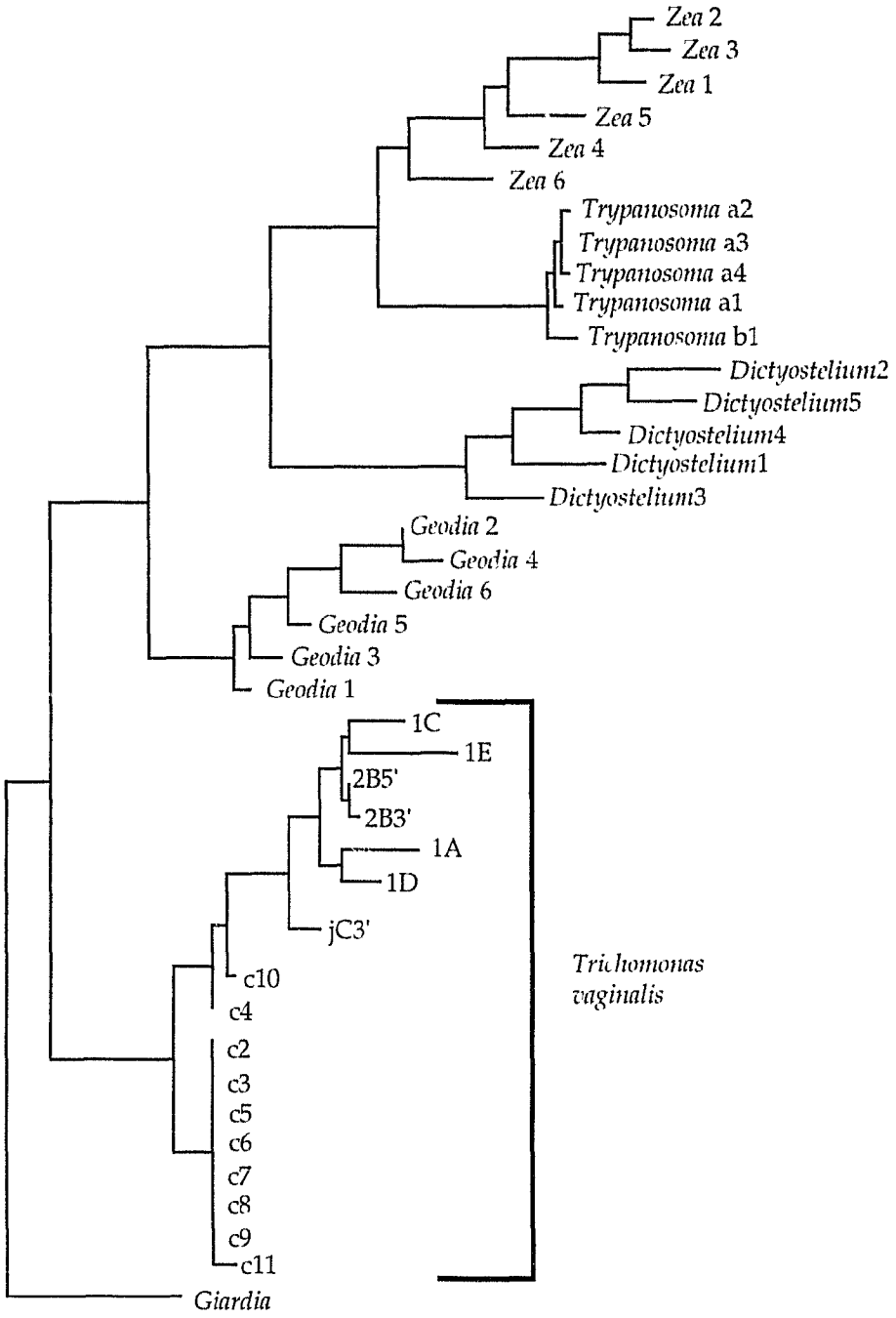


Figure 3-4. Parsimony tree of ubiquitin nucleotide sequences. Polyubiquitin genes were divided into individual monomers, aligned with the nucleotide sequences from *Trichomonas* and the most parsimonious tree found using PAUP under default conditions. The tree shows that sequences from a particular species form a coherent group. In addition, it can be seen that variable repeats c4 and c10 from the *Trichomonas* cDNA are more similar to the amplification products than the seven identical repeats.

***T. vaginalis* UBC1, a member of the E2 ubiquitin-conjugating enzyme family**. Downstream of the *Trichomonas* calmodulin gene reported in Chapter 2, there are 461 bp of extremely AT-rich non-coding DNA, which is followed by another open reading frame on the opposite strand. This open reading frame (which, like the calmodulin ORF, is truncated by a *Hin*dIII site) encodes a sequence with high similarity to E2 ubiquitin-conjugating (UBC) enzymes. These enzymes have until now only been found in animals, plants, and fungi, where they make up a large multi-gene family. The best sampling of diverse *UBC* genes is currently found in *S. cerevisiae*, where twelve individual members have been identified. In Figure 3-5 the inferred amino acid sequences of these twelve genes are aligned to the *T. vaginalis* sequence (named *Tv*UBC1 to conform to one existing nomenclature; Sullian and Vierstra, 1991). From the alignment the similarity of the inferred *T. vaginalis* amino acid sequence to other homologues can be seen to concentrate around the otherwise highly conserved domains, especially surrounding the catalytic cysteine residue at position 127 (boxed in Figure 3-5), which forms the actual thioester bond to ubiquitin (Sung *et al.*, 1990; Sullivan and Vierstra, 1991; Jentsch, 1992).

Different types of E2 ubiquitin-conjugating enzymes have been classified by different physical properties and activities, including the presence of a long, often acidic carboxy terminal extension, the ability to ubiquitinate histone, and the requirement for E3 ubiquitin-protein ligase (reviewed in Jentsch *et al.*, 1990; Qin *et al.*, 1991; Jentsch, 1992). However, there is a strong correlation between having a carboxy terminal extension and being able to ubiquitinate histones without E3, which suggests that some of these characters may be functionally related. It would be useful, therefore, to make a phylogenetic classification of ubiquitin-conjugating enzymes and in this way determine the relationship of the *T. vaginalis* sequence to other E2 sequences.

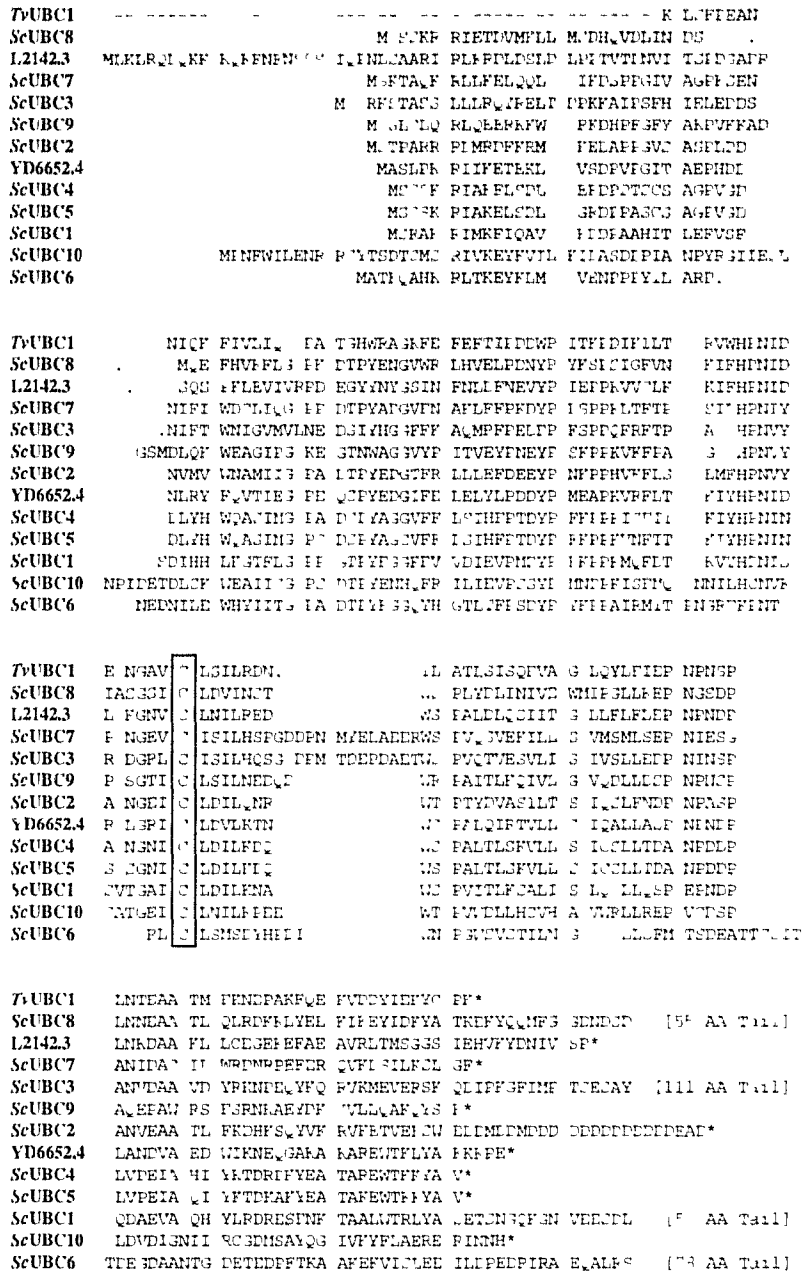


Figure 3-5. Alignment of *Tv*UBC1 amino acid sequence with that of twelve known *UBC* genes from *S. cerevisiae*. Genes are named where possible by one conventional nomenclature where the first two letters are the organism's initials (in this case *Sc* is *S. cerevisiae*) followed by UBC for ubiquitin-conjugating enzyme, and a number to distinguish paralogous enzymes from the same genome. Two homologues in *S. cerevisiae* that are only known from genome sequencing are also shown, L2142.3 and YD6652.4 (GenBank accessions U17247 and Z50111 respectively). Length heterogeneity is indicated by spaces, gaps in the alignment are shown as dots (.), missing data in *T. vaginalis* as dashes (-), and termination codons as asterisks (*). Four of the *S. cerevisiae* proteins, UBC1, 3, 6, and 8, have considerable carboxy terminal extensions, which are not shown, but the length is given in square brackets. The cysteine residue that forms the thioester bond with ubiquitin is boxed.

The alignment in Figure 3-5 is a sample of a larger one composed of 50 *UBC* genes from animals, plants, and fungi which was used to infer phylogenetic trees by both distance and parsimony methods. A distance tree based on 154 positions of the conserved UBC core is shown in Figure 3-6, where suggested subfamilies are indicated. Each subfamily comprises a group of sequences separated from other subfamilies by a highly significant branch. Also, the branching order of taxa within subfamilies does not strongly contradict what is known of the organismal phylogeny. For instance, the animals, plants, and fungi are not interspersed among one another. Different datasets varying in the inclusion of positions of ambiguous alignment were also examined (most excluding the short region missing from *Tv*UBC1), and in all cases the subfamilies shown were conserved while the order between families was slightly variable. Based on the low level of significance for all inter-subfamily nodes, it is doubtful that the branching order of subfamilies has much meaning. In general, however, the known *UBC* genes can at least be divided into nine subfamilies and a handful of highly divergent sequences that are difficult to place for lack of clear orthologues. These subfamilies are also supported by parsimony analysis, which resulted in 528 equally parsimonious trees, the strict consensus of which includes all these groups and maintains much of the same branching order between groups.

In Figure 3-6 the *Trichomonas* enzyme is affiliated with a *S. cerevisiae* open reading frame known only from the sequence of chromosome XII. This enzyme has not been functionally characterised and its role in the cell is completely unknown.

Two other clearly cohesive groups (represented by *Sc*UBC8 and *Sc*UBC1 proteins) are comprised entirely of enzymes with carboxy extensions, and several families are totally void of this type of enzyme. However, two subfamilies are composed of a mixture of enzymes with and without carboxy terminal extensions.

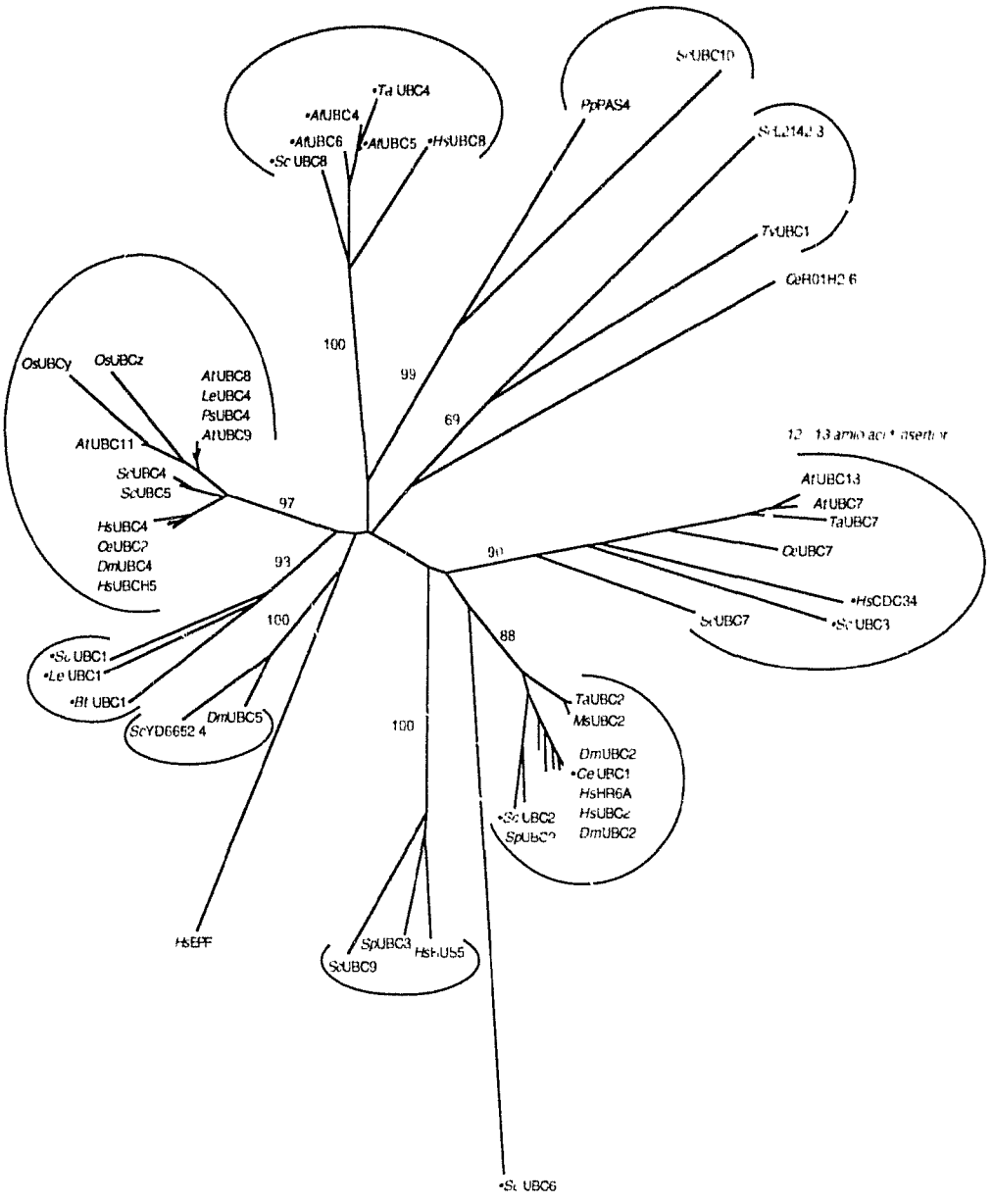


Figure 3-6. Neighbor joining tree of 50 UBC proteins with bootstrap percents shown for major nodes over 50%. Nine significant clusters are defined, each of which is bracketed. The group represented by *Sc*UBC3/7 is also defined by a shared 12-13 amino acid insertion unique to this group. Sequences that have carboxy terminal extensions are identified by a dot (•) preceding their name. The nomenclature is the same as in Figure 3-5, consisting of the initials of the organism (*Tv, T. vaginalis; Sc, S. cerevisiae; Sp, Schizosaccharomyces pombe; Pp, Pichia pastoris; Ce, Caenorhabditis elegans; Dm, Drosophila melanogaster; Hs, Homo sapiens; Bt, Bos taurus; At, Arabidopsis thaliana; Ls, Lycopersicon esculentum; Ts, Triticum aestivum; Os, Oryza sativa; Ms, Medicago sativa*) followed by their number where specified by the original authors in database entries. However, not all genes have been assigned such names: exceptions and their GenBank accession numbers are *Hs*HUS5 (U29092), *Hs*EPF (M91670), *Hs*HR6A (M74522), *Hs*CDC34 (L22005), *Ce*R01H2.6 (U00035), *Os*UBCy (U15971), *Os*UBCz (D17786), and *Pp*PAS4 (U12511).

This is particularly obvious in the group that includes *Sc*UBC3 and *Sc*UBC7; this family shares a conserved insertion in the UBC core, but only *Sc*UBC3 and *Hs*CDC34 have carboxy extensions. In general the enzymes with carboxy extensions are not all directly related to one another, which implies that classification based on these physical properties in some cases does not reflect relatedness.

Discussion

Ubiquitin has an enormous variety of roles in eukaryotic cell biology, acting in both highly specialised processes (Ball *et al.*, 1987; Kelly *et al.*, 1991; Früh *et al.*, 1994), and more general "housekeeping" pathways (Murti *et al.*, 1988; Davie and Murphy, 1990; Ghislain *et al.*, 1993; Sommer and Jentsch, 1993; Holloway *et al.* 1993). One of these housekeeping processes is the degradation of unfolded or improperly folded proteins by the proteasome, a large multisubunit particle that is activated by ubiquitin units covalently bound to the target protein. The proteasome has now also been discovered in the archaebacterium, *Thermoplasma acidophilum* (Zwickl *et al.*, 1991; Zwickl *et al.*, 1992), and there is evidence of genes with sequence similarity to proteasome subunits in eubacteria (Lupas *et al.*, 1995). However, it is not entirely clear what the role of ubiquitin is in these organisms, or if it even exists. A single short peptide corresponding to ubiquitin has been reported from *Thermoplasma* (Wolf *et al.*, 1993) and there have been similar reports of ubiqutin protein sequences from the cyanobacterium, *Anabaena variabilis* (Durner and Börger, 1995). However, neither the *A. variabilis* nor the *T. acidophilum* ubiquitin sequences have been verified by the actual identification of the gene, and considerable efforts to isolate ubiquitin from other archaebacteria have failed (Pühler *et al.*, 1994). Moreover, in neither of the two eubacterial genomes that have been completely sequenced (Fleischmann *et al.*, 1995; Fraser *et al.*,

1995), nor in the over one million base pairs known from the cyanobacterium *Synechocystis* PCC6803 (Kaneko *et al.*, 1995) is there any sequence resembling ubiquitin or anything closely related to the two classes of proteasome subunit.

The proteasome from *T. acidophilum* has been shown to have an increased activity on proteins that have been ubiquitinated (Wenzel and Baumeister, 1993), which circumstantially supports the notion that ubiquitin conjugation is also part of archaebacterial proteolysis. However, the proteasome selectively degrades misfolded or unfolded proteins, and ubiquitination has a chaotropic effect on protein structure, partly unfolding the target peptide (Wenzel and Baumeister, 1993). Therefore, the presence of ubiquitin in these *in vitro* studies may artifically increase the activity of the *Thermoplasma* proteasome simply by unfolding the substrate.

The uncertainty surrounding the presence of ubiquitin in archaebacteria will surely be resolved by the soon-to-be-released complete genome sequence of an archaebacterium, but at this time it appears likely that they do not use ubiquitin. Here the genes for ubiquitin were identified in two of the earliest-diverging eukaryotic lineages, *G. lamblia* and *T. vaginalis*, members of the Diplomonads and Parabasalia respectively. During the course of this work Krebber *et al.* (1994) characterised the ubiquitin gene complement of *G. lamblia* and demonstrated that this organism contains a single ubiquitin gene, which they sequenced. With the addition of these genes, ubiquitin has been identified in most major eukaryotic lineages, showing that it evolved prior to the divergence of extant eukaryotes.

The ubiquitin sequences from *Trichomonas vaginalis* are surprising for two reasons: in one sense they are extremely variable, while in another they are uncommonly conserved. In other polyubiquitin genes (except those of *Tetrahymena* and *Geodia*) all nucleotide variation occurs in the form of synonymous substitutions. This is also the case in the cDNA from *Trichomonas*, but several of

the PCR products vary from this sequence, resulting in seven variable sites, and a total of three different amino acid sequences.

These amino acid sequences are themselves interesting as the *Trichomonas* sequences contain numerous unique substitutions, many of which are either anisosteric or alter the charge. One of particular interest is N54, which is an otherwise highly conserved arginine that has been demonstrated through site-directed mutagenesis to be involved in the formation of the ubiquitin-adenylate intermediate in the conjugation of ubiquitin to the ubiquitin-activating enzyme E1 (Burch and Haas, 1994). The presence of an asparagine at this position in *Trichomonas* and a lysine in *Entamoeba* raises questions about the contribution of this residue to conjugation in these organisms.

By far the most unusual feature of the *Trichomonas* ubiquitin genes, however, is found at the nucleotide level. By analysing the pairwise distances between repeats, it is clear that there is generally a high degree of variability between repeats relative to that observed in other taxa. In contrast, the repeats at the 3' end of the polyubiquitin are remarkably homogeneous. Seven out of ten repeats in this region are absolutely identical, suggesting that these repeats are a special case, and their homogenisation most likely a recent event.

Homogenisation may result from unequal crossing over, gene conversion, and transposition. Unequal crossover events will usually affect tandem repeats, and lead to an allelic heterogeneity in the number of repeats. Gene conversion can also operate on non-allelic repeats, and can homogenise sequences without necessarily altering their frequency in the genome. Transposition (or episodic pseudogene formation) generally leads to changes in the number of repeats by the creation of new copies (Dover, 1982). Both gene conversion and unequal crossing-over have been described in polyubiquitin genes (Baker and Board, 1987; Sharp and Li, 1987), and both require multiple events to generate a tandem array of identical

repeats. In the case of gene conversion, this would also necessitate the repetitious involvement of a particular donor, but multiple unequal crossovers could easily yield tandem replications of the sort observed here with ut any special conditions. However, if the long stretches of homogeneity observed in the cDNA are likely the product of unequal crossing-over events, the pattern of identity between cDNA repeats c4 and c10 are almost certainly the product of gene conversion as two identical blocks of sequence are surrounded by somewhat different contexts. The only conclusion seems to be that the evolution of this locus involved a complicated series of events that probably included both gene conversion and unequal crossing-over.

The finding of E2 ubiquitin-conjugating enzyme in *T. vaginalis* was a nice surprise and significantly adds to the understanding of the role of ubiquitin in *Trichomonas*. This enzyme catalyses the formation of an isopeptide bond between ubiquitin and the target protein, in some cases through an E3-ubiquitin thioester intermediate (Scheffner *et al.*, 1995). There are numerous families of conjugating enzymes, and the substrate choice for ubiquitination is to some extent specified by the different physical characteristics and activities of the E2 involved in ubiquitination. Of the known *UBC* genes, the *T. vaginalis* E2 enzyme is most similar in sequence to L2142.3, an uncharacterised gene on chromosome XII of *S. cerevisiae*, so little can be inferred about the function of *Tv*UBC1. Nevertheless, the relatively firm relationship of *Tv*UBC1 with one particular *S. cerevisiae* sequence does give some indication that *T. vaginalis* likely also has multiple E2 ubiquitin-conjugating enzymes; otherwise the root of the tree would have to lie in the branch leading to *T. vaginalis*, and *Tv*UBC1 would be unlikely to have a specific affinity to any particular subfamily. This implies that some distribution of function among UBC proteins may be found in early diverging eukaryotes, but this

cannot be known with any certainty until more *UBC* genes from these taxa are identified and assigned to other subfamilies.

Chapter IV:

Tubulins

Introduction

The tubulin gene family consists of three distinct but highly conserved sub-families, alpha, beta and gamma-tubulin, each defined by sequence conservation, a wide distribution among eukaryotes and, where studied, a conservation of function. Of the three varieties, alpha and beta-tubulins are the most abundant in the eukaryotic cell and have been studied most extensively. Heterodimers of these two proteins are the primary constituents of microtubules, which in turn are central to the composition of eukaryotic flagella, cilia, mitotic spindles and the cytoskeleton. Gamma-tubulin was discovered much later (Oakley and Oakley, 1989) and its function is less clear, although it is known to be important in microtubule organising centres, or MTOCs (Oakley *et al.*, 1990; Zheng *et al.*, 1991), and has been implicated in several other processes (Gard, 1994; Lajoie-Mazenc *et al.*, 1994). Recently, two additional tubulin families have been proposed based on the identification of two unusual and highly divergent sequences, the so-called delta-tubulin found in *Caenorhabditis elegans*, and the epsilon-tubulin from *Saccharomyces cerevisiae* (Burns, 1995). While it is true that these sequences are very distant from other known tubulins, their apparent restriction to a single taxon each implies that they may not represent novel gene families but rather unique genes specific to the lineages in which they have been described.

Each tubulin orthologue is unique to eukaryotes, but the tubulin family as a whole does have a prokaryotic antecedent in the FtsZ protein, a component of the eubacterial cytokinesis system (see Bi and Lutkenhaus, 1991; Donachie, 1993). The notion that tubulins are derived from FtsZ was first put forward by Lutkenhaus from observation that the GTPase domains of both proteins share weak but

detectable sequence similarity and a few physica properties (Lutkenhaus, 1993). Since then functional and structural evidence for this relationship has accumulated appreciably: FtsZ has been found to assemble into tubules in a GTP-dependent process not unlike the polymerisation of microtubules (Bramhill and Thompson, 1994; Erickson *et al.*, 1996). Evidence for sequence similarity also derives from the identification of an archaebacterial homologue of FtsZ, which is closer in sequence to eubacterial FtsZ proteins, but is also more like tubulin than any previously characterised FtsZ homologue (Margolin *et al*, 1996).

Presumably the three tubulins diverged from a single ancestral FtsZ, but it is not known when this triplication took place or in which order the paralogues arose. Of alpha, beta and gamma, beta-tubulin currently enjoys the widest taxonomic representation. Beta-tubulin genes have been found even in the earliest-diverging eukaryotes (Kirk-Mason *et al.*, 1988; Katiyar and Edlind, 1994; Edlind *et al.*, 1996), demonstrating that this orthologue of the tubulin family predates the divergence of extant eukaryotes. However, the data on archezoal tubulins is restricted to beta: alpha and gamma-tubulins have been identified in a few protist lineages, but none that diverged so early in eukaryotic evolution (Lai *et al.*, 1988; Sanchez *et al.*, 1995). This leaves some uncertainty as to when alpha and gamma-tubulins diverged, before or after the appearance of extant eukaryotes.

Results

Identification of Tubulin Genes in Ancient Eukaryotic Lineages. A battery of universal (all tubulins) and gamma-tubulin-specific primers were used to try to identify even a small fragment of the gamma-tubulin gene in a variety of eukaryotes, but unfortunately with no success. Seven clones of the sizes expected from numerous primer combinations were isolated from *Trichomonas vaginalis*,

Giardia lamblia, *Nosema locustae*, and *Encephalitozoon hellem*, but none was found to encode tubulins when sequenced.

Similar attempts to identify alpha-tubulins were considerably more successful. Using ATUB-A and ATUB-B primers, products of the expected size were isolated from the Diplomonad *Hexamita* 50330, the Parabasalia *Trichomonas vaginalis*, *Tritrichomonas foetus*, *Trichomitus batrachorum*, and *Monocercomonas* sp., the Heterolobosean *Acrasis rosea*, and the Microsporidia *Nosema locustae*, *Encephalitozoon hellem*, and *Spraguea lophii*. These were cloned and the ends sequenced, revealing that each encoded a gene with a high resemblance to alpha-tubulin. The four parabasalian sequences (*Trichomonas vaginalis*, *Tritrichomonas foetus*, *Trichomitus batrachorum*, and *Monocercomonas* sp.) all proved to be extremely similar, so the sequencing of genes from *Trichomonas vaginalis* and *Tritrichomonas foetus* was not continued. In addition, two variants from *Monocercomonas* were found that differed at 17 positions (15 transitions and 2 transversions), resulting in two conservative amino acid substitutions (both due to transitions). The seven genes that were completed were all subcloned into fragments ranging in size from 200 bp to 800 bp in pBluescript using restriction enzymes appropriate for each individual gene. The actual sequencing was carried out either by manually sequencing subclones and gap filling using primers, or using ABI 373A or LiCor automated sequencing machines.

An alignment of the inferred amino acid sequences of these genes is shown in Figure 4-1. These genes are from taxa that are among the deepest-branching eukaryotes known according to molecular and ultrastructural data (see Cavalier-Smith, 1993), but are nevertheless extremely similar to known alpha-tubulin homologues. There are a number of conserved motifs that have a defined function in tubulin proteins (for review see Burns, 1991) that are also maintained in all these sequences except in those from the microsporidia, where there are two

```
H.30 LFCLEHGIHQDGQMPSDKSIGVAEDSFNTFFSETGAGKHVPRCVYIDLEPTVVDEVRAGAYRQIYHP  ISGKED
A.ro LYCLEHGIQPDGQMPSDKTIGVEDDAFNTFFSETGAGKHVPRAVFLDLEPTVIDEVRTGTYRQLFHPEQLISGKED
T.ba LYCLEHGIQPDGQMPSDKTIGICDDAFNTFFSETGACKHVPRAVMVDLEPTVVDEVRTGTYRQLWHPEQLINGKED
M. 1 LYCLEHGIQPDGQMPSDKTIGVCDDAFNTFFSETGAGKHIPRAVFVDLEPTVVDEVRTGTYRQLFHPEQIINGKED
M. 2 LYCLEHGIQPDGQMPSDKTIGVCDDAFNTFFSETGAGKHVPRAVFVDLEPTVVDEVRTGTYRQLFHPEQIINGKED
E.he LYCKEHGILPDGRLDQNRM..DDES.AESFFSQTSVGTYVPRTLMVDLEPGVLESIKTGKYRELYHPGQLISGKED
S.lo LYCKEHGILPDGTPDPNFN..DKESYSSTFFSETSGGNFVPRALMIDLEPGVIDSIKTSEYKNLYHPSQLIAGQED
N.lo LYCKEHNIRPDGTTGGV.....DDS.CSSFFIETSAGTYVPRTLMVDLEPGVIESIKNSEYRALYHPSSLINGKED

H.30 AANNYARGHYTVGKEVVDLVLDRLRKLADDCSGLQGFMLHHSFGGGTGSGLGSLILERLSVDYGRKTKLEFVIYPSL
A.ro AANNYARGHYTIGKEIVDLCLDRIRKLADNCTGLQGFLVFNSVGGGTGSGLGALLLERLSVDYGKKSKLGFTVYPSP
T.ba AANNYARGHYTVGKEIIDLTLDRIRKLADQCTGLQGFLIFHSFGGGTGAGFGSLLLERLSVDYGKKSKLEFTVYPAP
M. 1 AANNYARGHYTVGKEIIDLTLDRIRKLADQCTGLQGFLIFHSFGGGTGAGLGSLLLERLSVDYGKKSKLELTVYPAP
M. 2 AANNYARGHYTVGKEIIDLTLDRIRKLADQCTGLQGFLIFHSFGGGTGAGFGSLLLERLSVDYGKKSKLEFTVYPAP
E.he AANNYARGHYTVGKEIIEPVMEQIRRMADNCDGLQGFLIYHSFGGGTGSGFASLMMDRLAAEFGKKSKLEFSVYPAP
S.lo AANNYARGHYTAGKEIIEKVTDQIKRIAENCSGLQGFLVFHSFGGGTGSGFGALLMDRLSVEFGKKSKLEFAIYPSP
N.lo AANNYARGHYTVGKEIIEPVMEQIRRMADCCDGLQGFLIFHSFGGGTGSGFGGLLMDRLSQEFGKKSKLEFSVYPAP

H.30 SIAVSVVEPYNTVLAAHCMLEHSDCAFMIDNEAMYDICHRNLDIERCTYTNINRIVAQMISGMTASLRFDGALNVDL
A.ro QVATAVVEPYNSVLSTHALLEHTDVAVMLDNEAIYDICRRSLDIQRPTYTNLNRLVAQVISSLTCSLRFDGALDVDV
T.ba QVSTAVVEPYNSILATHAMIDHSDCAFMVDNEALYDLCRRALDIERPTYTNLNRLIGQVVSSLTASLRFDGALNVDF
M. 1 QVSTAVVEPYNSILATHAMIDHSDCAFMVDNEALYDLCRRALDIERPTYTNLNRLMGQVVSSLTASLRFDGALNVDF
M. 2 QVSTAVVEPYNSILATHAMIDHSDCAFMVDNEALYDLCRRALDIERPTYTNLNRLIGQVVSSLTASLRFDGALNVDF
E.he KIATAVVEPYNSILTTHTTLDYSDCSFLVDNEAIYDMC.RNLGIQRPYYTDINRIIAQVVSSITASLRFPGSLNVDL
S.lo RISTAVVEPYNSILTTHTTLNHFDCSFLVDNEAIYDIC.KNLGIAMPHANDLNKCITQVVSSITASLRFPGSLNVDL
N.lo RIATAVVEPYNSILTTHTTLDHSDCSFLVDNEAIYDMC.RNLGIERPKYKEINRVLAQVVSSITASLRFPGSLNVDL

H.30 TEFQTNLVPYPRVHFPFCSYAPLVSSEKAYHEKLTVAEITNSVFEPANMMVKCDPRHGKYMACCMMYRGDVVPKDVN
A.ro TEFQTNLVPYPRIHFMLCSIAPVISAEKAYHEQLSVAEITNSAFEPASMMAKCDPRHGKYMACCLMYRGDVVPKDVN
T.ba TEFQTNLVPYARIHFPICSYAPVISAEKAYHEQLTVAEVTNTLFEPANMMVKCDPRHGKYMACTLLYRGDVVPKDVS
M. 1 TEFQTNLVPYPRIHFPICSYAPVISAEKAYHEQLSVAEITNSLFEPANMMVKCDPRHGKYMACTLLYRGDVVPKDVG
M. 2 TEFQTNLVPYSRIHFPICSYAPVISAEKAYHEQLSVAEITNSLFEPANMMVKCDPRHGKYMACTLLYRGDVVPKDVG
E.he TEFQTNLVPYPRIHFPLVAYSPMLSKEKAAHEKLSVQEITNACFEPQSQMVRCDTRKGKYMACCLLFRGDVNPKDAN
S.lo TEFQTNLVPYPRIHFPLVAYFPMLSRERASHEQLSVQEITSACFDPENQMVKCDPRNGKYMACCLLFRGNVNPKDVN
N.lo TEFQTNLVPYPRIHFPLVAYAPMLSRNKASHEQLSVSEITSACFNPESQMVKCDPKKGKYMACCLLFRGDVQPKDVN

H.30 AAIAVIKTKRTIQFVDWCPTGFKVGINYQPPTVIPGGDLAKVQRAVLMISNSTAIAEVWSRTDHNFDLMYAKRAFVH
A.ro AAVATIKTKRTIQFVDWSPTGFKCGINYQPPIVVPGGDLAKIQRAVCMISNSTAIAEVFSRIDHKFDLMYAKRAFVH
T.ba AAIATIKTKRAIQFVDWCPTGFKIGINYQPPTVVPGGDLAKVQRAVCMLANTTAVAEAWSRLDHKFDLMYAKRAFVH
M. 1 AAVATIKTKRTIQFVDWCPTGFKIGINYQPPTVVPGGDLAKVQRAVCMLANTTAIAEAWSRLDHKFDLMYAKRAFVH
M. 2 AAVATIKTKRTIQFVDWCPTGFKIGINYQPPTVVPGGDLAKVQRAVCMLANTTAIAEAWSRLDHKFDLMYAKRAFVH
E.he TATANVKAKRTNQFVEWCPTGFKVGINSRKPTVLDGEAMAEVSRAVCALSNTTAISEAWKRLNNKFDLMFSKRAFVH
S.lo QATSLVKSKRANQFVEWCPTGFKIGINDRKPYVFEDGAMAPVDRAVCMLSNTTAISEAWKRLNRKFDLMFSKRAFVH
N.lo QAMAFVKAKRAAQFVEWCPTGFKIGMNSRKPTILDDDAMAPVSRAVCLLSNTTAIAEAWQRLNQKFDLMFSKRAFVH
```

Figure 4-1. Amino acid sequence of all alpha-tubulin genes reported in this section. In order, the taxa are: *Hexamita* 50330, *Acrasis rosea*, *Trichomitus batrachorum*, *Monocercomonas* sp. clone 1, *Monocercomonas* sp. clone 2, *Encephalitozoon hellem*, *Spraguea lophii*, and *Nosema locustae*.

noteworthy exceptions. The GTP-binding motif at positions 70 to 73 (numbered according to human) is generally LEPT in alpha-tubulins and LEPG in beta-tubulins, but in microsporidia, both alpha and beta-tubulin sequences contain LEPG. Also, the acetylatable lysine at position 40 of alpha-tubulins and the highly conserved region around it are both missing in microsporidia as they are in fungi, *Entamoeba histolytica* and *Dictyostelium discoideum* (Figure 4-2). It is not obvious why constraints on this otherwise highly conserved region have relaxed in these disparate taxa. One interesting correlation is that these organisms all lack flagella and cilia in all stages of their life cycle, although the same is true of some other organisms that have maintained the acetylation domain (for instance in plants, where some paralogues have not maintained the lysine residue). In any case, the role and importance of acetylation in tubulin function remains unclear, especially since it may be abolished without apparent consequence in *Chlamydomonas* and *Tetrahymena* (Kozminski *et al.*, 1993; Gaertig *et al.*, 1995), but is always observed when acetylatable alpha-tubulin is present in the cell.

Phylogeny Based on Alpha and Beta-Tubulins. Tubulin genes, for the most part beta-tubulin, have been used in the past to infer organismal relationships (Baldauf and Palmer, 1993; Edlind *et al.*, 1996), but the extreme conservation leaves few informative characters. Nevertheless, the utility of three alignable gene-families is attractive, and the substantial diversity of taxa previously known for beta-tubulin has now been roughly matched in the alpha-tubulin branch. From an amino acid alignment composed of 24 gamma, 42 beta and 81 alpha-tubulins, phylogenetic trees were inferred for each tubulin independently, and combined sets were used to reciprocally root one another.

To make the data more manageable, pairwise distance calculations were used to identify and eliminate closely related sequences. In this way the number of

Hexamita inflata	LFCLEHGIHHDGQ.....MPSD **K** SVGVSEDSFNTFFSETGAGKHVP
Spironucleus	LYCLEHGIHHDGQ.....MPSD **K** SIGVAEDSFNTFFSETGAGKHVP
Hexamita 50330	LFCLEHGIHQDGQ.....MPSD **K** SIGVAEDSFNTFFSETGAGKHVP
Drosophila	LYCLEHGIQPDGQ.....MPSD **K** TVGGGDDSFNTFFSETGAGKHVP
Human	LYCLEHGIQPDGQ.....MPSD **K** TIGGGDDSFNTFFSETGAGKHVP
Trichomitus	LYCLEHGIQPDGQ.....MPSD **K** TIGICDDAFNTFFSETGAGF 'P
*Monocercomonas*1	LYCLEHGIQPDGQ.....MPSD **K** TIGVCDDAFNTFFSETGAGI ..IP
*Monocercomonas*2	LYCLEHGIQPDGQ.....MPSD **K** TIGVCDDAFNTFFSETGAGKHVP
Physarum	LYCLEHGINPDGQ.....MPSD **K** SVGGGDDAFNTFFSETSSGKHVP
Naegleria	LYCLEHGIQPDGL.....MPSD **K** TIGVEDDAFNTFFSETGAGKHVP
Acrasis	LYCLEHGIQPDGQ.....MPSD **K** TIGVEDDAFNTFFSETGAGKHVP
Euplotes	LFCLEHGIQPDGQ.....MPSD **K** TIGGGDDAFNTFFSETGAGKHVP
Tetrahymena	LFCLEHGIQPDGQ.....MPSD **K** TIGGGDDAFNTFFSETGAGKHVP
Plasmodium	LFCLEHGIQPDGQ.....MPSD **K** ASRANDDAFNTFFSETGAGKHVP
Toxplasma	LFCLEHGIQPDGQ.....MPSD **K** TIGGGDDAFNTFFSETGAGKHVP
*Stylonychia*1	LFCLEHGIQPDGQ.....MPSD **K** TIGGGDDAFNTFFSETGAEKHVP
Euglena	LYCLEHGIQPDGS.....MPSD **K** AIGVEDDAFNTFFSETGAGKHVP
Leishmania	LFCLEHGIQPDGS.....MPSD **K** CIGVEDDAFNTFFSETGAGKHVP
Trypanosoma	LFCLEHGIQPDGA.....MPSD **K** TIGVEDDAFNTFFSETGAGKHVP
Chlamydomonas	LYCLEHGIQPDGQ.....MPSD **K** TIGGGDDAFNTFFSETGAGKHVP
Maize	LYCLEHGIQADGQ.....MPGD **K** TIGGGDDAFNTFFSETGAGKHVP
Arabidopsis	LYCLEHGIQPDGQ.....MPSD **K** TVGGGDDAFNTFFSETGAGKHVP
Anemia	LYCLEHGIQPDGQ.....MPSD **K** TVGGGDDAFNTFFSETGAGKHVP
Haemonchus	LYCLEHGIQPDGQ... ..MPSD **K** SLGGCDDSFSTFFSETGSGRHVP
Octopus	LYCLEHGIQPSGQ.....MPSD **K** AVGGKDDSFNTFFSETGSGKHVP
Schistosoma	LYCLEHGIQPDGQ.....MPSD **K** TIGGGDDSFNTFFSETGAGKHVP
Urechis	LYCLEHGIQPDGQ.....MPSD **K** TIGGGDDSFNTFFSETGAGKHVP
Rat	LYCLEHGIQPDGQ.....MPSD **K** TIGGGDDSFNTFFSETGAGKHVP
Entamoeba	LFCLEHGIQPDGTAIANSNEKRS V ITGGIDTAYNAFFQELQNGRHVP
Dictyostelium	LYCLEHGIERDGS.....IPAD R KQSSDNKDLGTFFSETNGKKVVP
*Emericella*1	LYCLEHGIQPDGY.....LTEE R KKEDPDHGFSTFFSETGQGKYVP
*Emericella*2	LYLLEHGLGADGR.....LDPK G EDINAGGSFETFFTETGCGKYVP
S.pombe 1	LYCLEHGIGPDGF..PTENSEVH K NNSYLNDGFGTFFSETGQGKFVP
*Saccharomyces*1	LYSLEHGIKPDGH.....LEDGL S KPKGGEEGFSTFFHETGYGKFVP
Pneumocystis	LYCLEHGIEPDGR.....LSPE K TTKPLDDGFSTFFSETGSGKYVP
*Schizophyllum*B	LYTLEHGLSPDGR.....LMDD S PSKH..DSGSTFFSETGQGKHVP
*Schizophyllum*A	LYTIEHGLSPDGR.....LSDD S PSKH.DDGFSTFFSETSSGKYVP
Neurospora A	LYCLEHGIQPDGY.....LTEE R KAADPDHGFSTFFSETGNGNTFP
*Histoplasma*1	LYCLEHGIQPDGY.....LTEE R KAADPDQGFNTFFSETGQGKYVP
*Histoplasma*2	TISGEHGVDGAGY.....YNGS L DIQL..ERMNVYFNEAAEKKYVP
S pombe 2	LYCLEHGIQPNGY.....MNPE T ASQNSDGGFSTFFSETGQGKYVP
Encephalitozoon	LYCKEHGILPDGR.....LDQN R MDD..ESAES.FFSQTSVGTYVP
Nosema	LYCKEHNIRPDGT.....TGGV . .....DDSCSSFFIETSAGTYVP
Spraguea	LYCKEHGILPDGT.....PDPN F NDK..ESYSSTFFSETSGGNFVP

Figure 4-2: Acetylation domain of 45 alpha-tubulin genes. Sequences correspond to 23-63 of the human sequence, and acetylation takes place at lysine 40 (singled out and in bold).

sequences was reduced to 58 and 40 for alpha and beta respectively. Unfortunately, this number is still prohibitively large for protein maximum likelihood analysis, and maximum parsimony analysis was also hampered by the impractically large number of equally parsimonious trees. However, it should be noted that the strict consensus of over 700 maximum parsimony trees of alpha-tubulin yielded a topology consistent with neighbor-joining trees. Trees were therefore constructed by neighbor-joining analysis of corrected distance measurements calculated according to the Dayhoff PAM250 substitution matrix. Significance of individual nodes on these trees was assessed by conducting 100 bootstrap resampling replicates, the results of which are also shown on each tree.

An alpha-tubulin tree is depicted in Figure 4-3. This tree is based on 406 positions, includes 58 sequences, and has been oriented with a diplomonad outgroup (diplomonads were chosen because they are consistently deep-branching eukaryotes in trees based on ribosomal RNA and EF-1α: Leipe *et al.*, 1993; Hashimoto *et al.*, 1994). Figure 4-4 is a beta-tubulin tree consisting of 431 positions from 40 sequences, and once again has a diplomonad outgroup. These trees share a number of features with other molecular phylogenies, including the presence of several monophyletic groupings such as animals, plants, fungi, and alveolates. It is also noteworthy that the alpha-tubulins of *Acrasis rosea* and *Naegleria gruberi* branch together, since these taxa are thought to belong to the phylum Heterolobosea (Page and Blanton, 1985), for which supporting molecular data has just been introduced (Roger *et al.*, 1996).

While these groups may be consistent with other data, alpha and beta-tubulin trees also mirror one another in several ways that are not generally supported by other data. Such anomalies might be overlooked as artifacts or the results of inappropriate data for the question, but being shared by both trees, these discrepancies do require some auxiliary explanation.

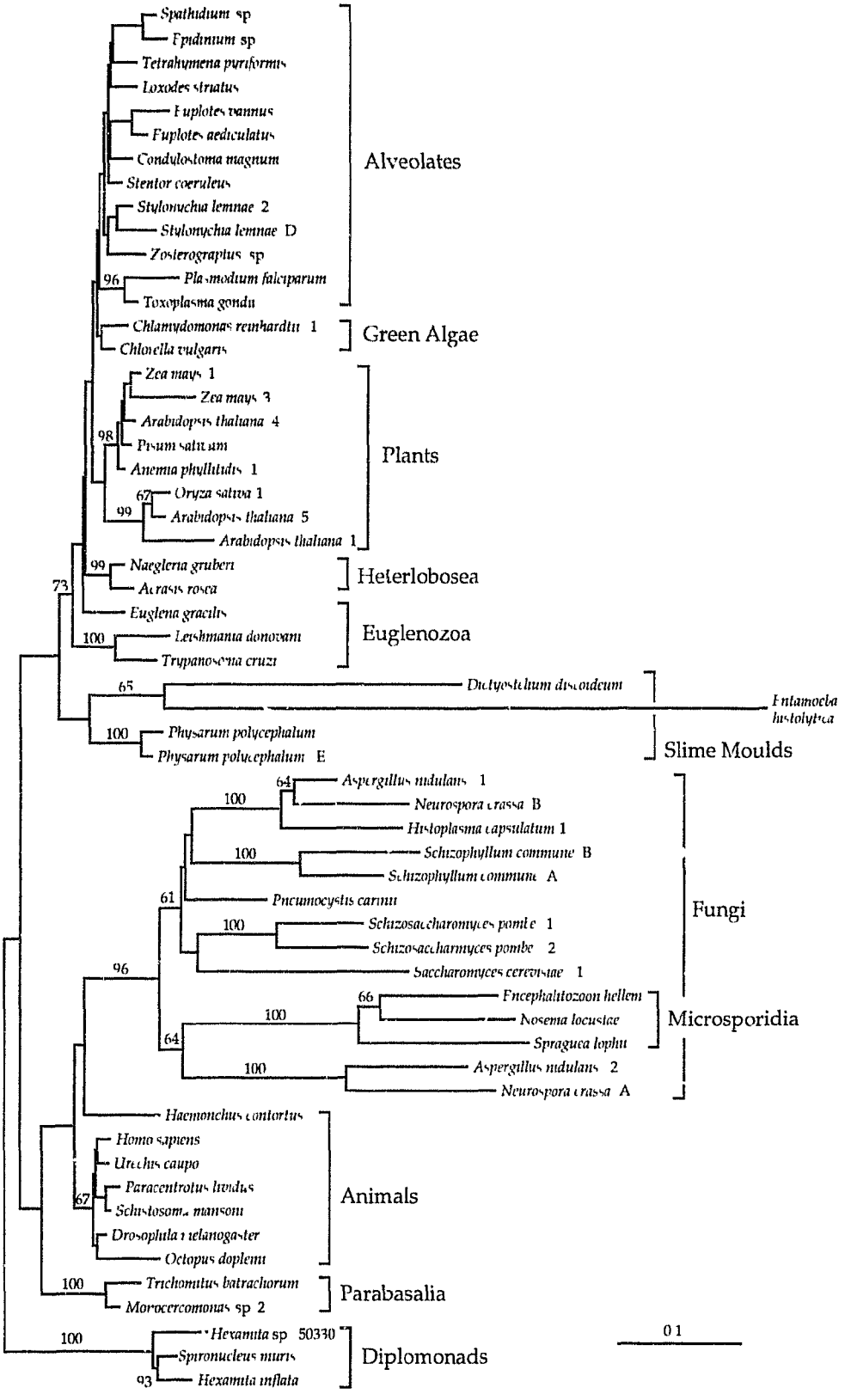


Figure 4-3. Neighbor-joining tree of alpha-tubulin.

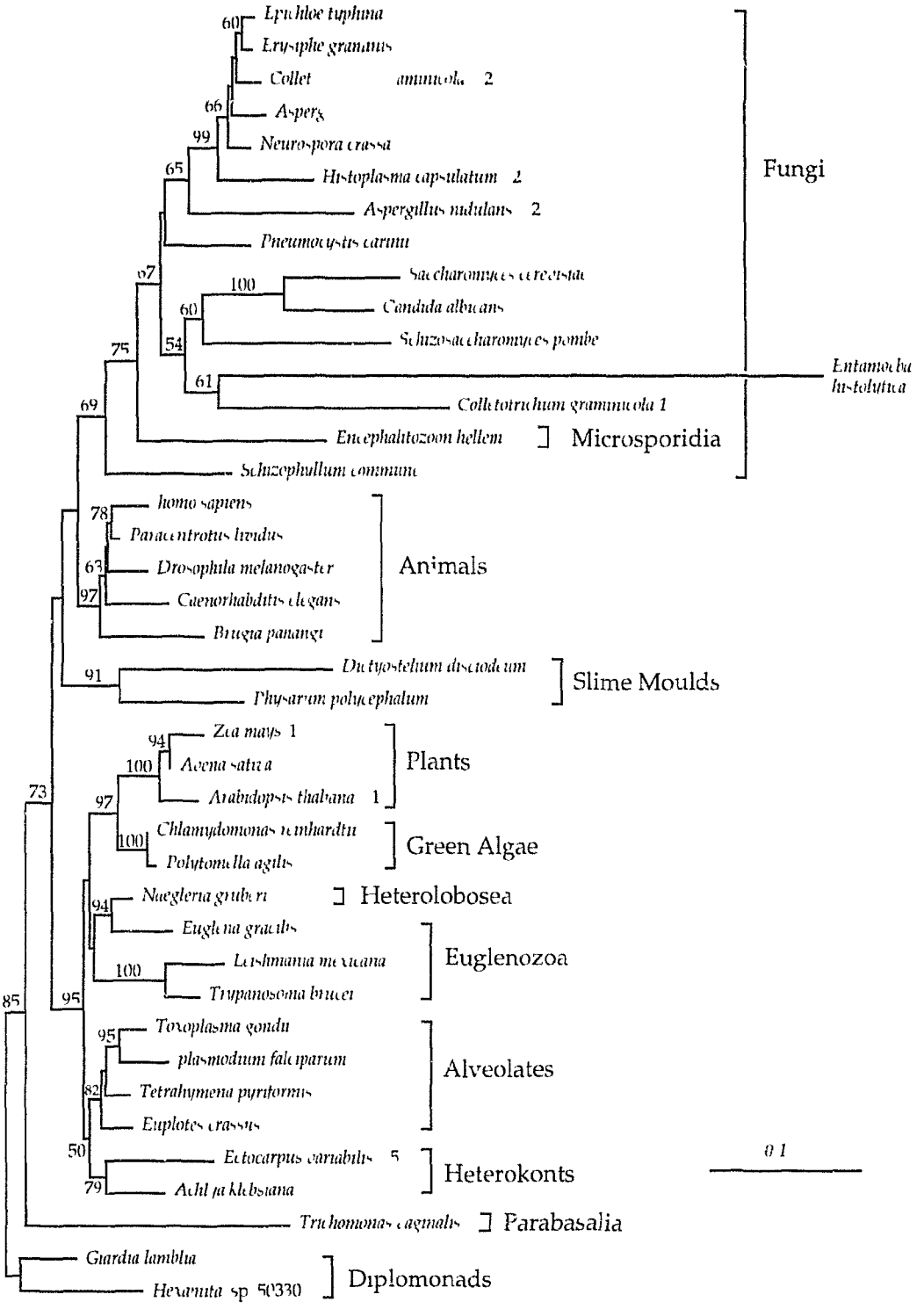


Figure 4-4. Neighbor-joining tree of beta-tubulin.

The first such characteristic is the position of the animals and fungi relative to parabasalia and diplomonads. When the diplomonads are used as an outgroup in these unrooted trees, the results is a deep split in the eukaryotes where animals and fungi fall on one side, and the plants, euglenozoa, alveolates and heterolobosea on the other (slime moulds and parabasalia cannot readily be classified into either category as they both branch close to diplomonads and their exact position is inconsistent). This topology, also found in beta-tubulin phylogeny by Edlind *et al.* (1996), is not a feature of ribosomal RNA or EF-1α phylogenies, in which diplomonads fall at or near the base of a comb-like distribution of taxa (Cavalier-Smith, 1993; Leipe *et al.*, 1993; Hashimoto *et al.*, 1994).

A second noteworthy characteristic of both trees is the position of microsporidia within the fungi. Microsporidia are generally thought to be archezoa, partly because they lack several cytological features also missing in other archezoa (see Cavalier-Smith, 1993), and partly because they normally branch very deeply in eukaryotic trees of ribosomal RNA or translation elongation factors (Vossbrinck *et al.*, 1987; Kamaishi *et al.*, 1996). Considering that fungi and microsporidia share a highly divergent acetylation domain in alpha-tubulin, these residues were excluded and the analysis repeated. The resulting topology was no different than that of Figure 4-3 (data not shown), suggesting that microsporidian and fungal alpha-tubulins do generally resemble one another outside the acetylation domain.

One last concern with these tree topologies is the position of *Entamoeba histolytica* and its alarmingly long branch. *Entamoeba* tubulins, although easily classifiable by family, are extremely divergent from other orthologues, resulting in a very long branch. The position of *Entamoeba* in these trees is therefore suspect, since it branches with the next longest branch in both alpha and beta-tubulin trees. This conclusion seems to be borne out by removing *Dictyostelium* from alpha-tubulin trees, which results in no change to the topology except that *Entamoeba*

moves to the next longest branch, at the base of the fungi (data not shown). In contrast, the removal of *Entamoeba* resulted in no change at all to the rest of the tree (data not shown). The affinity of *Entamoeba* appears to be for long branches.

Gamma, Delta, and Epsilon-Tubulins In addition to the well represented alpha and beta-tubulins, there are the more poorly represented gamma-tubulins, and two highly divergent tubulin-like sequences from *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. Based on their extreme distance from other tubulins, these two genes have prompted a proposal to expand the number of tubulin families from three to five, classifying *C. elegans* and *S. cerevisiae* sequences as delta and epsilon-tubulin respectively (Burns, 1995).

Distance notwithstanding, there are a number of facts that support the contrasting notion that these delta and epsilon-tubulins are not really novel families, but rather highly divergent orthologues of an existing family that are unique to the lineages where they have been observed. First, the now completed *S. cerevisiae* genome does not contain either a conventional gamma-tubulin gene or a so-called delta-tubulin gene, but only highly conserved alpha and beta tubulins and the so-called epsilon-tubulin. Similarly, searching the expressed sequence tag (EST) database for gamma, delta and epsilon-tubulins in *C. elegans* and *C. briggsae* yielded only alpha, beta, and the so-called delta-tubulin. The implication from these observations is that neither *S. cerevisiae* nor *C. elegans* contain either a conventional gamma-tubulin, or the supposed novel tubulin found in the other. Indeed, no other organism has ever been found to contain either of these genes except *C. briggsae,* which contains an EST almost identical to the *C. elegans* delta-tubulin.

Greater support for the gamma-tubulin provenance of these unusual sequences comes from phylogenetic reconstruction of all gamma-tubulins with the

delta and epsilon genes and outgroups chosen from the alpha and beta-tubulins. This tree (Figure 4-5) reveals that the unusual *Saccharomyces* and *Caenorhabditis* sequences branch with a high affinity to the gamma-tubulin lineage to the exclusion of either alpha or beta-tubulin. The basal position of *S. cerevisiae* and *C. elegans* is likely the result of an attraction to the other long-branches on the tree: those leading to *Reticulomyxa*, *Entamoeba* and *Plasmodium*. In analyses excluding these long-branches, or by simply excluding *Entamoeba* (data not shown), *S. cerevisiae* branches specifically with the fungi with high statistical support, and *C. elegans* branches specifically with the animals, although with much weaker support (Figure 4-6).

Lastly, and perhaps most conclusively, recent functional characterisation of the *Saccharomyces* gene product has provided excellent evidence that it is located at the spindle pole body (a MTOC), and that its disruption results in a phenotype similar to gamma-tubulin disruptions in other ascomycetes (Sobel and Snyder, 1995). Taken together, these observations leave little room to doubt the conclusion that both *Saccharomyces* and *Caenorhabditis* tubulin-like genes are lineage-specific, highly divergent orthologues of gamma-tubulin.

Rooting Tubulin Trees. Figure 4-7 shows the result of combining subsets of the alpha, beta, and gamma-tubulin alignments (chosen for representative diversity, but excluding the extremely diverse gamma-tubulins from *Saccharomyces* and *Caenorhabditis* discussed above). This tree is based on 310 positions and 70 sequences. Clearly the three families are each independent, monophyletic groups, and it is not obvious if any two are more similar than the third, reflecting the great inter-family distance relative to intra-family distances. The actual topologies within the alpha and beta subtrees differ only very subtly from the topologies yielded by individual analyses, but in both cases trees appear different on account of the root

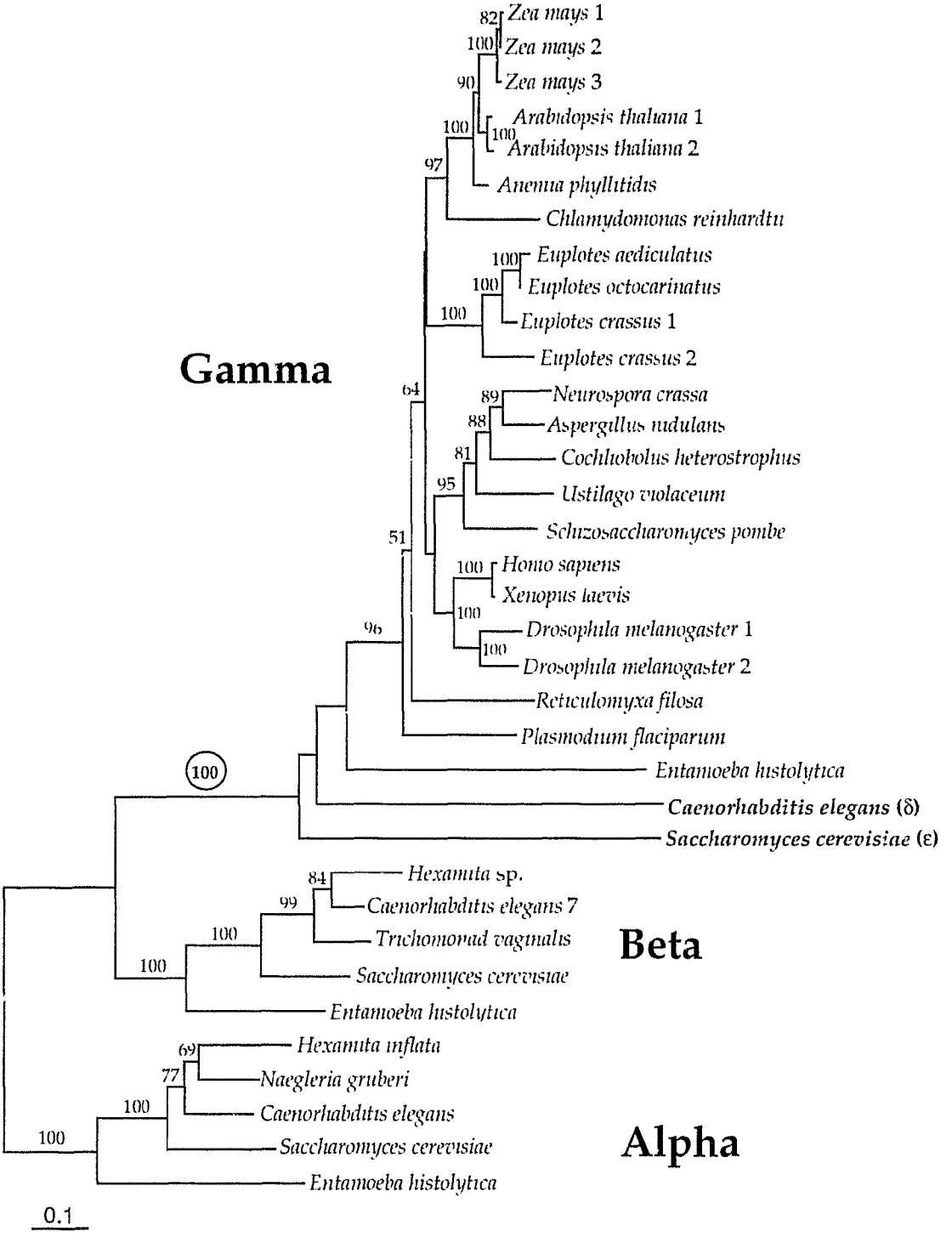


Figure 4-5. Neighbor-joining tree of gamma-tubulins, *Caenorhabditis elegans* "delta-tubulin" and *Saccharomyces cerevisiae* "epsilon-tubulin", all rooted with alpha and beta-tubulin sequences. The strong phylogenetic affinity of the so-called delta and epsilon-tubulins to gamma-tubulins is evident in the long branch uniting these sequences and the high support for that branch (circled).

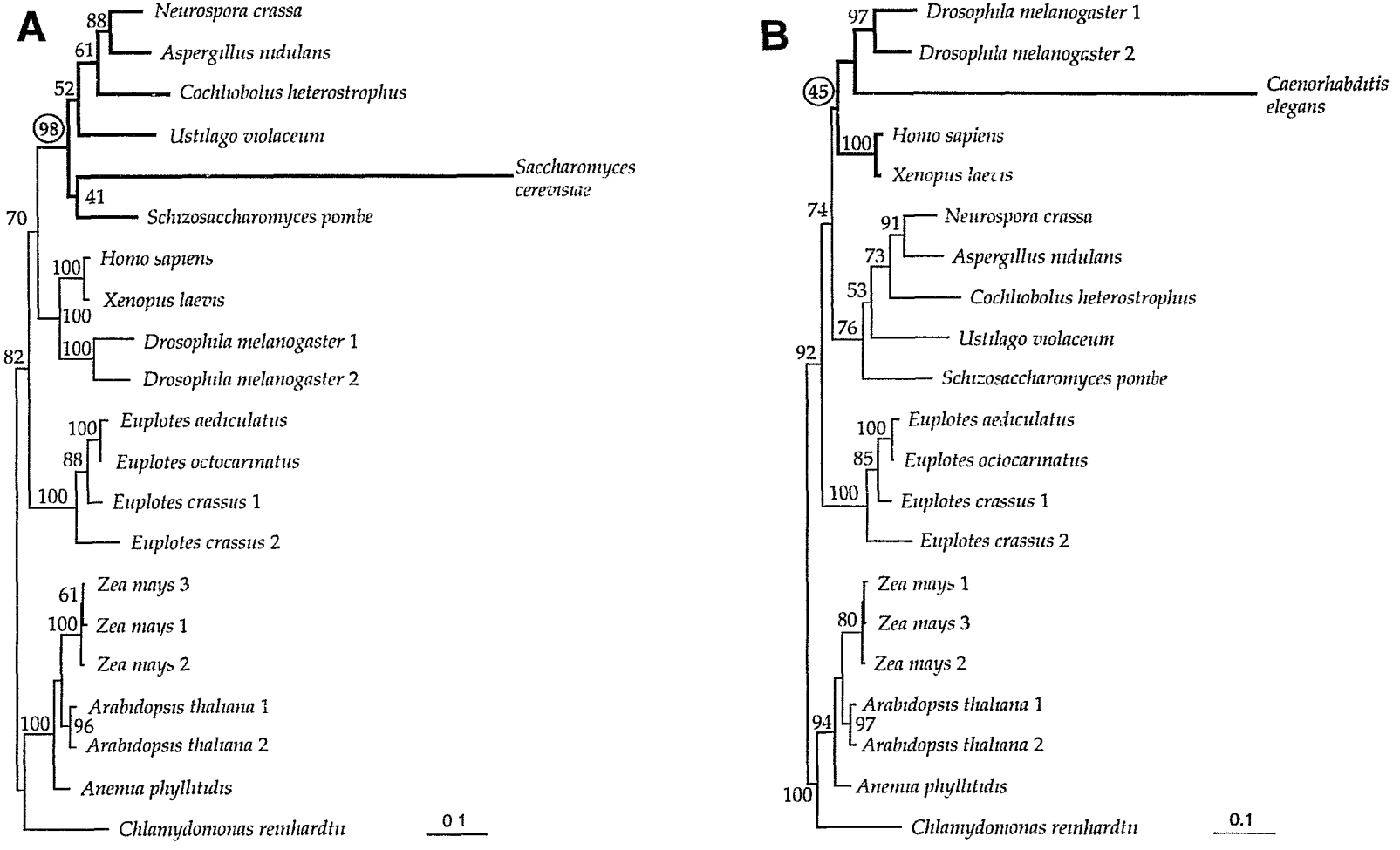


Figure 4-6. Neighbor-joining trees of gamma-tubulin showing the specific affinity of *Saccharomyces cerevisise* to other fungi (A) and *Caenorhabditis elegans* to other animals (B). Percent of bootstrap replicates supporting each node is also shown, those which support the position of *S. cerevisiae* and *C. elegans* within their respective taxonomic groups are circled.

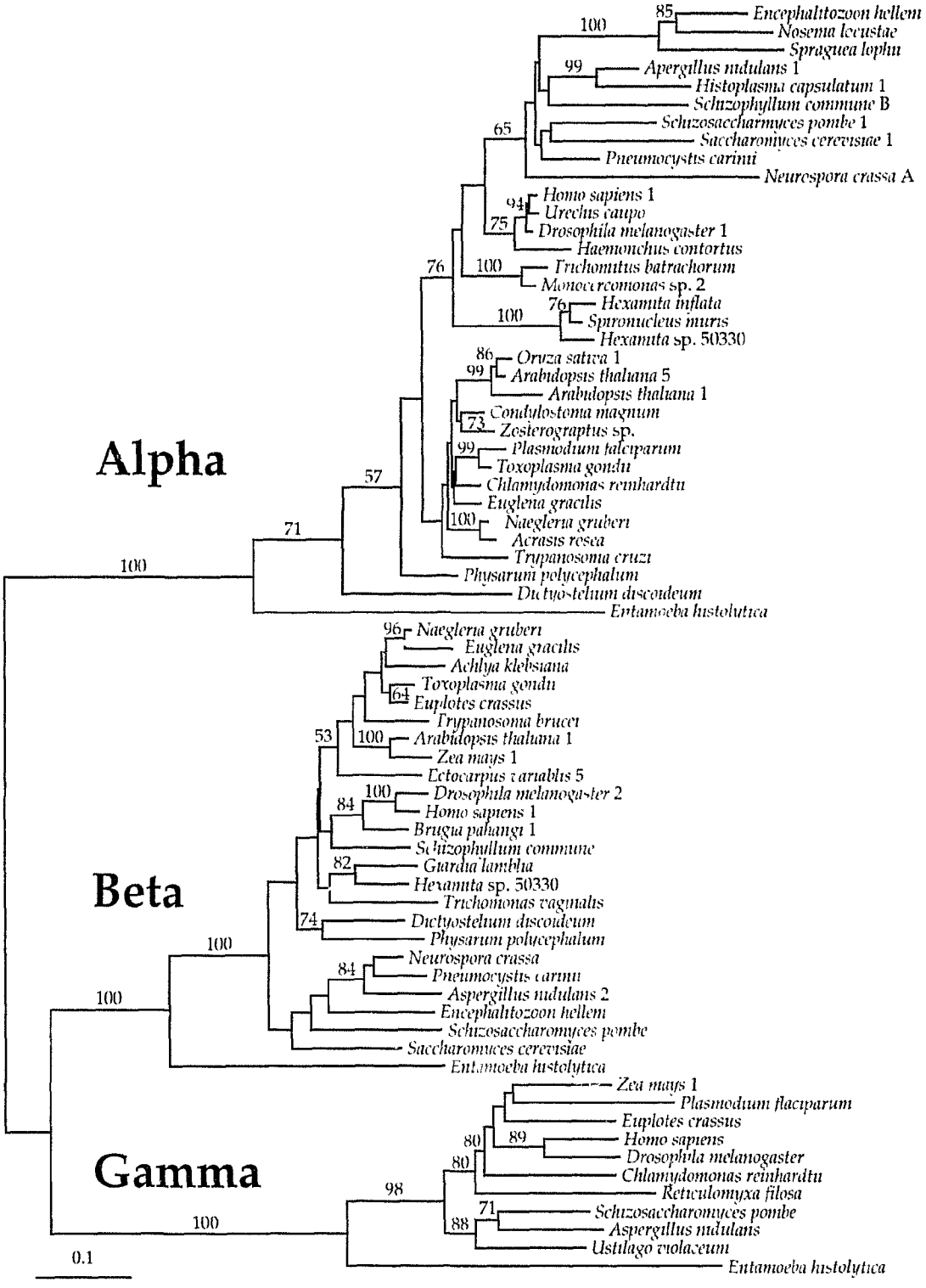


Figure 4-7. Neighbor-joining tree of alpha, beta and gamma-tubulins.

falling within the branch leading to *Entamoeba histolytica*. As discussed above, *Entamoeba* tubulins are quite divergent, so it is likely that this relationship is not legitimate, but is the product of a long branch attraction between *E. histolytica* and the branch leading to the other tubulin families. Once again, *E. histolytica* alpha and beta-tubulins were removed and the three way rooting repeated. In this tree the topology does not change within each subtree, but the position of the root does change, moving in both cases to the next longest branch in the fungi (data not shown). Therefore, the position of the root within each sub-tree appears to be largely dependent upon branch length and should therefore be considered highly suspect.

Discussion

Phylogenetic analysis of alpha-tubulin genes from diplomonads, microsporidia, parabasalia, and heterolobosea provides convincing evidence that alpha and beta-tubulins diverged prior to the divergence of extant eukaryotes. Moreover, since each of the three sub-trees is holophyletic, it appears that gamma-tubulin was also present by this time. If this were not the case, then barring any paralogue-specific rate acceleration, one would expect that the gamma subtree would branch from within one of the other two, and not from the basal position seen in Figure 4-7.

These new sequences also provide the opportunity to compare alpha and beta-tubulin phylogenies with an almost equal representation of major lineages. The alpha and beta-tubulin trees are nearly identical in topology, but in several major respects this topology is curiously inconsistent with the phylogeny inferred by other molecular markers. This is most evident in the relative branching order of the unrooted trees of Figures 4-3 and 4-4, where the animals and fungi branch closer to supposedly deep-branching protists such as diplomonads and parabasalia than they do conventionally (Cavalier-Smith, 1993; Leipe *et al.*, 1993; Kamaishi *et al.*, 1993;

Gunderson *et al.*, 1995). Another unexpected relationship mirrored by both molecules is the firm position of the microsporidia within the fungi. Conventionally, Microsporidia is seen as an ancient phylum, in part because of their very degenerate cytology, lacking numerous features also missing in other Archezoa, and in part from the consistently deep position of microsporidia in the few molecular trees where microsporidian data is available (Vossbrinck *et al.*, 1987; Kamaishi *et al.*, 1996).

The congruent alpha and beta-tubulin trees might be taken as independent support for these unusual results. However, since microtubules are composed of alternating units of closely packed alpha and beta-tubulin, these trees may instead reflect a strong tendency to co-variation between the two tubulin molecules. If there were a finite number of 'solutions" to the problem of satisfactory interactions between the proteins (which is supported by the extreme degree of both inter- and intra-family conservation in tubulins), then the appearance of a certain variant of one tubulin could strongly favor the co-variation of the other along predictable lines. Paralogy and loss could also be involved in such a process, but in such a case once again co-variance would have to be evoked to explain the congruent loss of paralogues in several supposedly unrelated lineages.

Even if the congruence between tubulin trees is not independent support for this topology, there are truly independent reasons to carefully consider the phylogenetic position of the Microsporidia within the fungi. The extremely derived, obligately parasitic lifestyle of these organisms has raised doubts as to whether their "primitive" cytology is ancestral or a relatively recent adaptation (Cavalier-Smith, 1993). Similarly, since microsporidian gene sequences are typically very divergent, the deep phylogenetic position of these sequences may be a consequence of attraction to other long branches. Alpha and beta-tubulins are not immune to this possibility either, and their position within the fungi may simply be due to the fact

that both microsporidian and fungal tubulins are diverging faster than other lineages, but there is other circumstantial evidence for a relationship between microsporidia and fungi. First, the ridged endospore wall of microsporidia is composed of chitin (see Canning, 1990), the same material that fungi utilise as a cell wall polymer (chitin is also found in numerous other unrelated lineages: Mulisch, 1993). Secondly, unlike other archezoa, which appear to be clonal (see Tabayrenc *et al.,* 1991 for review), microsporidia undergo a form of meiosis. This process is in itself a source of debate, however, and there are alternative arguments that it is either radically different from meiosis in other eukaryotes (Canning, 1988), or fundamentally the same process (Flegel and Pasharawipas, 1995). Curiously the argument that microsporidian meiosis is typically eukaryotic in form is based on similarities observed by the authors between the cell-cycle of microsporidia and fungi (Flegel and Pasharawipas, 1995). Lastly, although the molecular phylogeny of EF-1α supports the deep divergence of microsporidia (Kamaishi *et al.,* 1996), the only microsporidian EF 1α gene known to date, that of *Glugea plecoglossi*, contains an eleven-codon insertion at exactly the same position as a twelve codon insertion that has been argued to be a determinative feature of the animal-fungal clade (Baldauf and Palmer, 1993). An insertion in this general region of the protein may be a common event, but the *Glugea* insertion is at exactly the same location and also bears a weak resemblance to that of animals and fungi (see Figure 2 of Kamaishi *et al.,* 1996).

Individually each of these characters may be inadequate to argue strongly for any relationship between microsporidia and fungi, as each is also shared with other taxa. However, taken together, and considering the relatively strong support for the microsporidia-fungi clade in both alpha and beta-tubulin trees, the possibility that microsporidia are highly derived fungi certainly should be considered.

Chapter V:

The Genetic Code in Diplomonads

Introduction

Among the taxa for which alpha-tubulin genes were sequenced were two *Hexamita* strains (ATCC 50330 and 50380), blood-borne and muscle parasites of Pacific and Atlantic salmon respectively. Surprisingly these genes were found to contain numerous in-frame termination (TAA and TAG) codons. These were shown to be sense codons in these genomes by identifying cognate tRNA genes, and a survey of tRNA genes throughout several diplomonads revealed that another species, *Hexamita inflata*, also likely uses TAA and TAG (TAR) glutamine codons.

While almost all known genomes employ the same ancestral genetic code, variant codes have been identified in one bacterial genome, three eukaryotic nuclear lineages, and in mitochondria. One theory that provides a plausible route for the evolution of such variants from the universal code is codon capture. This model, developed predominantly by Osawa and Jukes (reviewed in Osawa *et al.*, 1992), has the important advantage that it avoids selectively disadvantaged transition stages through a series of neutral steps. First, either by mutation pressure or chance, certain codons disappear altogether from all genes in a genome. Once this occurs, the genes encoding the tRNAs or release factor previously required to read these missing codons are superfluous, and may be inactivated or lost. The codon is now "unassigned" but may reappear in the genome if a new tRNA that can recognise it fortuitously arises (for instance if a duplicate of a functional tRNA gene acquires an anti-codon mutation). Such a tRNA gene suppresses the lethal effects of chance or pressure-driven mutations that reintroduce the missing triplet, thus "capturing" the codon and establishing a new code.

The particular variation of the genetic code observed in *Hexamita* has previously been observed only in very AT-rich nuclear genomes where it is thought to have been favored by the same directional mutation pressure that biased the genome's composition (Osawa and Jukes, 1989). There is no evidence of such AT pressure on the GC-rich genomes of these diplomonads. However, even in the absence of directional mutation pressure, mutations converting glutamine to amber or ochre termination codons are expected to occur with a higher than average frequency because they involve only a single transition each. The lesson of these diplomonad genomes appears to be that there is no unique force required to change the genetic code, but that any mutation occurring at a sufficiently high frequency has the potential to motivate a codon capture event.

Results

The unusual tubulin gene from *Hexamita* (ATCC 50330). The sequence of the alpha-tubulin gene from *Hexamita* 50330 revealed an interesting feature of this organism that deserves special attention. Amino acids 59 and 223 of the *Hexamita* 50330 sequence in Figure 4-2 are shown as glutamine (Q) but the actual sequence of the genes at these positions did not contain the universal glutamine codons (CAR) but instead had amber (TAA) termination codons (Figure 5-1). That these codons really existed in the genome (and were not erroneously incorporated by Taq polymerase) was confirmed by re-amplifying the gene, re-cloning, and sequencing one of these independant clones to show that it had exactly the same sequence (throughout the gene) as the original two clones.

Such an observation could be attributed to a variety of artifacts or biological oddities, including pseudogenes, RNA editing, or a non-canoncal genetic code. To examine the first of these possibilities, the genes for elongation factor-1 alpha (EF-1α) and beta-tubulin were amplified, cloned, and sequenced (using the primers

Alpha-Tubulin	59	223
***Hexamita* 50330**	IYHPE * LISGK	DLTEF * TNLVP
Drosophila	LFHPE Q LITGK	DLTEF Q TNLVP
Homo	LFHPE Q LITGK	DLTEF Q TNLVP
Trichmitus	LWHPE Q LINGK	DFTEF Q TNLVP
Plasmodium	LFHPE Q LISGK	DVTEF Q TNLVP
Naegleria	LFHPE Q LITGK	DVTEF Q TNLVP
Acrasis	LFHPE Q LISGK	DVTEF Q TNLVP
Chlamydomonas	LFHPE Q LISGK	DITEF Q TNLVP
Tetrahymena	LFHPE Q LISGK	DITEF Q TNLVP
Trypasoma	LFHPE Q LISGK	DLTEF Q TNLVP
Arabidopsis	LFHPE Q LISGK	DITEF Q TNLVP
Saccharomyces	LFHPE Q LLSGK	DLNEF Q TNLVP
Nosema	LYHPG Q LISGK	DLTEF Q TNLVP

Figure 5-1. Aligned amino acid sequence of alpha-tubulins surrounding TAG termination codons at positions 59 and 223 of the *Hexamita* 50330 gene fragment.

BTUB-A and BTUB-B for beta-tubulin and EF1F and EF7R for the 5' end, and SEF3 and EF8R for the 3' end of EF-1α), whereupon it was found that each contained TAA or TAG termination codons, all at several positions where glutamine is found conserved among diverse homologues (Figure 5-2). To further extend this observation, the alpha-tubulin gene was also isolated and sequenced from another *Hexamita* strain (ATCC 50380), parasitic in Atlantic salmon. This second tubulin gene proved to be very similar to the first, differing at only three out of 1153 positions. Interestingly, two of these substitutions are synonymous transitions, while the third interconverts glutamine and amber, also by a transition (Figure 5-3). The explanation that satifies all these observations is that these two organisms share the same variant genetic code, and this last substitution is also synonymous.

TAR codons also specify glutamine in *Acetabularia* and certain ciliates, where the conclusion that these triplets (in addition to CAG and CAA) code for glutamine has been confirmed by comparisons of gene and protein sequences (Schneider *et al.*, 1989) and most convincingly by the finding of $tRNA^{Gln}$ species with UUA and CUA anticodons in *Tetrahymena thermophila* (Hanyu *et al.*, 1986). Similar confirmation that the amino acid sequences shown here bespeak a similar variant code was sought by searching these *Hexamita* genomes for genes that encode novel tRNAs able to decode UAG and UAA. Using the primers Q-F and Q-R, such genes should generate PCR products of 83 bp, with the anticodon located 13 bp from the terminus of the 5' primer, and should be foldable into cloverleaf tRNA structures.

Indeed, products matching these criteria were readily obtained, and sequencing indicated that they fell into three distinct groups, two of $tRNA^{Gln}$-like genes with the anticodons CTA and TTA, and one $tRNA^{Gly}$-like gene with the anticodon GCC. When aligned to a sample of all types of tRNAs from diverse organisms, the nearest relatives to the putative $tRNA^{Gln}$ fragments were glutamine

A.

EF-1α

```
G.la      STLTGHLIYKCGGILQRTIDEYEKRATEMGKGSFKYAWVLDQLKDERERGITINIALWKFETKKYIV
H.30  NGKSTLTGHLIYKCGGID*RTLDEYEKRANEMGKGSFKYAWVLDQLKDERERGITINIALWKFETKKFTV

G.la  TIIDAPGHRDFIKNMITGTSQADVAILVVAAGQGEFEAGISKDGQTREHATLANTLGIKTMIICVNKMDD
H.30  TIIDAPGHRDFIKNMITGTSQADVAILVIASGQGEFEAGISKEGQTREHATLAHTLGIKTLIVCVNKMDD

G.la  GQVKYSKERYDEIKGEMMKQLKNIGWKKAEEFDYIPTSGWTGDNIMEKSDKMPWYEGPCLIDAIDGLKAP
H.30  PQVNYSEARYKEIKEEMQKNLKQIGYKKWDEFDFIPTSGWTGDSIMEKSPNMPWYSGPCLIDAIDGLKAP

G.la  KRPTDKPLRLPIQDVYKISGVGTVPAGRVETGELAPGMKVVFAPTSQVSEVKSVEMHHEELKKAGPGDNV
H.30  KRPTDKPLRLPIQDVYKINGVGTVPAGRVESGLLIPNMTVVFAPSTTTAEVKSVEMHHEELPQAGPGDNV

G.la  GFNVRGLAVKDLKKGYVVGDVTNDPPVGCKSFTAQVIVMNHPKKIQPGYTPVIDCHTAHIACQFQLFLQK
H.30  GFNVRGIAAKDIKKGYVVGDTKNDPPVGCKSFTAQVIIMNHPKKIQPGYSPVIDCHTAHIACKFDAFLQK

G.la  LDKRTLKPEMENPPDAGRGDCIIVKMVPQKPLCCETFNDYAPLGPFAVR
H.30  LNARTLKPEMENPTEASRGECIVVRMVPSKPLSCESFNDYAALGRFAVR
```

Beta Tubulin

```
G.la  MREIVHIQAGQCGNQIGAKFWEVISDEHGVDPSGEYRGDSELQIERINVYFNEAAGGRYVPRAILVDLEP
H.30                IGAKFWEVISDEHGIDPSGEYRGDSELQIERVNVYYNEATGGRYVPRAVLVDLEP

G.la  GTMDSVRAGPFGQIFRPDNFVFGQSGAGNNWAKGHYTEGAELVDAVLDVVRKESEACDCLQGFQICHSLG
H.30  GTMDSVRAGPFGQLFRPDNFVFGQSGAGNNWAKGHYTEGAELVDAVLDTVRKEAEACDCLQGFQLVHSLG

G.la  GGTGAGMGTLLTAKIREEYPDRMMCTFSVVPSPKVSDTVVEPYNATLSVHQLVEHADEVFCIDNEALYDI
H.30  GGTGSGMGTLLMAKIREEYPDRMMCPFSIVPSPKVSDTVVEPYNATLSVHQLVENADEVFCIDNEALYDI

G.la  CFRTLKLTCPTYGDLNHLVSLVMSGCTSCLRFPGQLNADLRKLAVNLIPFPRLHFFLVGFAPLTSRGSQI
H.30  CFRTLKLTCPTYGDLNHLVSLVMSGITCCLRFPGQLNADLRKIAVNLVPFPRLHFFVAGFAPLTSRGSQI

G.la  YRALTVPELVSQMFDNKNMMAASDPRHGRYLTAAAMFRGRMSTKEVDEQMLNIQNKNSSYFVEWIPNNMK
H.30  YRALTVPELFSQMFDNKNMMAASDPRHGRYLTCLTLIRGRVSTKEVDEQDHNIQNKNSSYFVEWIPRNIM

G.la  VSVCDIPPRGLKMAATFIGNSTCIQELFKRVGEQFSAMFRRKAFLHWYTGEGMDEMEFTEAESNMNDLVS
H.30  VGICDIPPRGLKMSGTFIGNTTAI*ELFKRVGEQFTAMFRRKAFLHWYTGEG
```

B.

Beta-Tubulin	310	EF-1α	19
***Hexamita* 50330**	NTTAI * ELFKR	***Hexamita* 50330**	CGGID * RTLDE
Giardia	NSTCI Q ELFKR	***Giardia***	CGGID Q RTIDE
Drosophila	NSTAI Q ELFKR	***Euglena***	CGGID K RTIEK
Homo	NSTAI Q ELFKR	***Tetrahymena***	CGGID K RVIEK
Saccharomyces	NSTSI Q ELFKR	***Homo***	CGGID K RTIEK
Trypanosoma	NNTCI Q EMFRR	***Mucor***	CGGID K RTIEE
Arabidopsis	NSTSI Q EMFRR	***Saccharomyces***	CGGID K RTIEK
Chlamydomonas	NSTAI Q EMFKR	***Arabidopsis***	LGGID K RVIER
Tetrahymena	NSTAI Q EMFKR	***Entamoeba***	CGGID Q RTIEK
Euglena	NNTAI Q EMFKR	***Staphlyococcus***	LGLVD Q KTIQM

Figure 5-2. (A) Amino acid sequence of EF-1α and beta-tubulin from *Hexamita* 50330 aligned with those of *G. lamblia*. (B) Aligned amino acid sequence surrounding termination codons found at position 310 and 19 of beta tubulin and EF-1α respectivly in *Hexamita* 50330. The terminator in beta-tubulin is TAG while EF-1α contains a TAA.

```
      1/1                                                    38/13
         L   F   C   L   E   H   G   I   H   Q   D   G   Q   M   P   S
50330 A CTT TTC TGC CTT GAA CAC GGT ATC CAC CAG GAC GGC CAG ATG CCT TCT
50380 A CTC TTC TGC CTT GAA CAC GGT ATC CAC CAG GAC GGC TAG ATG CCT TCT
         L   F   C   L   E   H   G   I   H   Q   D   G   *   M   P   S
```

Figure 5-3. First 49 nucleotides and inferred amino acids of alpha-tubulin genes from *Hexamita* strains 50330 and 50380. The C-T mismatch at position 38 results in a stop codon in 50380 and a glutamine codon in 50330. Another C-T transition can be seen at postion 4, this is a silent third position substitution.

tRNAs from other eukaryotes. The primary and predicted secondary structures show many other tRNA-like features including the presence of an invariant U33 residue that is necessary to allow wobbic pairing at position 34, a pyrimidine at positi:n 32 and a purine at 37, as well as scattered conserved nucleotides and an anticodon stem, all spaced exactly as expected of a eukaryotic $tRNA^{Gln}$ (Figure 5-4).

The genetic code of other diplomonads. To see whether this curious trait is restricted to the closely related *Hexamita* strains, genes for α-tubulin and EF-1α were amplified and sequenced from *Hexamita inflata* ar 1 *Spironucleus muris*. These are also shown in Figure 5-5, where it can be seen that no termination codons were observed in 21 glutamine codons of *H. inflata* or 16 glutamine codons of *S. muris*. However, the tRNA genes from these diplomonads are more revealing.

Even a slight decrease in the frequency of TAR use observed in the *Hexamita* strains (1 out of 10 glutamine codons for *Hexamita* 50330) could render them difficult to detect in protein-coding sequences. In *H. inflata*, for instance, the frequency of TAR glutamine codons could be as high as 1 out of 7, and it would still not be unlikely that none was observed in the 21 glutamine codons encountered (based on a Poisson distribution of hits). Transfer RNA genes were therefore amplified from *S. muris, H. inflata* and also the human parasite *G. lamblia,* for which there is enough molecular data (including data on termination codons used at the ends of open reading frames) to conclude that it uses the universal code.

The amplification products that appear to correspond to tRNA genes are shown in Figure 5-6. *G. lamblia* yielded only a single $tRNA^{Gln}_{UUG}$ as well as a $tRNA^{Pro}_{UGG}$, but no non-canonical tRNAs. *S. muris* yielded only a single unambiguous product, corresponding to $tRNA^{Gln}_{CUG}$. The most interesting results

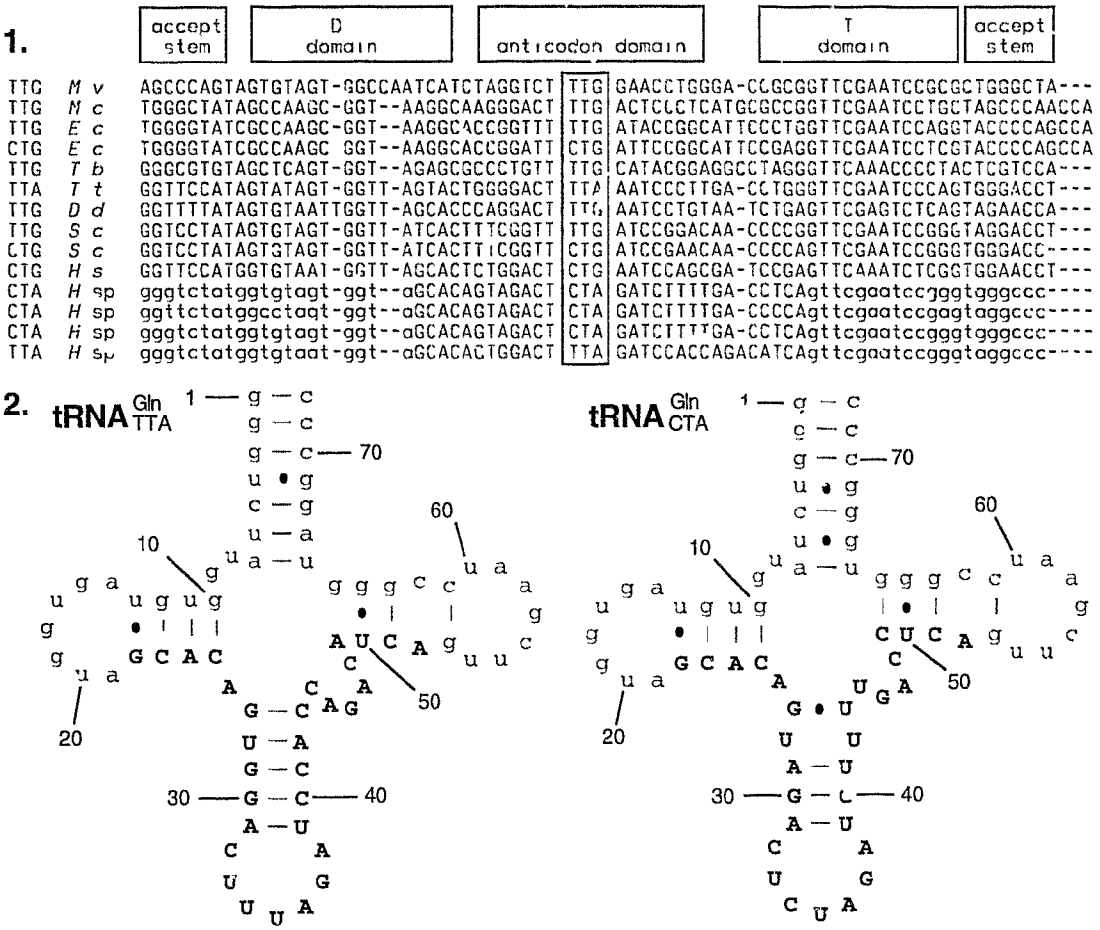


Figure 5-4. Primary and secondary structure model of putative tRNAGln species from *Hexamita* (1) Nucleotide sequences from selected tRNAGln genes aligned with amplification products from *Hexamita* 50330. Primer sequences are shown in lower case, domains indicated above the sequences, and the anticodon distinguished by a box. (2) Proposed cloverleaf structures for tRNA^{Gln} UUA and tRNA^{Gln} CUA numbered according to the scheme of Sprinzl *et al.* (1989). Primer sequences are in lower case, amplified sequences are uppercase and boldface. The primers have been included to give the structural context of the amplification product. Abbreviations: *M. v.*, *Methanococcus vannielii; M. c.*, *Mycoplasma capricolum; E. c.*, *Escherichia coli; T. b.*, *Trypanosoma bruceii; T. t.*, *Tetrahymena thermophila; D. d.*, *Dictyostelium discoideum; S. c.*, *Saccharomyces cerevisiae; H. s.*, *Homo sapiens; H.* sp., *Hexamita* ATCC 50330.

Figure 5-5. Amino acid sequences of alpha-tubulin and EF-1 α genes from *S. muris* (*S. mu*) and *H. inflata* (*H. in*) aligned with homologues from *G. lamblia* (*G. la*), *Hexamita* 50330 (*H.* 30) and *Hexamita* 50380 (*H.* 80).

Alpha-Tubulin

```
H.30   LFCLEHGIHQDGQMPSDKSIGVAEDSFNTFFSETGAGKHVPRCVYIDLEPTVVDEVRAGAYRQIYHPE*L
S.mu   LYCLEHGIHHDGQMPSDKSIGVAEDSFNTFFSETGAGKHVPRAVFIDLEPTVVDEVRAGTYRQIYHPEQL
H.in   LFCLEHCIHHDGQMPSDKSVGVSEDSFNTFFSETGAGKHVPRAVFIDLEPTVVDETRACAYRQLYHPEQL
H.80   LFCLEHGIHQDG*MPSDKSIGVAEDSFNTFFSETGAGKHVPRCVYIDLEPTVVDEVRAGTYRQIYHPE*L

H.30   ISGKEDAANNYARGHYTVGKEVVDLVLDRLRKLADDCSGLQGFMLIHSFGGGTGSGLGSLILERLSVDYG
S.mu   LSGKEDAANNYSRGHNTIGKEVVDLVLDRIRKLADDCSGLQGFIVFHSFGGGTGSGLGSLLLERLSVDYG
H.in   ISGKEDAANNYSRGHNTIGKEVVDLVLDRIRKLADDCSGLQGFMMYHAFGGGTGSGLGSLLLERLSVDY
H.80   ISGKEDAANNYARGHYTVGKEVVDLVLDRLRKLADDCSGLQGFMLHHSFGGGTGSGLGSLILERLSVDYG

H.30   RKTKLEFVIYPSLSIAVSVVEPYNTVLAAHCMLEHSDCAFMIDNEAMYDICHRNLDIERCTYTNINRIVA
S.mu   RKTKLEFVIYPSIHISVSVVEAYNTVHAAHVMLEHSDCAFMVDNEAMYDICHRNLDIERCTYTNINRIIA
H.in   RKTKLEFVIYPSVHIAVSVVEAYNTVHAAHCMLEHSDCAFMVDNEAMYDICHRNLDIERCTYTNINRIIG
H.80   RKTKLEFVIYPSLSIAVSVVEPYNTVLAAHCMLEHSDCAFMIDNEAMYDICHRNLDIERCTYTNINRIVA

H.30   QMISGMTASLRFDGALNVDLTEF*TNLVPYPRVHFPFCSYAPLVSSEKAYHEKLTVAEITNSVFEPANMM
S.mu   QMISGITASLRFDGALNVDLTEFQTNLVPYPRVHFPFCSYAPLVSSEKAYHEKLTVAEITNSVFEPANMM
H.in   QMVSAMTASLRFDGALNVDLTEFQTNLVPYPRVHFPFCSYAPLVSSDKAYHEKLTVAEITNAVFEPANMM
H.80   QMISGMTASLRFDGALNVDLTEF*TNLVPYPRVHFPFRSYAPLVSSEKAYHEKLTVAEITNSVFEPANMM

H.30   VKCDPRHGKYMACCMMYRGDVVPKDVNAAIAVIKTKRTIQFVDWCPTGFKVGINYQPPTVIPGGDLAKVQ
S.mu   VKCDPRHGKYMACCMMYRGDVVPKDVNPAIAVIKTKRTIQFVDWCPTGFKVGINYQPPTVIPGGDLAKVQ
H.in   VKCDPRHGKYMACCMMYRGDVVPKDVNAAIAVIKTKRTIQFVDWCPTGFKVGINYQPPTVIPGGDLAKVQ
H.80   VKCDPRHGKYMACCMMYRGDVVPKDVNAAIAVIKTKRTIQFVDWCPTGFKVGINYQPPTVIPGGDLAKVQ

H.30   RAVLMISNSTAIAEVWSRTDHNFDLMYAKRAFVH
S.mu   RACLMISNSTAIAEVCSRTDKNFDLISAKRAFRH
H.in   RACLMISNSTAIAEVWSRTDKNFDLMFAKRAFVH
H.80   RAVLMISNSTAIAEVWSRTDHNFDLMYAKRAFVH
```

EF-1α

```
G.la   ...STLTGHLIYKCGGIDQRTIDEYEKRATEMGKGSFKYAWVLDQLKDERERGITINIALWKFETKKYIV
H.30   NGKSTLTGHLIYKCGGID*RTLDEYEKRANEMGKGSLKYAWVLDQLKDERERGITINIALWKFETKKFTV
S.mu   NGKSTLTGHLIFKCGGIDKRTIEEYEKKAAEIGKGSFKYAWVLDQLKDERERGITINIALWKFETKNYIV
H.in   NGKSTLTGHLIYKCGGIDQRTLEDYEKKANEIGKGSFKYAWVLDQLKDERERGITINIALWKFETKKFIV

G.la   TIIDAPGHRDFIKNMITGTSQADVAILVVAAGQGEFEAGISKDGQTREHATLANTLGIKTMIICV.NKMD
H.30   TIIDAPGHRDFIKNMITGTSQADVAILVIASGQGEFEAGISKEGQTREHATLAHTLGIKTLIVCV.NKMD
S.mu   TIIDAPGHRDFIKNMITGTAQADVAILVIAAGQGEFEAGISKDGTAREHATLANTLGIRT.IICAINKMD
H.in   TIIDAPGHRDFIKNMITGTSQADVAILVVAAGQGEFEAGISSEGQTREHATLANTLGIKTMIV.AVNKMD

G.la   DGQVKYSKERYDEIKGEMMKQLKNIGW.KKAEEFDYIPTSGWTGDNIMEKSDKMPWYEGPCLIDAIDGLK
H.30   DPQVNYSEARYKEIKEEMQKNLKQIGY.KKWDEFDFIPTSGWTGDSIMEKSPNMPWYSGPCLIDAIDGLK
S.mu   S..IKYDQKRYTEIMEEMKKLLKSIGYGKKAEEFHYIPVSGWIGDNIMEKSENMPWYTGKCLIEAIDELK
H.in   DPQVNYSEARYTEIKTEMQKTFKQIGF.KHWEEFDFVPLSGWTGDNIMEASPKTPWYKGKCLIECIDGLK

G.la   APKRPTDKPLRLPIQDVYKISGVGTVPAGRVETGELAPGMKVVFAPTSQVSEVKSVEMHHEELKKAGPGD
H.30   APKRPTDKPLRLPIQDVYKINGVGTVPAGRVESGLLIPNMTVVFAPSTTTAEVKSVEMHHEELPQAGPGD
S mu   PPKRPTDKPLRLPLQDVYKISGIGTVPAGRVESGVLKPGQIVVFAPSDESGEVKSVEMHHESLPQAVPGE
H.in   APKRPNDKPLRLPIQDVYKINGVGTVPAGRVESGELIPGMMVVFAPAGEKTEVKSVEMHHELAKAGPGD

G.la   NVGFNVRGLAVKDLKKGYVVGDVTNDPPVGCKSFTAQVIVMNHPKKIQPGYTPVIDCHTAHIACQFQLFL
H.30   NVGFNVRGIAAKDIKKGYVVGDTKNDPPVGCKSFTAQVIIMNHPKKIQPGYSPVIDCHTAHIACKFDAFL
S.mu   MVGSN.....
H.in   NVGFNIKGLSAKDIKKGYVVGDVNNDAPKGCEYFKANVIIMNHPKKINPGYTPVLDCHTSHLAWKFDKFL

G.la   QKLDKRTLKPEMENPPDAGRGDCIIVKMVPQKPLCCETFNDYAPLGPFAVR
H.30   QKLNARTLKPEMENPTEASRGECIVVRMVPSKPLSCESFNDYAALGRFAVR
S.mu   .....
H.in   AKLNSRTFKVEIENPTEAVRGECVLMQIVPTKPLCVESFEQYPALGRFAVR
```

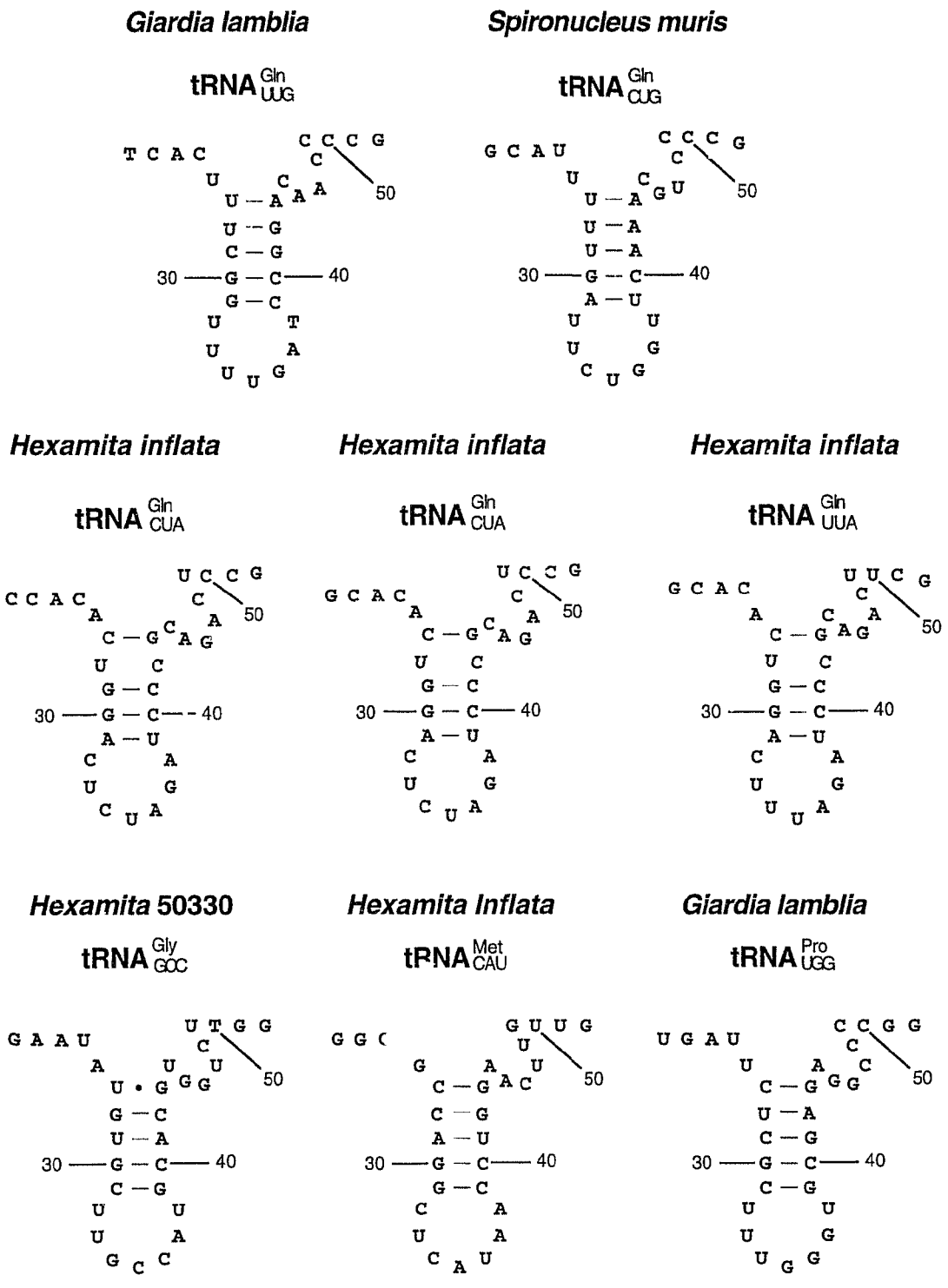


Figure 5-6. Primary and putative secondary structure of tRNA-like amplification products from *G. lamblia, S. muris* and *H. inflata*. In each case the sequence can be folded into stem loop structures resembling anticodon stems with the triplet 13 bp from the 5' end in the anticodon position.

were obtained from the free living diplomonad, *H. inflata*. In this case, despite the fact that no termination codons were observed in the 783 codons comprising α-tubulin and EF-1α, genes for tRNAs that decode both TAA and TAG were readily obtained (in this case, as in *Hexamita* 50330, no canonical tRNAs were observed, but a $tRNA^{Met}_{CAU}$ was spuriously amplified).

Stop codon usage in *Hexamita* genes. The use of this variant code requires that all legitimate termination codons be TGA. To see if this prediction holds in these organisms, the carboxy terminis of several genes from *Hexamita* species were sought. Specific oligonucleotides (HA550f, SART, HSEF3, and SPEF3) were used in combination with a random primer (uniN) to generate amplification products that extend beyond the known sequence. Fragments generated in such reactions were chosen for characterisation based only on their length exceeding the predicted end of the gene.

Clones of this nature were sequenced from alpha-tubulin and EF-1α-specific reactions from both *Hexamita* 50380 and *H. inflata*. Amazingly, all of the products obtained were artifacts generated by the specific primer alone, and yet two of these proved to be exactly the fragments that were sought. These were the 3' ends of the *H. inflata* EF-1α and *Hexamita* 50380 alpha-tubulin genes. The *H. inflata* alpha-tubulin carboxy terminus was therefore sought using a specific primer alone (A860f), and once again a product was cloned and sequenced and found to be the expected fragment of the genome.

In each of the three cases a considerable overlap with the target gene allows some estimate as to whether the fragments come from identical alleles. The alpha-tubulin fragments from both *Hexamta* 50380 and *H. inflata* were identical throughout overlaps of 523 and 328 bp respectively. However, the *H. inflata* EF-

1α fragment differed at 5 sites out of 556 bp of overlap suggesting that this fragment comes from a recently duplicated but extremely similar gene.

Following the region of overlap each fragment also contained a short stretch of coding region missing from the original PCR clone (137, 131, and 56 bp for *Hexamita* 50380 alpha-tubulin, *H. inflata* alpha-tubulin and EF-1α respectively) all followed by a TGA termination codon as predicted.

Downstream of each termination codons was also a length of non-coding DNA. The nature of the non-coding DNA was unexpected as it is quite high in AT pairs in contrast to the generally high GC content of the coding regions (Table 5-1).

Table 5-1. Characteristics of *Hexamita* coding and non-coding DNA

	50380 α-tubulin	*H. inflata* α-tubulin	*H.inflata* EF-1α
coding			
length	1291	1285	1255
GC %	50.0	53.5	51.0
non-coding			
length	96	378	337
GC %	26.0	28.0	38.0
GC % ratio	**1.92**	**1.91**	**1.34**

Phylogenetic relationships among the diplomonads: Phylogenetic trees based on EF-1α amino acid sequences were inferred using maximum likelihood, parsimony and distance (Figure 5-7). The large dataset analysed by parsimony and distance methods confirms the very early divergence of diplomonads, and argues very strongly that diplomonads are a monophyletic taxon. In addition, all methods gave the same topology for the diplomonads, although the statistical support for the relationship between *H. inflata* and *Hexamita* 50330 is very weak in parsimony and

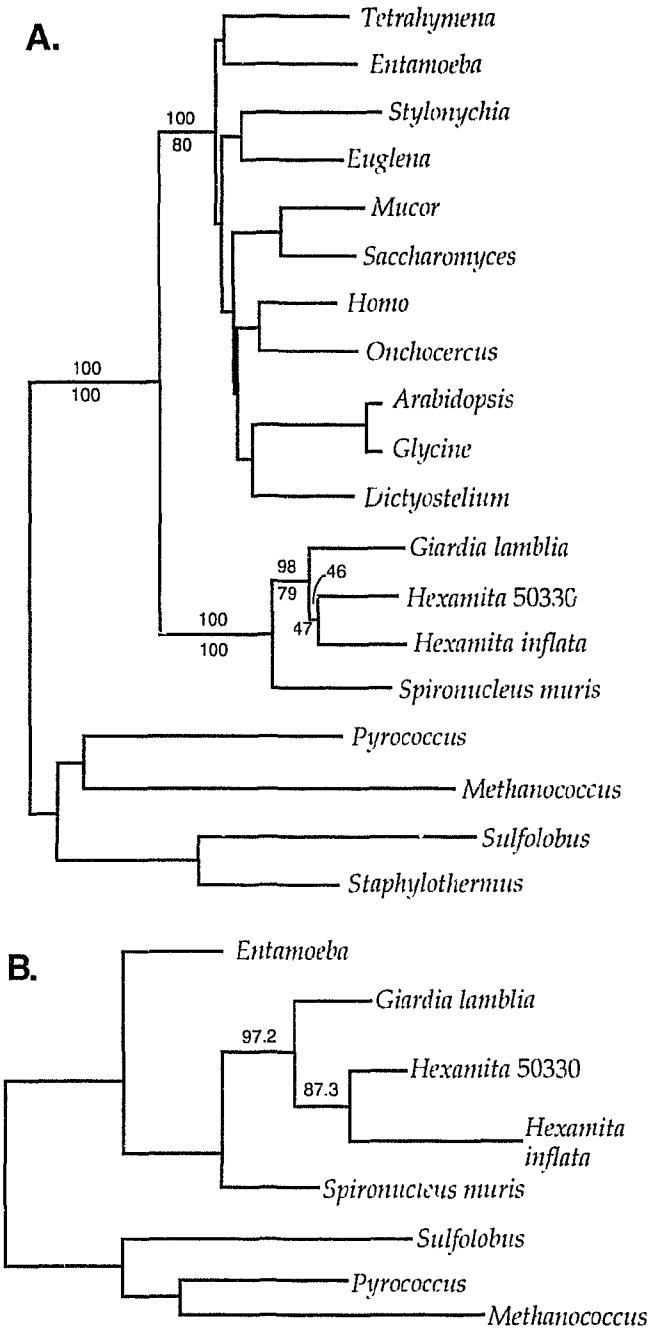


Figure 5-7. Phylogenetic tree of selected EF-1α sequences. (A) Parsimony and distance topology shown with distance branch lengths. Bootstrap percents for nodes within and immediately surrounding the diplomonads are shown on the figure (distance above the node, parsimony below), others are excluded for clarity. (B) Maximum likelihood topology of a restricted dataset. Estimated bootstrap percent is shown for all unconstrained nodes.

distance analyses (47% and 46% respectively). Nevertheless, the support for this relationship is highly significant in maximum likelihood (estimated bootstrap of 87%, and a 96% confidence that this topology is superior to any other), which has been shown to be more consistently correct in inferring relationships when rates are unequal between taxa or between sites within a taxon (Hasegawa and Fujiwara, 1993). In general, EF-1α phylogeny tends to support the conclusion that the two taxa with the non-canonical genetic code are themselves a clade. This has also now been seen in phylogenies based on GAPDH (Rozario *et al.*, 1996), but there must be more taxa included in both datasets before any firm conclusions can be drawn. In ciliates it has become clear that the same variant code evolved in several groups independently (Tourancheau *et al.*, 1995), but this does not seem to be the case in these diplomonads.

Discussion

Changes to the genetic code in the nucleus are very rare. This case is only the fifth to be discovered, and interestingly, three of the others also involve TAA and TAG stop codons (both) changing to glutamine. The common involvement of termination codons in code alterations might be explained by their relatively low frequency, their functional redundancy, and the fact that occasional failure to terminate translation of some proteins following loss of release factors should be less detrimental than the failure to complete translation of some proteins because of loss of tRNAs -- factors that mitigate the effects of their loss. However, in eukaryotes the specific and *simultaneous* capture of TAA and TAG by glutamine suggests that some further special relationships exist between these codons: no variant codes in which either TAR codon has been replaced by an amino acid other than glutamine have been described, and TAA and TAG seem always to be replaced together. These issues are addressed in turn.

Why always glutamine? The glutamine-encoding TAA and TAG in *Hexamita* genes presumably arose from CAA and CAG codons. Other organisms (ciliates and *Acetabularia*) where TAR codes for glutamine are very AT-rich (as high as 76%: Schneider *et al.*, 1989; Prescott, 1994), and this has led to the suggestion (Osawa and Jukes, 1989; Osawa *et al.* 1992) that the AT mutation pressure which biased the overall composition of these genomes has also driven the conversion of many CAR codons to TAR, once the original chain-terminating TARs had been fortuitously reduced in number to the point where release factors recognising them could be lost with impunity. However, the genome of *Hexamita* appears to be GC-rich: the overall and third position GC content of these genes from *Hexamita* 50330 are 53% and 63% respectively and those from *H. inflata* are 52% and 64% respectively. This argues that AT mutation pressure is not necessary to explain the appearance of TAA and TAG glutamine codons.

If directional mutation pressure is not a requirement (although it may contribute in other situations), then the answer might lie in the fact that canonical glutamine CAA and CAG codons and anticodons are, together with TTG tryptophan codons, the only sense triplets that can be converted to TAA or TAG by single *transitions*. Novel tRNA genes arising by chance duplication and base substitutions in the anticodon will not be maintained by selection until codons that require their services have also arisen by chance within coding regions. Both these events will generally take place most frequently when they result from transitions rather than transversions (Kimura, 1980), so modifications to the specificity of TAR codons will tend to involve glutamine, regardless of directional mutation pressure.

Capture of TAA and TAG by glutamine could be further facilitated if G-U pairing would allow a single tRNA with anticodon UUA to recapture both UAA and UAG as Gln codons, and this has been suggested (Osawa *et al.*, 1992).

However, uridine residues in the first position of NNR decoding tRNAs are usually modified to one of several derivatives that pair strongly with A but weakly with G (Björk, 1995), necessitating a second tRNA to decode NNG. Even in *Tetrahymena*, where the first position U exhibits a rare modification that does allow both A and G to be recognised (Schull and Beier, 1994), there are still two tRNAs to decode the variant Gln codons UAA and UAG (Hanyu *et al.*, 1986). Based on the results described here, it appears that *Hexamita* 50330 and *H. inflata* use both isoacceptors to decode TAR. This may point to a deeper general reason why two tRNAs are always required to decode TAR, but it is simpler to suppose that species of *Hexamita* modify U34 in a more conventional fashion than *Tetrahymena*. This supposition is supported by the presence of a $tRNA^{Gln}CUG$ in *S. muris*, which suggests that in other diplomonads both CUG and UUG isoacceptors are used to decode CAR.

Why TAA and TAG together? Without a tRNA that efficiently recognises both codons, it is unlikely that the conversion of TAA and TAG to Gln codons took place simultaneously. A more plausible scenario is that two separate iterations of the codon capture process took place, each involving one unassigned codon. If this is correct, then the fact that in eukaryotes TAA and TAG are always reassigned together may mean that both have to be lost as functioning nonsense codons before either can be recaptured as sense.

If, for instance, some activity of the eukaryotic peptide release mechanism was common to termination exclusively by TAA and TAG, then neither of these codons could appear as sense within the coding regions of genes until that activity was rendered superfluous, and lost. This in turn could not take place until neither codon was absolutely essential for the termination of any gene. A possible role for release factors in this process is also suggested by the phylogenetic restriction of glutamine-specifying TAR codons to the eukaryotic nucleus. The eubacterial

peptide termination system does not appear to be homologous to that of eukaryotes (Frolova *et al.*, 1994; Zhouravela *et al.*, 1995) and uses two codon-specific factors, one recognising UAA and UAG, and the other recognising UAA and UGA (Scolnick *et al.*, 1968; Caskey *et al.*, 1968). This redundancy makes the loss of UAA (and because of wobble this extends to UAG) very unlikely in eubacteria. Indeed, only UGA has been lost in any eubacterial or organellar system, in contrast to the nucleus where changes involving UAA and UAG are the most common alteration to the nuclear genetic code.

Chapter VI:

Eukaryotic Triosephosphate Isomerase Originated with the Mitochondrial Symbiont

Introduction

For at least two decades, we have known that animals, plants, fungi and most protists are chimeric: their DNA-containing organelles have evolutionary histories different from that of the nucleus (Gray, 1992). Mitochondria are the degenerate descendants of once free-living eubacteria that engaged in an endosymbiotic association with a "primitive" and presumably organelle-free nucleated host cell, perhaps two billion years ago. The genes retained in the mitochondrial genome show that this eubacterium was what we would now call an alpha-proteobacterium, a relative of modern genera such as *Rhizobium*, *Agrobacterium* and *Rickettsia* (Yang *et al.*, 1985). Similarly, plastid genes derive from the genome of a photosynthetic endosymbiont whose nearest modern relatives are cyanobacteria (Bonen and Doolittle, 1975).

The remaining component, the early nucleated host that welcomed the first (proto-mitochondrial) endosymbiont was itself related to the ancestor of modern archaebacteria. Rooted phylogenetic trees based on translation elongation factors and aminoacyl-tRNA synthetases show that the archaebacteria are the sister group of the eukaryotes (Iwabe *et al.*, 1989; Brown and Doolittle, 1995) and the sequences of many other essential components of the transcription and translation apparatus also reveal a strong archaebacterial-eukaryotic affinity (Zillig *et al.*, 1993; Kletzin, 1992; Keeling *et al.*, 1996). In many other instances, transcription and translation factors that are found in both archaebacteria and eukaryotes are altogether absent from eubacteria (for review see Keeling and Doolittle, 1995).

The general belief that eukaryotic nuclear genomes share a close ancestry with archaebacteria while the eukaryotic organellar genomes are eubacterial in nature admits to two exceptions. First, some lineages thought to have diverged early in eukaryotic evolution (Diplomonads and Microsporidia) in fact have no mitochondria. These lineages, which Cavalier-Smith has called Archezoa (Cavalier-Smith, 1983), may have never acquired mitochondria, and would therefore represent the original condition of the host. Second, many genes determining proteins that function in mitochondria or plastids actually reside in the eukaryotic nuclear genome. These genes most often resemble eubacterial homologues and are thought to have been transferred to the nucleus from the symbiont genome, in most cases soon after the endosymbiosis was established. Isolated instances of organelle to nucleus transfer occurring more recently in evolution can still be documented for both mitochondria and plastids (Baldauf & Palmer, 1990; Covello & Gray, 1992; Grohman *et al.*, 1992; Nugent & Palmer, 1991).

In almost all widely-accepted instances of such transfer, the product of the transferred gene still functions in the organelle in which it originally resided. We are aware of only one case in which an organelle gene seems to have replaced a nuclear homologue and assumed its cytosolic function. This is in chlorophytes where there are two nuclear-encoded phosphoglycerate kinase (PGK) genes, one specific for the cytosol and one targeted to the plastid. In land plants the cytosol-specific gene is significantly more similar to the choroplast-specific gene (and thus to eubacterial genes) than to other eukaryotic cytosol-specific genes. This was originally attributed to a high level of intergenic recombination (Longstaff *et al.*, 1989), but the data are more consistent with the nuclear-encoded chloroplast-targeted gene having duplicated at some point after the divergence of land plants from chlorophyte algae, but before dicots and monocots diverged, and having replaced its nuclear-

encoded cytosol-specific counterpart (Brinkmann and Martin, 1996). This two-step process could lead to the take-over of cytosolic function by an organellar gene.

Here we present data consistent with another such take-over, involving triosephosphate isomerase (TPI). This enzyme is central to glucose metabolism, and is exclusively cytosolic in function (except for plastid isoforms). A preliminary analysis of a limited number of TPI sequences by Schmidt and co-workers (1995) indicated that eukaryotic TPIs branched with Gram-negative bacteria (although these authors did not comment on this result). We reasoned that if TPIs were in fact of mitochondrial origin, then a better phylogenetic spread of TPI sequences ought to reveal a specific, close relationship to proteobacteria, or more specifically to the alpha subgroup from which mitochondria likely arose.

We isolated and sequenced TPI genes from three diverse eubacteria, the gamma-proteobacterium *Francisella tularensis*, the green non-sulfur bacterium *Chloroflexus aurentiacus*, and the alpha-proteobacterium *Rhizobium etli*, predicting that the eukaryotes should branch at least weakly with *R. etli*. Phylogenetic analyses including these new sequences confirmed the association between the eukaryotes and proteobacteria, and did indeed place *R. etli* alone as the outgroup to eukaryotes. Of all the prokaryotes, the archaebacteria (represented by a single sequence from *Pyrococcus woesei*) actually branch most distantly from the eukaryotes. Since cytosolic genes tend to be most closely related to archaebacteria, and mitochondria are of alpha-proteobacterial ancestry, it seems most parsimonious to assume that these eukaryotic TPI genes were transferred into the eukaryotic nuclear genome from the genome of the mitochondrial endosymbiont. Such an assumption has ramifications for current theories about early eukaryote evolution (Cavalier-Smith, 1983 & 1993) and for arguments based on TPI that have been used in the "introns early *vs.* introns late" debate (Gilbert *et al.*, 1987; Tittiger *et al.*, 1993; Stoltzfus *et al.*, 1994; Logsdon *et al.*, 1995; Kwaitowski *et al.*, 1995).

Results

Identification of TPI genes from eubacteria. PCR amplification reactions on genomic DNA from a diverse selection of eubacteria were carried out using a battery of primers specific for highly conserved portions of TPI. In many cases these reactions resulted in major products of the expected size which were cloned and sequenced; however, most of these proved to be unidentifiable, and were discarded. A small fragment of the TPI gene from *Helicobacter pylori* was obtained using primers TF4 and TR2, and several other clones proved to be identifiable, but were not TPI. Noteworthy among the latter two *R. prowezekii* genes, UDP-N-acetylglucosamine pyrophosphorylase, and an uncharacterised ORF that maps close to the replication origin in *E. coli*. Also found were an *fts*H homologue from the cyanobacterium *Prochloron*, NADH dehydrogenase subunit 5 from *Agrobacterium radiobacter*, and the glucose 6-phosphate isomerase gene of *Helicobacter pylori*. These hits are shown in Appendix E, including the complete sequence of the 750 bp clone of *H. pylori* glucose 6-phosphate isomerase.

Amplification products covering over 90% of the TPI gene were isolated from *Chloroflexus aurantiacus, Francisella tularensis*, and *Rhizobium etli* (Figure 6-1). In the former two species the sequence was isolated as a single 730 bp fragment using primers TF1 and TR1. In both cases two clones were chosen and sequenced on both strands and were found to be identical. This prime combination in *R. etli* failed, so the same portion of the gene was amplified in two overlapping pieces using TF1 and TR2 for the 5' end and TF4 and TR1 for the 3' end. Six clones of the 3' end were sequenced and found to be identical, but of the six clones isolated and sequenced from the 5' end, three distinct TPI coding sequences were identified. The three variants, type 1, 2 and 3, share between 72.5 to 65.3% identity at the amino acid level. Divergence notwithstanding, these three sequences are extremely similar to one another compared to other genes from other species,

```
G.lamb   MPARRPFIG GNFKCNGSLD FIKSHVASIA SY.KIPESVD VVVAFSFVHL STAIAAN... .TSKCLKIAA
H.sapi   MAPSRKFFVG GNWKMNGPKQ SLGELIGTLN AA.KVPADTE VVCAPPTAYI DF\RQKL... ..DPKIAVAA
S.cere     MARTFFVG GNFKLNGSKQ SIKEIVERLN TA.SIPENVE VVICPPATYL DYSVSLV... .KKPQVTVGA
P.falc     MARKYFVA ANWKCNGTLE SIKSLTNSFN NLDFDPSKLD VVVFPVSVHY DHTRKLL... ..QSKFSTGI
R.etli1  ---------- ----MNPMQA DAKQLLQEFK QLLQENEITE EKCLAPVTLA LTSTQAELAN AARSVF.TVA
R.etli2  - -------- ----MNPLQT DAQTLLRGVK DLLESTPISA EKCHLGVAVA IALTQVQAEL ASAVRVYTVA
R.etli3  ---------- ----MNPMQA NAQQLIQDLK QRLLQEVVSE QDCHIGIAIS IALLSVKAQL DDASVIATVA
F.tula   ---------- ----MNGNST SIKELCSGIS QVQYDTSRVA IAVFPSSVYV KEVISQLPE. ....KVGVGL
E.coli     MRHPLVM GNWKLNGSRH MVHELVSNLR KELAGVAGCA VAIAPPEMYI DMAKREAEG. ...SHIMLGA
H.pylo   ---------- ---------- ---------- ---------- ---------- ---------- ----------
B.subt     MRKPIIA GNWKMNKTLG EAVSFVEEVK SSIPAADKAE AVVS.PALFL EKLASAVKG. ...TDLKVGA
C.aure   ---------- ----MYKTVG EATTLVRDLL AGLGELSDRE AIVCPPFTAL AAVALVADS. ....PLGLGA
T.mari     ITRKLILA GNWKMHKTIS EAKKFVSLLV NELHDVKEFE IVVCPPFTAL SEVGEILSG. ...RNIKLGA

G.lamb   QNVYLEGN.G AWTGETSVEM LLDMGLSHVI IGHSERRRIM GETNEQSAKK AKRALDKGMT VIFCTGETLD
H.sapi   QNCYKVTN.G AFTGEISPGM IKDCGATWVV LGHSERRHVF GESDELIGQK VAHALAEGLG VIACIGEKLD
S.cere   QNAYLKAS.G AFTGENSVDQ IKDVGAKWVI LGHSERRSYF HEDDKFIADK TKFALGQGVG VILCIGETLE
P.falc   QNVSKFGN.G SYTGEVSAEI AKDLNIEYVI IGHFERRKYF HETDEDVREK LQASLKNNLK AVVCFGESLE
R.etli1  QDVSRFAH.G AYTGEVSAEL LKDSQIEYVL IGHSERREYF AESAAILNAK AQNALNAGLK VIYCVGESLE
R.etli2  QDVSRIAG  AYTGEVSAEL LADSGIGYVL VGHSERREIF GESREILNTK IKNALNAGLT VIYCVGESLE
R.etli3  RDVSRMAGIG AYTGEVSADL LVDSGIQFVL IGHSERREIF GDNPQILSDK IHYALNANMT IIYCVGESLE
F.tula   QNITFYDD.G AYTGEISARM LEDIGCDYLL IGHSERRSLF AESDEDVFKK LNKIIDTTIT PVVCIGESLD
E.coli   QNVNLNLS.G AFTGETSAAM LKDIGAQYII IGHSERRTYH KESDELIAKK FAVLKEQGLT PVLCIGETEA
H.pylo   ---------- -----ITSQH LEELKIHTLL IGHSERRTLL KESPSFLKEK FDFFKSKNFK IVYCIGEELI
B.subt   QNMHFEES.G AFTGEISPVA LKDLGVDYCV IGHSERREMF AETDETVNKK AHAAFKHGIV PIICVGETLE
C.aure   QNLYPEAQ.G AFTGEVSPPM LVDIGCRYVI IGHSERRQYF GESDAFVNRK LRAALAHGLR PIVCVGESKP
T.mari   QNVFYEDQ.G AFTGEISPLM LQEIGVEYVI VGHSERRRIF KEDDEFINRK VKAVLEKGMT PILCVGETLE

G.lamb   ERKANNTMEV NIAQLEALKK EIGESKKLWE NVVIAYEPVW SIGTGVVATP EQAEEVHVGL RKWFAEKVCA
H.sapi   EREAGITEKV VFEQTKVIAD NVK...D.WS KVVLAYEPVW AIGTGKTATP QQAQEVHEKL RGWLKSNVSD
S.cere   EKKAGKTLDV VERQLNAVLE EVK...D.WT NVVVAYEPVW AIGTGLAATP EDAQDIHASI RKFLASKLGD
P.falc   QREQNKTIEV ITKQVKAFVD LI....DNFD NVILAYEPLW AIGTGKTATP EQAQLVHKEI RKIVKDTCGE
R.etli1  QRESGQAEVV VLQQICDLAS VVT..AEQWP HIVIAYEPIW AIGTGKTASP EDAQTMHAKI REGLTQITSH
R.etli2  QREAGQAEAV VLQQICDIAA VVE..AEQWK NIVIAYEPI- ---------- ---------- ----------
R.etli3  QRESGQAEQI VLQQICDVAS VVK..AEQWH NIIIAYEPI- ---------- ---------- ----------
F.tula   DRKSGKLKQV LATQLSLILE NLS..VEQLA KVVIAYEPVW AIGTGVVASL EQIQETHQFI RSLLAKV.DE
E.coli   ENEAGKTEEV CARQIDAVLK TQG..AAAFE GAVIAYEPVW AIGTGKSATP AQAQAVHKFI RDHIAKV.DA
H.pylo   TREKG..FKA VKEFLSEQLE NID...LSYS NLIVAYEPI- ---------- ---------- ----------
B.subt   EREAGKTNDL VADQVKKGLA GLS..EEQVA ASVIAYEPIW AIGTGASTA KDANDVCAHI RKTVAESFSQ
C.aure   QRDAGQAEPI VTAQVRAALL EVP..PDQMA NVVIAYEPIW AIGTGDTATP ADAQAMHAAI RATLAELYGS
T.mari   EREKGLTFCV VEKQVREGFY GLD..KEEAK RVVIAYEPVW AIGTGRVATP QQAQEVHAFI RKLLSEMYDE

G.lamb   EGAQHIRIIY GGSANGSNCE KLGQCPNIDG FLVGGASLKP EFTTMIDILA KTRA
H.sapi   AVAQSTRIIY GGSVTGATCK ELASQPDVDG FLVGGASLKP EF...VDIIN AKQ
S.cere   KAASELRILY GGSANGSNAV TFKDKADVDG FLVGGASLKP EF...VDIIN SRN
P.falc   KQANQIRILY GGSVNTENCS SLIQQEDIDG FLVGNASLKE SF...VDIIK SAM
R.etli1  GA..NMAILY GGSVKAENAV ELAACPDING AL-------- ---------- ------
R.etli2  ---------- ---------- ---------- ---------- ---------- ------
R.etli3  ---------- ---------- ---------- ---------- ---------- ------
F.tula   RLAKNIKIVY GGSLKAENAK DILSLPDVDG GL-------- ---------- ------
E.coli   NIAEQVIIQY GGSVNASNAA ELFAQPDIDG ALVGGASLKA DAFAVIVKAA EAAKQA
H.pylo   ---------- ---------- ---------- ---------- ---------- ------
B.subt   EAADKLRIQY GGSVKPANIK EYMAESDIDG ALVGGASLEP QSFVQLLEEG QYE
C.aure   EIAATVRIQY GGSVKPDNID ELMAQPDIDG A--------- ---------- ------
T.mari   ETAGSIRILY GGSIKPDNFL GLIVQKDIDG GLVGGASLK. ESFIELARIM RGVIS
```

Figure 6-1. Amino acid sequences inferred from TPI genes of *R. etli* (*R.etli*1-3), *F. tularensis* (*F.tula*), *C. aurentiacus* (*C.aure*), and *H. pylori* (*H.pylo*) aligned with those of *E. coli* (*E.coli*), *T. maritima* (*T.mari*), *B. subtilis* (*B.subt*), *G. lamblia* (*G.lamb*), *S. cerevisiae* (*S.cere*), *P. falciparum* (*P.falc*), and human (*H.sapi*). Dots represent gaps in the alignment, spaces represent length heterogeneity, and dashes represent missing data.

and also contain shared insertions and deletions, strongly supporting the notion that they are recently diplicated genes. Of the three, type 1 was identical over the 282 bp overlap region with the six 3' clones, so these have been treated as two fragments of a single gene.

Phylogeny of TPI genes. Phylogenetic trees based on TPI sequences from 22 eukaryotes, 18 eubacteria, and one archaebacterium were inferred using unweighted parsimony, corrected distance and maximum likelihood methods. Parsimony and neighbor-joining trees based on 265 sites are shown in Figure 6-2 and 6-3 respectively, each with bootstrap proportions for nodes over 30%. Parsimony analysis yielded five trees of identical length (2745 steps) that differed only in minor characteristics of the relative branching order within the eukaryotes. The neighbor-joining and parsimony trees also differed in the order of a handful of branches, but none that was significantly supported in either analysis, and none that is central to the questions posed here. Two constant features of all the preferred trees is the association between eukaryotes and proteobacteria, specifically the alpha-proteobacterium *R. etli*, and the relatively great distance between eukaryotic and archaebacterial sequences. Protein maximum likelihood analysis was conducted on 213 positions by constraining the topology of the eukaryotes to match that shown in Figure 6-2, and dividing the prokaryotes into nine groups: *Rhizobium*, gamma-proteobacteria, mycoplasmas, low GC Gram-positive bacteria, Actinomycetes, *Thermotoga*, *Chloroflexus*, *Synechocystis*, and *Pyrococcus*. The branching order of these ten groups was then exhaustively searched, but in both cases, the best 100 trees were virtually indistinguishable from one another statistically (all but a few being within 1.98 standard errors from the best tree). Once again these trees were consistent in that the eukaryotes were quite distant from

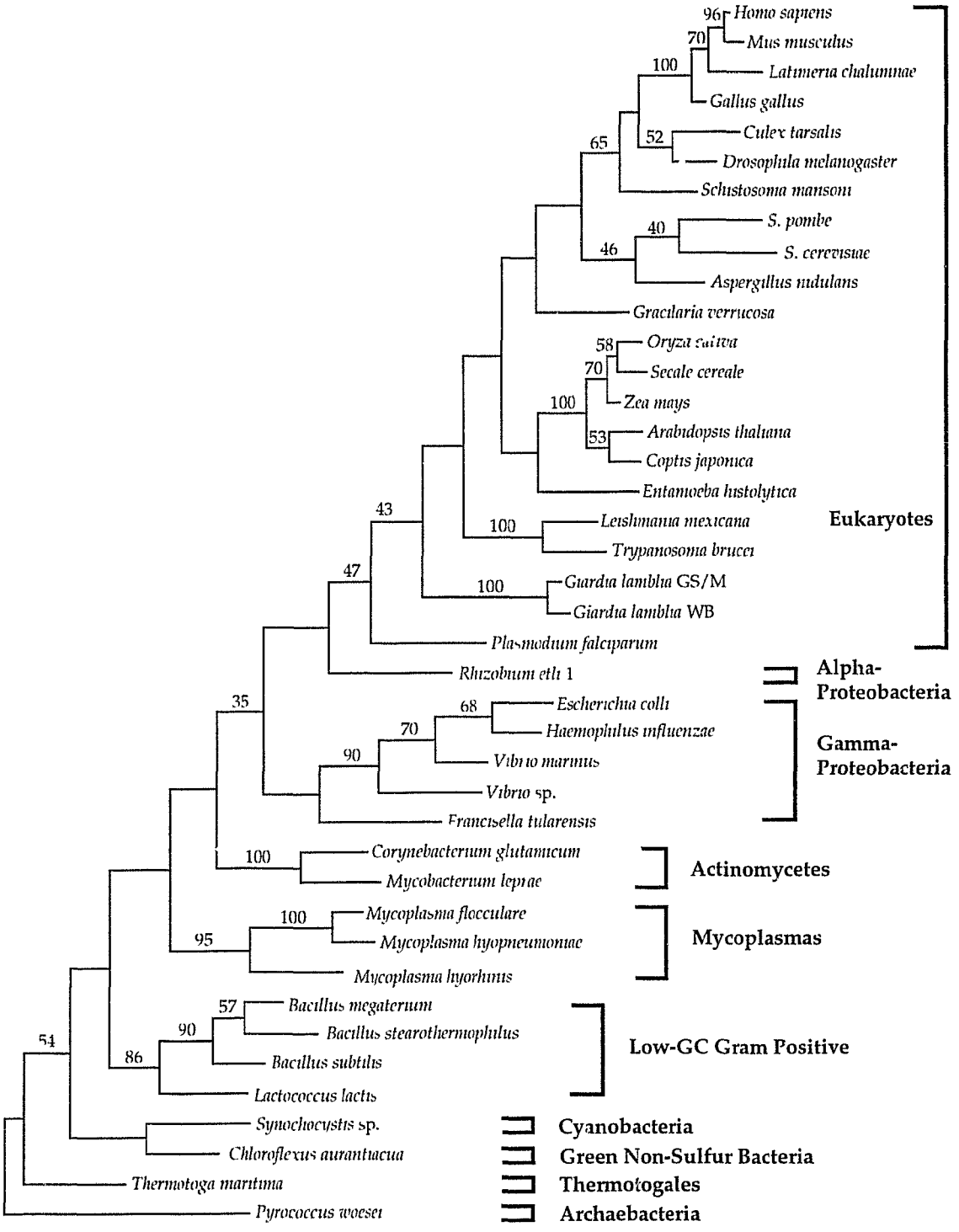


Figure 6-2. Unweighted parsimony tree of TPI amino acid sequences from eukaryotes, eubacteria, and *P. woesei*. Bootstrap support is shown for nodes where it is over 30%. Eukaryotes and major subdivisions of eubacteria are delineated to the right by brackets.

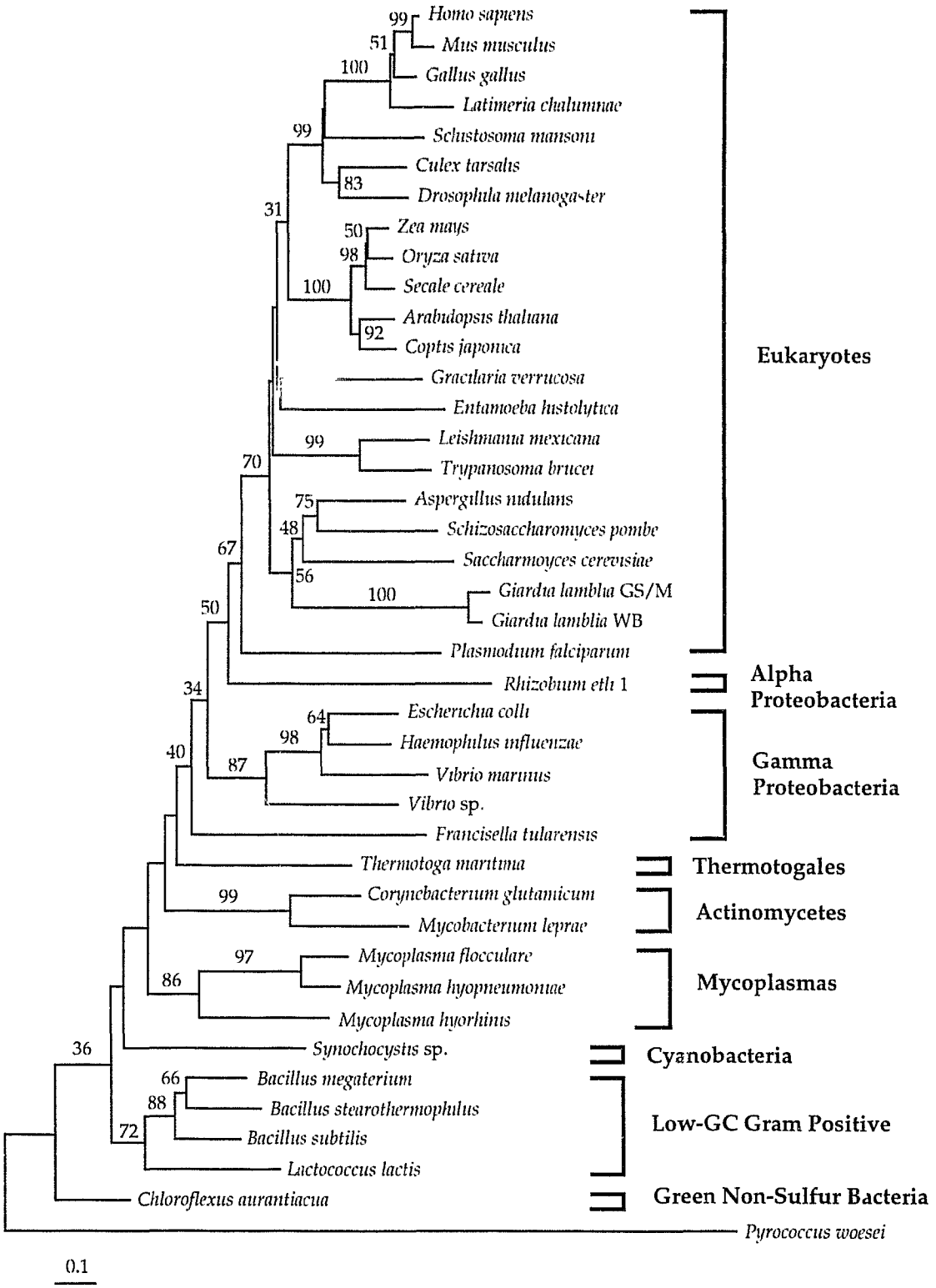


Figure 6-3. Neighbor-joining tree of TPI amino acid sequences from eukaryotes, eubacteria, and *P. woesei*. Bootstrap support is shown for nodes where it is over 30%. Eukaryotes and major subdivisions of eubacteria are delineated to the right by brackets. Scale represents an estimated 0.1 substitutions per site.

the archaebacterium, and close to the proteobacteria, although in this case the closest relationship was to the gamma subdivision (data not shown).

There is now a great deal of support both from molecular phylogeny and molecular biology that among all prokaryotes, the archaebacteria are the closest relatives of the eukaryotes. The expected phylogenetic relationship of a eukaryotic cytosolic enzyme is therefore exactly the opposite of that observed here: the eukaryotes *ought* to branch closer to the archaebacteria than they do to eubacteria. The highly divergent nature of the archaebacterial enzyme (Kohlhoff *et al.*, 1996) is a concern as it could conceivably lead to an erroneous phylogeny, so all three analyses were repeated excluding the *Pyrococcus* sequence from the alignment. This deletion had little effect on the outcome of the trees; once again parsimony and distance analyses showed *Rhizobium* as the immediate outgroup of eukaryotes, with the gamma-proteobacteria next followed by the remaining eubacteria, in the same order as the trees shown in Figure 6-2 (not shown), and protein maximum likelihood showed a general affinity between eukaryotes and proteobacteria, but with little support to distinguish one tree over any other.

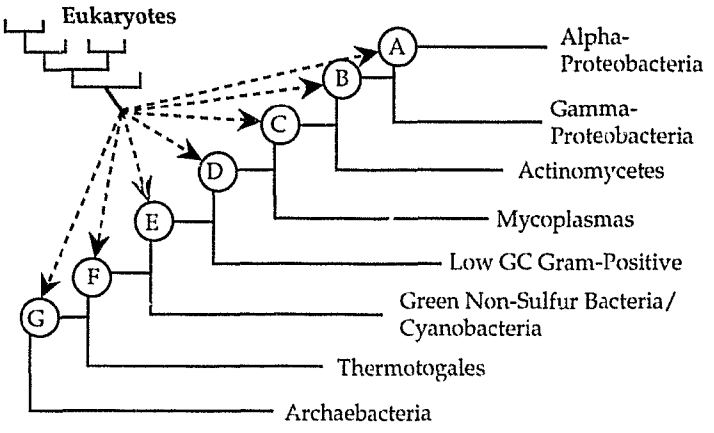
Significance of the TPI tree topology. The significance of the relationship between eukaryotic and proteobacterial TPI genes was tested first by performing 100 bootstrap replicates on the distance and parsimony analyses. In both cases the bootstrap support was almost universally low for all nodes but those between the most closely related taxa. TPI is a relatively small protein, with apparently little power to resolve organismal phylogeny. Nonetheless, the significance and relatively high bootstrap support for the position of the eukaryotes within the eubacteria was specifically tested by comparing the "TPI tree" (using the topology inferred by both parsimony and neighbor-joining) to alternative trees, according to the method described by Templeton (1983). The alternative topologies were chosen

by rooting the eukaryotes in each of the main inter-group nodes within the prokaryotes for both neighbor-joining and parsimony topologies. Figures 6-4 shows the results of these tests for the topology predicted by parsimony and Figure 6-5 the results for the neighbor-joining topology. In all tests the "TPI tree" is superior to all alternatives, and is significantly so in all cases except that where the outgroup of eukaryotes is all proteobacteria. Topology H in Figure 6-4 and G in Figure 6-5 are of particular interest, as these trees are a very close approximation to the topology of universal trees predicted by other molecular markers (archaebacteria as sisters to eukaryotes), and yet these trees are between 2 and 4.4 standard errors *worse* in both tests than the topologies actually inferred from TPI. Once again these tests were repeated on the tree topologies derived by excluding *Pyrococcus*, and as before the archaebacterial sequence was not seen to be unduly affecting the analysis: the ultimate result was the same and statistical significance was changed very little (data not shown).

The possibility that the branching order defined for the eukaryotes was having some negative effect on the likelihood of alternative topologies was examined by conducting 100 independent maximum likelihood replicates with a random pair of eukaryotic sequences (so that there is only one topology) and 16 eubacteria. In every one of these 100 replicates the TPI topology was the best, although once again the difference between the *R. etli* specifically or all the proteobacteria was often insignificant.

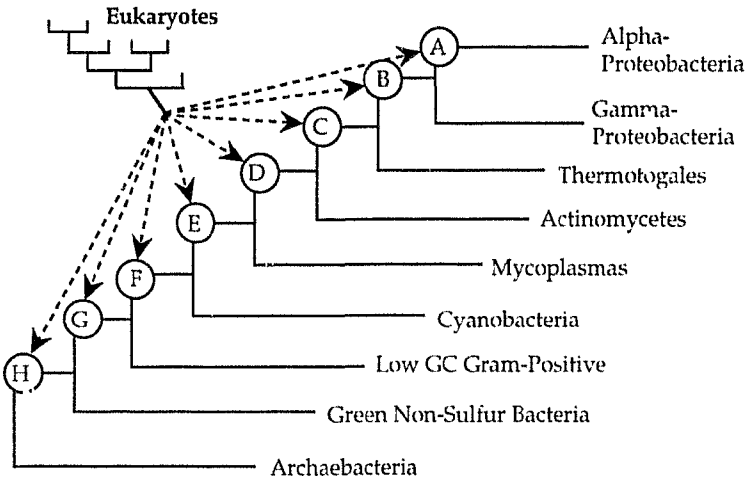
Discussion

A relationship between eukaryotic and proteobacterial TPI sequences has been supported by the addition of TPI genes from both early-diverging (*Chloroflexus*) and later-diverging (*Rhizobium* and *Francisella*) eubacterial lineages. Furthermore, the relationship between *Rhizobium* and the eukaryotes inferred from TPI data



Position of Eukaryotes	Parsimony	Maximum Likelihood
A	Length 2745 **BEST**	-lnL 12459.5 **BEST**
B	Length 2756 Δ length 11 SE 3.5 insignificant	-lnL 12466.4 Δ lnL 6.9 SE 4.7 insignificant
C	Length 2762 Δ length 17 SE 4.9 **WORSE**	-lnL 12480.1 Δ lnL 20.7 SE 9.3 **WORSE**
D	Length 2767 Δ length 22 SE 6.0 **WORSE**	-lnL 12485.0 Δ lnL 25.5 SE 10.8 **WORSE**
E	Length 2771 Δ length 26 SE 6.4 **WORSE**	-lnL 12493.9 Δ lnL 34.4 SE 12.5 **WORSE**
F	Length 2766 Δ length 21 SE 6.3 **WORSE**	-lnL 12494.0 Δ lnL 34.6 SE 13.2 **WORSE**
G	Length 2765 Δ length 20 SE 6.5 **WORSE**	-lnL 12497.3 Δ lnL 37.8 SE 13.4 **WORSE**

Figure 6-4. Parsimony topology alternatives tested by parsimony and maximum likelihood. SE is calculated as described in text. Criteria for insignificant or significantly worse are greater than 1.97 SE, or 95% confadence.



Position of Eukaryotes	Parsimony	Maximum Likelihood
A	Length 2778 **BEST**	-lnL 12503.7 **BEST**
B	Length 2786 Δ length 8 SE 3.9 insignificant	-lnL 12517.4 Δ lnL 8.7 SE 5.6 insignificant
C	Length 2791 Δ length 13 SE 5.3 **WORSE**	-lnL 12530.1 Δ lnL 26.4 SE 9.6 **WORSE**
D	Length 2792 Δ length 14 SE 5.9 **WORSE**	-lnL 12543.5 Δ lnL 39.8 SE 12.5 **WORSE**
E	Length 2794 Δ length 16 SE 6.6 **WORSE**	-lnL 12548.7 Δ lnL 45.0 SE 13.7 **WORSE**
F	Length 2795 Δ length 17 SE 7.1 **WORSE**	-lnL 12556.7 Δ lnL 53.0 SE 15.0 **WORSE**
G	Length 2794 Δ length 16 SE 7.1 **WORSE**	-lnL 12560.1 Δ lnL 56.3 SE 15.0 **WORSE**
H	Length 2793 Δ length 15 SE 7.5 **WORSE**	-lnL 12561.4 Δ lnL 57.7 SE 15.1 **WORSE**

Figure 6-5. Neighbor-joining topology alternatives tested by parsimony and maximum likelihood. SE is calculated as described in text. Criteria for insignificant or significantly worse are greater than 1.97 SE, or 95% confadence.

suggests that the eukaryotic gene derives specifically from an alpha-proteobacterium.

Associations between eukaryotes and contemporary alpha-proteobacteria that might encourage some sort of lateral gene transfer are of course common and well known: symbionts such as *Rhizobium,* intracellular parasites such as *Rickettsia* and *Ehrlichia,* and *Agrobacterium,* which has a sophisticated mechanism for transferring certain genes to the nucleus of its eukaryotic host (Zambryski, 1988). However, the TPI gene transfer must have occurred very early in eukaryotic evolution, and it seems simplest to assume that the genome of the alpha-proteobacterial symbiont that became the mitochondrion was the proximal source. Many instances of both ancient and recent organelle-to-nucleus transfer have been identified (Gray, 1992; Baldauf and Palmer, 1990; Covello and Gray, 1992; Nugent and Palmer, 1991) and organelle-specific proteins (including plastid TPI) have on occasion been replaced by their cytosolic counterparts (Bububenko *et al.*, 1994; Brown *et al.*, 1994; Henze *et al.*, 1994; Schmidt *et al.*, 1995), so there is also no *a priori* reason to suppose that organelle-derived genes could not assume a cytosolic role. This has, in fact, been proposed for plastid PGK, where the chloroplast-targeted gene has duplicated and replaced its cytosol-specific counterpart (Brinkmann and Martin, 1996).

If eukaryotic TPI genes are indeed of mitochondrial origin, then we must rethink some aspects of early eukaryote evolution. Trees in Figures 6-2 amd 6-3 include two amitochondrial eukaryotes, *Entamoeba histolytica* and *Giardia lamblia*, which both have TPI genes similar to those of other eukaryotes. Interestingly *Entamoeba* lacks cytologically defined mitochondria and is strictly anaerobic, but genes whose products are normally targeted to the mitochondrion have been identified in its nuclear DNA, providing direct evidence for mitochondrial loss (Clark and Roger, 1995). Similar suggestions have been made for *Giardia*, but on

weaker evidence: from glyceraldehyde-3-phosphate dehydrogenase (Henze *et al.*, 1995) and immuno-crosreactivity with mitochondrial Cpn60 (Soltys and Gupta, 1994). *Giardia* is a member of the Archezoa, a group of organisms that, unlike *Entamoeba*, is thought to have diverged before the acquisition of mitochondria (Cavalier-Smith, 1983). A common origin of eukaryotic TPI genes from the mitochondrion would support the notion that, like *Entamoeba*, *Giardia* and possibly other so-called Archezoa may have once had mitochondria.

Regardless of whether a mitochondrial origin of TPI genes can be more precisely defined, these genes appear at least to have a non-nuclear, proteobacterial provenance. TPI has been used extensively as a model for correlations between protein structure and gene structure in the debate over the origin of spliceosomal introns. It has been claimed that intron positions in TPI genes represent the boundaries between domains or modules in proteins (Gilbert *et al.*, 1986; Tittiger *et al.*, 1993), and that some of these introns are shared between distantly related eukaryotic groups (Gilbert *et al.*, 1986; Gilbert and Glynias, 1993). Opponents of the exon theory of genes have disputed these claims, arguing that intron positions are random with respect to protein structure (Stoltzfus *et al.*, 1994; Logsdon *et al.*, 1995), and that the distribution of introns on phylogenetic grounds is more parsimonious with a model which includes their relatively late insertion (Logsdon *et al.*, 1995; Kwiatowski *et al.*, 1995). If TPI sequences presently encoded in eukaryotes were actually obtained from a eubacterium, then their introns (if ancient) must also have been inherited from that bacterium. This in turn demands that these introns have been lost many times over independently, in many eubacterial lineages. Late intron insertion offers a far more parsimonious view.

Appendix A: Media Formulations (1 litre)

ATCC Culture Medium 111 RHIZOBIUM X MEDIUM

Yeast extract, 1.0 g
Mannitol, 10.0 g
Adjust pH to 7.2.
Soil Extract: 200.0 ml
African violet soil, 77.0 g
Na2CO3, 0.2 g
Distilled water, 200.0 ml
Autoclave for one hour. Filter before using.

LB

Tryptone, 10g
Yeast Extract, 5g
NaCl, 5g

2YT

Bacto-Tryptone, 16 g
Bacto-Yeast Extract, 10 g
NaCl, 10 g

NZY

Bacto-Yeast Extract, 5 g
NaCl, 5 g
$MgSO_4$-$7H_2O$, 2 g
NZ Amine (casein hydrolysate), 10 g

ATCC Culture Medium 1404 KEISTER'S MODIFIED TYI-S-33 MEDIUM

Casein Digest (BBL 97023), 20.0 g
Yeast Extract (BBL 11928), 10.0 g
Dextrose, 10.0 g
Bovine Bile (Sigma B-8381), 0.75 g
NaCl, 2.0 g
L-Cysteine-HCl (Sigma C7880), 2.0 g
Ascorbic Acid (J.T. Baker B581-5), 0.2 g
K_2HPO_4, 1.0 g
KH_2PO_4, 0.6 g
Ferric Ammonium Citrate (Mallinckrodt 0658), 22.8 mg
Adjust pH to 7.0 - 7.2 with 1 N NaOH and filter-sterilize. Aseptically add 100.0 ml heat-inactivated bovine serum.

Halophile Rich Medium

NaCl, 206 g
$MgSO_4$-$7H_2O$, 37 g
KCl, 3.7 g
Bacto-Yeast Extract, 3 g
Bacto-Tryptone, 5 g
1.7 ml of a 75 mg per litre solution of $MnCl_2$
50 ml of 1 M Tris-HCl (pH 7.2)
5 ml of 10% $CaCl_2$-$2H_2O$
Tryptone and yeast extract, sucrose, agar, $CaCl_2$, and basal salt solutions are prepared separately and combined after autoclaving.

Appendix B: Primers used in PCR

Ubiquitin:

UB-A	CGGGATCCCGATGCARATDTTYGTNAA
UB-3	CGGGATCCCTCYTCNARYTGYTTNCC
RUB-1	GYTGRYTCGACKTCGATKGTG
RUB-2	CARGAYARRGAAGGTATTCC

Calmodulin:

CAM-1	TGGGGTACCCAAGATATGATHAAYGARGT
CAM-2	GGACTAGTATCATTTCRTCNACYTCYTC
I-CAM-1	CCTTCATCTGACGGTTCATC
I-CAM-2	CAAGATTACAGCTGCAGAGC

Triosephosphate Isomerase:

TF1†	ACGTCTCGAGTTCGGTGGNAAYTGGAA
TF4†	CGAGAATTCAACGGTGCATTYACNGGNGA
TR1†	ATCTCTAGAAGTGATGCNCCNCCNAC
TR2†	AGCTCTAGACCTGTNCCDATNGCCCA
RhTP2	GGTTTATACCGTTGCTCAGG
RhTP3	GTTAAAGCACAGTTAGATG

Tubulin:

ATUB-A†	TCCGAATTCARGTNGGAAYGCNTGYTGGGA
ATUB-B†	TCCAAGCTTCCATNCCYTCNCCNACRTACCA
BTUB-A†	TCCTGCAGGNCARTGYGGNAAYCA
BTUB-B†	TCCTCGAGTRAAYTCCATYTCRTCAT
SART	CCCAGATGATCTCTGGTATGACTGC
HA550f	CTCACTGCATGCTTGA

Elongation Factor-1 alpha:

EF1F*	CGAGGATCCGTTATTGGNCAYGTNGA
EF7R*	ACGTTGGATCCAACRTTRTCNCC
EF8R*	GGTCGCGACAGTYTGNCTCATRTC
SEF3P	TGATGCCATCGACGGACTCAAGGC
HSEF3P	GACAAGCCACTCCGTCTCCCA

Glutamine tRNA:

Q-F	GGTACCGGKYCYATGGYSTARTGGTA
Q-R	GGTACCGGGCCYRSYSGGATTCGAAC

† Courtesy of A.J.Roger

* Courtesy of S.L.Baldauf

Appendix C: Taxonomy of Species Used

Parabasalia

Eukaryotae; Parabasalidea; Trichomonadida;
Trichomonas vaginalis
Monocercomonas sp.
Trichomitus batrachorum
Tritrichomonas foetus

Diplomonads

Eukaryotae; Diplomonadida; Hexamitidae;
Giardia lamblia
Hexamita inflata
Hexamita ATCC50330
Hexamita ATCC50380
Spironucleus muris

Microsporidia

Eukaryotae; Microsporidia; Microsporea; Microsporida; Pansporablastina;
Spraguea lophii
Eukaryotae; Microsporidia; Microsporea; Microsporida; Apansporoblastina;
Nosematidae; *Nosema locustae*
Eukaryotae; Microsporidia; Microsporea; Microsporida; Apansporoblastina;
Unikaryonidae; *Encephalitozoon hellum*

Heterlobosea

Eukaryotae; Mitochondrial Eukaryotes; Acrasida; *Acrasis rosea*
Eukaryotae; Mitochondrial Eukaryotes; Schizopyrenida; Vahlkampfiidae; *Naegleria fowleri*

Archaea

Archaea; Euryarchaeota; Halobacteriales; Halobacteriaceae;
Haloferax volcanii
Haloarcula hispanica

Eubacteria

Eubacteria; Chloroflexaceae/Deinococcaceae group; Chloroflexaceae;
Chloroflexaceae;*Chloroflexus aurantiacus*
Eubacteria; Cyanobacteria; Prochlorophytes; Prochloroaceae; *Prochloron* sp.
Eubacteria; Proteobacteria; alpha subdivision; Rhizobiaceae; *Rhizobium etli*
Eubacteria; Proteobacteria; alpha subdivision; Rickettsiaceae; *Rickettsia prowazekii*
Eubacteria; Proteobacteria; alpha subdivision; Rhizobiaceae; *Agrobacterium radiobacter*
Eubacteria; Proteobacteria; epsilon subdivision; *Helicobacter pylori*
Eubacteria; Proteobacteria; delta subdivision; Francisella group; *Francisella tularensis*

Appendix D: GenBank Submissions.

Ubiquitin:

U27577 - *Trichomonas vaginalis* polyubiquitin (UbA).

U28008 - *Trichomonas vaginalis* ubiquitin 1A (Ub1A).

U28009 - *Trichomonas vaginalis* ubiquitin 1C (Ub1C).

U28010 - *Trichomonas vaginalis* ubiquitin 1D (Ub1D).

U28011 - *Trichomonas vaginalis* ubiquitin 1E (Ub1E).

U28012 - *Trichomonas vaginalis* ubiquitin dimer 2B (Ub2B).

U28013 - *Trichomonas vaginalis* polyubiquitin junction JC (UbJC).

Calmodulin:

U38787 - *Naegleria fowleri* calmodulin (CAM).

U38788 - *Acrasis rosea* calmodulin (CAM).

U38786 - *Trichomonas vaginalis* calmodulin and E2 ubiquitin-conjugating enzyme.

Alpha-Tubulin:

U29440 - *Hexamita* sp. 50330 alpha-tubulin.

U30664 - *Hexamita* sp. 50330 alpha-tubulin.

U37080 - *Hexamita inflata* alpha-tubulin.

U37079 - *Spironucleus muris* alpha-tubulin.

Beta-Tubulin:

U29441 - *Hexamita* sp. 50330 beta-tubulin.

Elongation Factor-1 alpha:

U29442 - *Hexamita* sp. 50330 elongation factor-1 alpha.

U37081 - *Hexamita inflata* elongation factor-1 alpha.

U37078 - *Spironucleus muris* elongation factor-1 alpha.

Triosephosphate Isomerase:

U31597 - *Helicobacter pylori* triosephosphate isomerase (TPI).

Glucose-6-Phosphate Isomerase:

U31596 - *Helicobacter pylori* glucose 6-phosphate isomerase (GPI).

Appendix E: Spurious amplification products with recognisable similarity to known genes.

R. prowazekii TPI 4-1.16 (187 bp) with hit to *H. influenzae* ORF HI0056 (also hits unidentified ORFs in *E. coli* and *Klebsiella pneumoniae*).

```
Rp4-1.16  187 FITVMVYKLPKHQQNKAMILGLGLAMIARIGLLGSLFFISHLQKPLFAIAGMSFSWRDVVLL 1
                FI ++V +LP+ Q+     ILGL LAM+ RI LL SL +I  L  PLF +      S RD++LL
HI0056     31 FINILVGRLPERQRQSGRILGLALAMLTRILLLMSLAWIMKLTAPLFTVFNQEISGRDLILL 92
```

R. prowazekii TPI 1-1.2 (136 bp) with hit to *E. coli* UDP-N-acetylglucosamine pyrophosphorylase (also hits same gene in numerous proteobacteria).

```
Rp1-1.2 45 MIRNDANNQIIILAAGKGTRMESDLPKVMHK 136
              +    ++ILAAGKGTRM SDLPKV+H
Ec UAGP     1 MLNNAMSVVILAAGKGTRMYSDLPKVLHT 29
```

Agrobacterium radioresistans TPI 1-1.2 (50 bp) with hit to *Aulonocarabus kurilensis* mitochondiral gene, NADH dehydrogenase subunit 5 (also hits same gene from other eukaryotes).

```
Ar1-1.2     2 YLLLISTFILLALLIL 50
              YL+LI  +I++ LLIL
NADH5     338 YLVLIILWIIILLLIL 353
```

Prochloron sp. TPI 1-1.1 (145 bp) with hit to *Synechocystis* sp. *fts*H (also hits same gene from other eubacteria).

```
Ps1-1.1    10 VLTERARNMVTRFGMSDLGPVALENGNNQVFLSNNWMNRAEYSEEIA 140
              +TE AR MVTRFGMS+LGP++LE+   +VFL    MNR+EYSEE+A
Ss ftsH   514 QVTEMARQMVTRFGMSNLGPISLESSGGEVFLGGGLMNRSEYSEEVA 560
```

Helicobacteri pylori glucose 6-phosphate isomerase aligned with that of *E. coli*.

```
EcGPI LVDYSKNRITEETLAKLQDLAKECDLAGAIKSMFSGEKINRTENRAVLHVALRNRSNTPILVDGKDVMPEVNA
       +DYSKNR+ + TL  L +LA +C L   I +MF GEKIN TE RAVLH ALR+ ++T IL+D  +V   V +
HpGPI SLDYSKNRLNDTTLKLLFELANDCSLKEKIEAMFKGEKINTTEKRAVLHTALRSLNDTEILLDNMEVLKSVRS

EcGPI VLEKMKTFSEAIISGEWKGYTGKAITDVVNIGIGGSDLGPYMVTEALRPY.KNHLNMHFVSNVDGTHIAEVLK
      VL  M  FS+++ SG   GYT   ITD+VNIGIGGSDLG  MV  AL  Y    L MHFVSNVDGT I +VL
HpGPI VLKRMRAFSDSVRSGKRLGYTNQVITDIVNIGIGGSDLGALMVCTALKRYAHPRLKMHFVSNVDGTQILDVLE

EcGPI KVNPETTLFLVASKTFTTQETMTNAHSARDWFLKAAGDEKHVAKHFAALSTNAKAVGEFGIDTANMFEFWDW
      K+ P  TLF-VASKTF TQET+TNA  AR WF+  +GDEKH+AKHF A+STN  AV  FGID  NMFEFWD
HpGPI KLSPASTLFIVASKTFSTQETLTNALTARKWFVERSGDEKHIAKHFVAVSTNKEAVQQFGIDEHNMFEFWDF
```

References

Adachi, J. and Hasegawa, M. (1992) *Computer science monographs, No. 27. MOLPHY: programs for molecular phylogenetics, I. - PROTML: maximum likelihood inference of protein phylogeny*. Institute of Statistical Mathematics, Tokyo.

Alam, M. and Oesterhelt, D. (1984) Morphology, function, and isolation of halobacterial flagella. *J. Mol. Biol.* **176**, 459-475.

Allardet-Servent, A., Michaux-Charachon, S., Jumas-Bilak, E., Karayan, L. and Ramuz, M. (1993) Presence of one linear and one circular chromosome in the *Agrobacterium tumefaciens* C58 genome. *J. Bacteriol.* **175**, 7869-7874.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Amiri, K. A. (1994) Fibrillarin-like proteins occur in the domain archaea. *J. Bacteriol.* **176**, 2124-2127.

Arndt, E., Kromer, W. and Hatakeyama, T. (1990) Organization and nucleotide sequence of a gene cluster coding for eight ribosomal proteins in the archaebacterium *Halobacterium marismortui*. *J. Biol. Chem.* **265**, 3034-3039.

Bailey, C. C. and Bott, K. F. (1994) An unusual gene containing a *dnaJ* N-terminal box flanks the putative origin of replication in *Myycoplasma genetalium*. *J. Bacteriol.* **176**, 5814-5819.

Baker, R. T. and Board, P. G. (1989) Unequal crossover generates variation in ubiquitin coding unit number at the human *UbC* polyubiquitin locus. *Am. J. Hum. Genet.* **44**, 534-542.

Baldauf, S. L. and Palmer, J. D. (1990) Evolutionary transfer of the chloroplast *tuf*A gene to the nucleus. *Nature* **344**, 262-265.

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

Baldauf, S. L., Palmer, J. D. and Doolittle, W. F. (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**, 7749-7754.

Ball, E., Karlik, C. C., Beall, C. J., Saville, D. L., Sparrow, J. C., Bullard, B. and Fyrberg, E. A. (1987) Arthrin, a myofibrillar protein of insect flight muscle, is an actin-ubiquitin congugate. *Cell* **51**, 221-228.

Bartig, D., Lmekemeier, K., Frank, J., Lottspeich, F. and Klink, F. (1992) The archaebacterial hypusine-containing protein. *Eur. J. Biochem.* **204**, 751-758.

Bi, E. F. and Lutkenhaus, J. (1991) FtsZ ring structure associated with division in *Escherichia coli*. *Nature* **354**, 161-164.

Björk, G. R. (1995) Biosynthesis and function of modified nucleosides. In *tRNA Structure, Biosynthesis and Function*. D. Söll and U. RajBhandary, Ed., Wachington, DC, American Society for Microbiology. 165-205.

Bork, P., Sander, C. and Valencia, A. (1992) An ATPase domaine common to prokaryotic cell cycle proteins, sugar kinases, and hsp70 heat shock proteins. *Proc. Natl. Acad. Sci. USA* **89**, 7290-7294.

Braithwaite, D. K. and Ito, J. (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res.* **21**, 787-802.

Bramhill, D. and Kornberg, A. (1988) Duplex opening by DnaA protein at novel sequences in initiation of replication at the origin of the *E. coli* chromosome. *Cell* **52**, 143-755.

Bramhill, D. and Thompson, C. M. (1994) GTP-dependent polymerization of *Escherichia coli* FtsZ protein to form tubules. *Proc. Natl. Acad. Sci. USA* **91**, 5813-5817.

Brenner, D. J., Fanning, G. R., Johnson, K. E., Citarella, R. V. and Falkow, S. (1969) Polynucleotide sequence relationships among members of Enterobacteriaceae. *J. Bacteriol.* **98**, 637-650.

Brewer, B. J. (1993) Intergenic DNA and the sequence requirements for replication initiation in eukaryotes. *Curr. Opin. Genet. Dev.* **4**, 196-202.

Brinkmann, H. and Martin, W. (1996) Higher plant chloroplast and cytosolic 3-phosphoglycerate kinases: A case of endosymbiotic gene replacement. *Plant Mol. Biol.* **30**, 65-75.

Brown, J. R. and Doolittle, W. F. (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., Daniels, C. J. and Reeve, J. N. (1989) Gene structure, organization, and expression in archaebacteria. *CRC Crit. Rev. Microbiol.* **16**, 287-338.

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

Brown, J. W. and Reeve, J. N. (1985) Polyadenylated, noncapped RNA from the archaebacterium *Methanococcus vannielii*. *J. Bacteriol.* **162**, 909-917.

Bubunenko, M. G., Schmidt, J. and Subramanian, A. R. (1994) Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. *J. Mol. Biol.* **240**, 28-41.

Burch, T. J. and Haas, A. L. (1994) Site-directed mutagenesis of ubiquitin. Differential roles for arginine in the interaction wiht ubiquitin-activating enzyme. *Biochemistry* **33**, 7300-7308.

Burhans, W. C. and Huberman, J. A. (1994) DNA replication origins in animal cells: a question of context? *Science* **263**, 639-640.

Burland, V., III, G. P., Daniels, D. L. and Blatner, F. R. (1993) DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics* **15**, 551-561.

Burns, R. G. (1991) Alpha-, beta-, and gamma-tubulins: sequence comparisons and structural constraints. *Cell Motil. Cytoskeleton* **20**, 181-189.

Burns, R. G. (1995) Identification of two new members of the tubulin family. *Cell Motil. Cytoskeleton* **31**, 255-258.

Bushman, J. L., Asuru, A. I., Matts, R. L. and Hinnebusch, A. G. (1993) Evidence that GCD6 and GCD7, translational regulators of *GCN4*, are subunits of the guanosine nucleotide exchange factor for eIF-2 in *Saccharomyces cerevisiae*. *Molec. Cell Biol.* **13**, 1920-1932.

Calcutt, M. J. and Schmidt, F. J. (1992) Conserved gene arrangement in the origin region of the *Streptomyces coelicolor* chromosome. *J. Bacteriol.* **174**, 3220-3226.

Canning, E. U. (1988) Nuclear division and chromosome cycle in microsporidia. *Biosystems* **21**, 333-340.

Canning, E. U. (1990) Phylum Microsporidia. In *Handbook of Protoctista*. L. Margulis, J. O. Corliss, M. Melkonian and D. J. Chapman, Ed., Boston, Jones and Bartlett Publishers. 53-72.

Cao, G.-J. and Sarkar, N. (1992) Poly(A) RNA in *Escherichia coli*: nucleotide sequence at the junction of the *lpp* transcript and the polyadenylate moiety. *Proc. Natl. Acad. Sci. USA* **89**, 7546-7550.

Caskey, C. T., Tompkins, R., Scolnick, E., Caryk, T. & Nirenberg, M. 1968. Sequential translation of trinucleotide codons for the initiation and termination of protein synthesis. *Science* **162**:135-138.

Cavalier-Smith, T. (1983) A o-kingdom classification and a unified phylogeny. In *Endocytobiology*, W. Schwemmler and H. E. A. Schenk, Ed., Berlin, Walter de Gruyter & Co. 1027-1034.

Cavalier-Smith, T. (1991) Cell evolution. In *Evolution of Life*, S. Osawa and T. Honjo, Eds., Tokyom Springer-Vrelag. 271-304.

Cavalier-Smith, T. (1993) The kingdom Protozoa and its 18 phyla. *Microbiol. Rev.* **57**, 953-994.

Charlebois, R. L., Schalkwyk, L. C., Hofman, J. D. and Doolittle, W. F. (1991) Detailed physical map and set of overlapping clones covering the genome of the archaebacterium *Haloferax volcanii* DS2. *J. Mol. Biol.* **222**, 509-524.

Chen, C. W., Yu, T. W., Lin, Y. S., Kieser, H. M. and Hopwood, D. A. (1993) The conjugative plasmid SLP2 of *Streptomyces lividans* is a 50 kb linear molecule. *Mol. Microbiol.* **7**, 925-932.

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

Cline, S. W. and Doolittle, W. F. (1992) Transformation of members of the genus *Haloarcula* with shuttle vectors based on *Halobacterium halobium* and *Haloferax volcanii* plasmid replicons. *J. Bacteriol.* **174**, 1076-1080.

Coen, E. S.. Thoday, J. M. and Dover, G. (1982) Rate of turnover of structural variants in the rDNA gene family of *Drosophila melanogaster*. *Nature* **295**, 564-568.

Copeland, H. F. (1938) The kingdoms of organisms. *Quart. Rev. Biol.* **13**, 383-420.

Copeland, H. F. (1947) Progress report on basic classification. *Amer. Nature.* **81**, 340-361.

Covello, P. S. and Gray, M. W. (1992) Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: evidence for RNA-mediated gene transfer. *EMBO J.* **11**, 3815-20.

Cregg, J. M., Barringer, K. J., Hessler, A. Y. and Madden, K. R. (1985) *Pichia pastoris* as a host system for transformations. *Mol. Cell. Biol.* **5**, 3376-3385.

Crespi, M., Messens, E., Caplan, A. B., van Montagu, M. and Desomer, J. (1992) Fasciation induction by the phytopathogen *Rhodococcus fascians* depends upon a linear plasmid encoding a cytokinin synthase gene. *EMBO J.* **11**, 795-804.

Davie, J. R. and Murphy, L. C. (1990) Level of ubiquitinated histone HB2 in chromatin is coupled to ongoing transcription. *Biochemistry* **29**, 4752-4757.

Delwiche, C. F., Kuhsel, M. and Palmer, J. D. (1995) Phylogenetic analysis of tufA sequences indicates a cyanobacterial origin of all plastids. *Mol. Phylogenet. Evol.* **4**, 110-128.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensives set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12**:387-395.

Doi, R. H. and Igarashi, R. T. (1965) Conservation of ribosomal and messenger ribonucleic acid cistrons in *Bacillus* species. *Proc. Natl. Acad. Sci. USA* **90**, 386-390.

Donachie, W. D. (1993) The cell cycle of Escherichia coli. *Annu. Rev. Microbiol.* **47**, 199-230.

Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111-117.

Dunn, R., McCoy, J., Simsek, M., Majumdar, A., Chang, S. H., RajBhandary, U. L. and Korana, H. G. (1981) The bacteriorhodopsin gene. *Proc. Natl. Acad. Sci. USA* **78**, 6744-6748.

Durner, J. and Börger, P. (1995) Ubiquitin in the prokaryote *Anabaena variabilis*. *J. Biol. Chem.* **270**, 3720-3725.

Durovic, P. and Dennis, P. P. (1994) Separate pathways for excision and processing of 16S and 23S rRNA from the primary rRNA operon transcript from the hyperthermophilic archaebacterium *Sulfolobus acidocaldarius*: similarities to eukaryotic rRNA processing. *Mol. Microbiol.* **13**, 229-242.

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

Eisen, J. A. (1996) 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.

Erickson, H. P. (1995) FtsZ, a prokaryotic homolog of tubulin? *Cell* **80**, 367-370.

Erickson, H. P., Taylor, D. W., Taylor, K. A. and Bramhill, D. (1996) Bacterial cell division protein FtsZ assembles into protofilament sheets and minirings, structural homologs of tubulin polymers. *Proc. Natl. Acad. Sci. USA* **93**, 519-523.

Faguy, D. M., Jarrell, K. F., Kuzio, J. and Kalmokoff, M. L. (1994) Molecular analysis of archaeal flagellins: similarity to the type IV pillin-transport superfamily widespread in bacteria. *Can. J. Microbiol.* **40**, 67-71.

Falkow, S. (1965) Nucleic acids, genetic exchange and bacterial speciation. *Am. J. Med.* **39**, 753-765.

Falkow, S. and Formal, S. B. (1969) Restriction in genetic crosses between *Escherichia coli* and *Shigella flexneri*. *J. Bacteriol.* **100**, 540-541.

Feierabend, J., Kurzok, H. G. and Schmidt, M. (1990) Genetics and evolution of chloroplast isozymes of triosephosphate isomerase. *Prog. Clin. Biol. Res.* **344**, 665-682.

Felsenstein, J. (1985) Confidence limits on phylogenies with a molecular clock. *Systematic Zoology* **34**, 152-161.

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

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

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C. A., Gocayne, J. D., Scott, J. D., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., McDonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C.

(1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Fox, G. E., Marum, L. M., Balch, W. E., Wolfe, R. S. and Woese, C. R. (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc. Natl. Acad. Sci. USA* **74**, 4537-4541.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J. L., Nguyen, D. T., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A. I. and Venter, J. C. (1995) The minimal gene complememt of *Mycoplasma genitalium*. *Science* **270**, 397-403.

Frolova, L., Goff, X. L., Rasmussen, H. H., Cheperegin, S., Drugeon, G., Kress, M., Arman, I., Haenni, A.-L., Cells, J. E., Phillippe, M., Justesen, J. and Kisselev, L. (1994) A highly conserved eukaryotic protein family possing properties of polypeptide chain release factor. *Nature* **372**, 701-703.

Früh, K., Gossen, M., Wang, K., Bujard, H., Peterson, P. A. and Yang, Y. (1994) Displacement of housekeeping proteasome subunits by MHC-encoded LMPs: a newly discovered mechanism for modulating the multicatalytic proteinase complex. *EMBO J.* **13**, 3236-3244.

Fujita, M. Q., Yoshikawa, H. and Ogasawara, N. (1990) Structure of the *dnaA* region of *Micrococcus luteus*: conservation and variations among the eubacteria. *Gene* **93**, 73-78.

Fujita, M. Q., Yoshikawa, H. and Ogasawara, N. (1992) Structure of the *dnaA* region in the *Mycoplasma capricolum* chromosome: conservation and variations in the course of evolution. *Gene* **110**, 17-23.

Funnell, B. E., Baker, T. A. and Kornberg, A. (1987) *In vitro* assembly of a prepriming complex at the origin of the *Escherichia coli* chromosome. *J. Biol. Chem.* **262**, 10327-10334

Gaertig, J., Cruz, M. A., Bowen, J., Gu, L., Pennock, D. G. and Gorovsky, M. A. (1995) Acetylation of lysine 40 in alpha-tubulin is not essential in *Tetrahymena thermophila*. *J. Cell Biol.* **129**, 1301-1310.

Gard, D. L. (1994) Gamma-tubulin is asymmetrically distributed in the cortex of *Xenopus* oocytes. *Dev. Biol.* **161**, 131-140.

Gilbert, W. and Glynias, M. (1993) On the ancient nature of introns. *Gene* **135**, 137-144.

Gilbert, W., Marchionni, M. and McKnight, G. (1986) On the antiquity of introns. *Cell* **46**, 151-154.

Gish, W. and States, D. J. (1993) Identification of protein coding regions by database similarity search. *Nat. Genet.* **3**, 266-272.

Goebl, M. G., Yochem, J., Jentsch, S., McGrath, J. P., Varshavsky, A. and Byers, B. (1988) The yeast cell cycle gene CDC34 encodes a ubiquitin-conjugating enzyme. *Science* **241**, 1331-1335.

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

Goldknopf, I. L. and Busch, H. (1977) Isopeptide linkage between nonhistone and histone 2A polypeptides of chromosomal conjugate-protein A24. *Proc. Natl. Acad. Sci. USA* **74**, 864-868.

Goldstein, G., Scheid, M., Hammerling, U., Boyse, E. A., Schlesinger, D. H. and Niall, H. D. (1975) Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. *Proc. Natl. Acad. Sci. USA* **72**, 11-15.

Gonzalez, I. L., Petersen, R. and Sylvester, J. E. (1989) Independent insertion of Alu elements in the human ribosomal spacer and their concerted evolution. *Mol. Biol. Evol.* **6**, 413-423.

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

Grohmann, L., Brennicke, A. and Schuster, W. (1992) The mitochondrial gene encoding ribosomal protein S12 has been translocated to the nuclear genome in *Oenothera*. *Nucleic Acids Res.* **20**, 5641-5646.

Gunderson, J., Hinkle, G., Leipe, D., Morrison, H. G., Stickel, S. K., Odelson, D. A., Breznak, J. A., Nerad, T. A., Muller, M. and Sogin, M. L. (1995) Phylogeny of trichomonads inferred from small-subunit rRNA sequences. *J. Eukaryot. Microbiol.* **42**, 411-415.

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

Hanner, M., Mayer, C., Köhrer, C., Golderer, G., Gröbner, P. and Piendl, W. (1994) Autogenous translational regulation of the ribosomal MvaL1 operon in the archaebacterium *Methanococcus vanielii*. *J. Bacteriol.* **176**, 409-418.

Hanyu, N., Kuchino, Y. and Nishimura, S. (1986) Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* $tRNAs^{Gln}$. *EMBO J.* **5**, 1307-1311.

Hasegawa, M. and Fujiwara, M. (1993) Relative efficiencies of the maximum likelihood, maximum parsimony and neighbor joining methods for estimating protein phylogeny. *Mol. Phyl. Evol.* **2**, 1-5.

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

Hausner, W., Frey, G. and Thomm, M. (1991) Control regions of an archaeal gene. *J. Mol. Biol.* **222**, 495-508.

Heizmann, C. W. and Hunziker, W. (1991) Intracellular calcium-binding proteins: more sites than insights. *Trends Biochem. Sci.* **16**, 98-103.

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

Henze, K., Schnarrenberger, C., Kellermann, J. and Martin, W. (1994) Chloroplast and cytosolic triosephosphate isomerases from spinach: purification, microsequencing and cDNA cloning of the chloroplast enzyme. *Plant Mol. Biol.* **26**, 1961-1973.

Herreros, E., Garcia-Saez, M. I., Nombela, C. and Sanchez, M. (1992) A reorganized *Candida albicans* DNA sequence promoting homologous non-integrative genetic transformation. *Mol. Microbiol.* **6**, 3567-3574.

Hinkle, G. & Sogin, M. L. 1993. The evolution of the Vahlkampfiidae as deduced from 16S-like ribosomal RNA analysis. *J. Eukaryot. Microbiol.* **40**:599-603.

Holloway, S. L., Glotzer, M., King, R. W. and Murray, A. W. (1993) Anaphase is initiated by proteolysis rather than by the inactivation of a maturation-promoting factor. *Cell* **73**, 1393-1402.

Holmes, M. L. and Dyall-Smith, M. L. (1991) Mutations in DNA gyrase result in Novobiocin resistance in halophilic archaebacteria. *J. Bacteriol.* **173**, 642-648.

Hüdepohl, U., Reiter, W.-D. and Zillig, W. (1990) *In vitro* transcription of two rRNA genes of the archaebacterium *Sulfolobus* sp.B12 indecates a factor requirement for specific initiation. *Proc. Natl. Acad. Sci. USA* **87**, 5851-5855.

Huet, J., Schnabel, R., Sentenac, A. and Zillig, W. (1983) Archaebacteria and eukaryotes posses DNA-dependent RNA polymerases of a common type. *EMBO J.* **2**, 1291-1294.

Ito, J. and Braithwaite, D. K. (1991) Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res.* **19**, 4045-4057.

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

Jacob, F. (1977) Evolution and tinkering. *Science* **196**, 1161-1166.

Jentsch, S. (1992) The ubiquitin-conjugation system. *Annu. Rev. Genet.* **26**, 179-207.

Jentsch, S., McGrath, J. P. and Varshavsky, A. (1987) The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme. *Nature* **329**, 131-134.

Johnson, L. H. and Barker, D. G. (1987) Characterization of an autonomously replicating sequence from the fission yeast, *Schizosaccharomyces pombe*. *Mol. Gen. Genet.* **207**, 161-164.

Kamaishi, T., Hashimoto, T., Nakamura, Y., Nakamura, F., Murata, S., Okada, N., Okamoto, K.-I., Shimzu, M., and Hasegawa, M. (1996) Protein phylogeny of translation elongation factor EF-1α suggests Microsporidians are extremely ancient eukaryotes. *J. Mol. Evol.* **42**, 257-263.

Kang, H. A. and Hershey, J. B. (1994) Effect of initiation factor eIF-5A depletion on protein synthesis and proliferation of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **269**, 3934-3940.

Katiyar, S. K. and Edlind, T. D. (1994) Beta-tubulin genes of *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* **64**, 33-42.

Keeling, P. J. and Doolittle, W. F. (1995) An archaebacterial eIF-1A: new grist for the mill. *Mol. Microbiol.* **17**, 399-400.

Keeling, P. J. and Doolittle, W. F. (1995) Archaea: narrowing the gap between prokaryotes and eukaryotes. *Proc. Natl. Acad. Sci. USA* **92**, 5761-5764.

Keeling, P. J., Bauldauf, S. L., Doolittle, W. F., Zillig, W. and Klenk, H.-P. (1996) An *inf*B-homologue in *Sulfolobus acidocaldarius*. *System. App. Microbiol.*, in press.

Keeling, P. J., Charlebois, R. L. and Doolittle, W. F. (1994) Archaebacterial genomes: eubacterial form and eukaryotic content. *Curr. Opin. Genet. Dev.* **4**, 816-822.

Kelly, A., Powis, S. H., Glynne, R., Radley, E., Beck, S. and Trowsdale, J. (1991) Second proteasome-related gene in the human MHC class II region. *Nature* **353**, 667-668.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitution through comparitive studies of nucleotide sequences. *J. Mol. Evol.* **17**, 110-113.

Kirk-Mason, K. E., Turner, M. J. and Chakraborty, P. R. (1988) Cloning and sequence of beta tubulin cDNA from *Giardia lamblia*. *Nucleic Acids Res.* **16** 2733.

Kishino, H. & Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in the Hominoidea. *J. Mol. Evol.* **29**, 170-179.

Klee, C. B., Crouch, T. H. and Richman, P. G. (1980) Calmodulin. *Annu. Rev. Biochem.* **49**, 489-515.

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

Kletzin, A. (1992) Molecular characterisation of a DNA ligase of the extremely thermophilic archaeon *Desulfurolobus ambivalens* shows close phylogenetic relationship to eukaryotic ligases. *Nucleic Acids Res.* **20**, 5389-5396.

Kohlhoff, M., Dahm, A. and Hensel, R. (1996) Tetrameric triosephosphate isomerase from hyperthermophilic Archaea. *FEBS Lett.* **383**, 245-250.

Koonin, E. V., Tatusov, R. L. and Rudd, K. E. (1995) Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications. *Proc. Natl. Acad. Sci. USA* **92**, 11921-11925.

Korman, S. H., Le Blancq, S. M., Deckelbaum, R. J. and Van der Ploeg, L. H. (1992) Investigation of human giardiasis by karyotype analysis. *J. Clin. Invest.* **89**, 1725-1733.

Kozak, M. (1983) Comparison of initiation of protien synthesis in procaryotes, eucaryotes, and organelles. *Micorbiol. Rev.* **47**, 1-45.

Kozminski, K. G., Diener, D. R. and Rosenbaum, J. L. (1993) High level expression of nonacetylatable alpha-tubulin in *Chlamydomonas reinhardtii*. *Cell Motil. Cytoskeleton* **25**, 158-170.

Krebber, H., Wöstmann, C. and Bakker-Grunwald, T. (1994) Evidence for the existance of a single ubiquitin gene in *Giardia lamblia*. *FEBS Lett.* **343**, 234-236.

Kwiatowski, J., Krawczyk, M., Kornacki, M., Bailey, K. and Ayala, F. J. (1995) Evidence against the exon theory of genes derived from the triose- phosphate isomerase gene. *Proc. Natl. Acad. Sci. USA* **92**, 8503-8506.

Lai, E. Y., Remillard, S. P. and Fulton, C. (1988) The alpha-tubulin gene family expressed during cell differentiation in *Naegleria gruberi*. *J. Cell Biol.* **106**, 2035-2046.

Lajoie-Mazenc, I., Tollon, Y., Detraves, C., Julian, M., Moisand, A., Gueth-Hallonet, C., Debec, A., Salles-Passador, I., Puget, A., Mazarguil, H., Raynaud-Messina, B. and Wright, M. (1994) Recruitment of antigenic gamma-tubulin during mitosis in animal cells: presence of gamma-tubulin in the mitotic spindle. *J. Cell Sci.* **107**, 2825-2837.

Lanzendörfer, M., Palm, P., Grampp, B., Peattie, D. A. and Zillig, W. (1992) Nucleotide sequence of the gene encoding the largest subunit of the DNA-dependent RNA polymerase III of *Giardia lamblia*. *Nucleic Acids Res.* **20**, 1145.

Lawson, F. S., Charlebois, R. L., and Dillon, J.-A. R. (1996) Phylogenetic analysis of carbamoylphosphate synthetase genes: evolution involving multiple gene duplications, gene fusions, and insertions and deletions of surrounding sequences. *Mol. Biol. Evol.* in press.

Le Blancq, S. M., Kase, R. S. and Van der Ploeg, L. H. (1991) Analysis of a *Giardia lamblia* rRNA encoding telomere with [TAGGG]n as the telomere repeat. *Nucleic Acids Res.* **19**, 5790.

Le Blancq, S. M., Korman, S. H. and Van der Ploeg, L. H. (1991) Frequent rearrangements of rRNA-encoding chromosomes in *Giardia lamblia*. *Nucleic Acids Res.* **19**, 4405-4412.

Leipe, D. D., Gunderson, J. H., Nerad, T. A. and Sogin, M. L. (1993) Small subuint ribosomal RNA of *Hexamita inflata* and the quest for the first branch of the eukaryotic tree. *Mol. Biochem. Parasitol.* **59**, 41-48.

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

Longstaff, M., Raines, C. A., McMorrow, E. M., Bradbeer, J. W. and Dyer, T. A. (1989) Wheat phosphoglycerate kinase: evidence for recombination between the genes for the chloroplastic and cytosolic enzymes. *Nucleic Acids Res.* **17**, 6569-6580.

Ludwig, W. and Schleifer, K. H. (1994) Bacterial phylogeny based on 16S and 23S rRNA sequence analysis. *FEMS Microbiol. Rev.* **1527**, 155-173.

Lupas, A., Zwickl, P. and Baumeister, W. (1994) Proteasome sequences in eubacteria. *Trends Biochem.***19**, 533-534.

Luria, S. E. and Burrous, J. W. (1957) Hybridization between *Escherichia coli* and *Shigella*. *J. Bacteriol.* **74**, 461-476.

Lutkenhaus, J. (1993) FtsZ ring in bacterial cytokinesis. *Mol. Microbiol.* **9**, 403-409.

Marahrens, Y. and Stillman, B. (1992) A Yeast chromosomal origin of DNA replication defined by multiple functional elements. *Science* **255**, 817-823.

Marczynski, G. T. and Shapiro, L. (1992) Cell cycle control of a cloned chromosomal origin of replication from *Caulobacter crescentus*. *J. Mol. Biol.* **226**, 959-977.

Margolin, W., Wang, R. and Kumar, M. (1996) Isolation of an *ftsZ* homologue from the archaebacterium *Halobacterium salinarium*: Implications for the evolution of FtsZ and tubulin. *J. Bacteriol.* **178**, 1320-1327.

Margulis, L. (1981) *Symbiosis in cell evolution*, W. H. Freeman and Co, San Francisco.

Marmur, J., Seaman, E. and Levine, J. (1962) Interspecific transformation in *Bacillus*. *J. Bacteriol.* **85**, 461-467.

Marsh, T. L., Reich, C. I., Whitlock, R. B. and Olsen, G. J. (1994) Transcription factor IID in the archaea: sequences in the *Thermococcus celer* genome would

encode a product closely related to the TATA-binding protein of eukaryotes. *Proc. Natl. Acad. Sci. USA* **91**, 4180-4184.

Matsuoka, M., Matsubara, M., Daidoh, H., Imanaka, T., Uchida, K. and Aiba, S. (1993) Analysis of regions essential for the function of chromosomal replicator sequences from *Yarrowia lipolytica*. *Mol. Gen. Genet.* **237**, 327-333.

Maundrell, K., Hutchison, A. and Shall, S. (1988) Sequence analysis of ARS elements in fission yeast. *EMBO J.* **7**, 2203-2209.

Maundrell, K., Wright, A. P. H. and Shall, S. (1985) Evaluation of heterologous ARS activity in *S. cerevisiae* using cloned DNA from *S. pombe*. *Nucleic Acids Res.* **13**, 3711-3721.

McCarthy, B. J. and Bolton, E. T. (1963) An approach to the measurement of genetic relatedness among organisms. *Proc. Natl. Acad. Sci. USA* **50**, 160-164.

McCarthy, J. E. G. and Gualerzi, C. (1990) Translational control of prokaryotic gene expression. *Trends Genet.* **6**, 78-85.

Means, A. R. and Dedman, J. R. ('1980) Calmodulin--an intracellular calcium receptor. *Nature* **285**, 73-77.

Merrick, W. C. (1992) Mechanism and regulation of eukaryotic protein synthesis. *Microbiol. Rev.* **56**, 291-315.

Mescher, M. F. and Strominger, J. L. (1976) Structural (shape-maintaining) role of the cell surface glycoproteins of *Halobacterium slainarium*. *Proc. Natl. Acad. Sci. USA* **73**, 2687-2691.

Michelson, A. M. and Orkin, S. H. (1983) Boundries of gene conversion within the duplicated human alpha-globin genes. Concerted evolution by segmental recombination. *J. Biol. Chem.* **258**, 15245-15254.

Miyata, M., Sano, K.-I., Okada, R. and Fukumura, T. (1993) Mapping of replication initiation site in *Mycoplasma capricolum* genome by two-dimensional gel-electrophoretic analysis. *Nucleic Acids Res.* **21**, 4816-4823.

Moriya, S., Atlung, T., Hansen, F. G., Yoshikawa, H. and Ogasawara, N. (1992) Cloning of an autonomously replicating sequence (ARS) from the *Bacillus subtilis* chromosome. *Mol. Microbiol.* **6**, 309-315.

Moriya, S., Firshein, W., Yoshikawa, H. and Ogasawara, N. (1994) Replication of a *Bacillus subtilis oriC* plasmid *in vitro*. *Mol. Microbiol.* **12**, 469-478.

Moriya, S., Fukuoka, T., Ogasawara, N. and Yoshikawa, H. (1988) Regulation of initiation of the chromosomal replication origin by DnaA-boxes in the origin region of the *Bacillus subtilis* chromosome. *EMBO J.* **7**, 2911-2917.

Mulisch, M. (1993) Chitin in protistan organisms. *Europ. J. Protistol.* **29**, 1-18.

Munoz, M. L., Weinbach, E. C., Wieder, S. C., Claggett, C. E. and Weinbach, E. C. (1987) *Giardia lamblia*: detection and characterization of calmodulin. *Experimental Parasitology* **63**, 42-48.

Murti, K. G., Smith, H. T. and Freid, V. A. (1988) Ubiquitin is a component of the microtubule network. *Proc. Natl. Acad. Sci. USA* **85**, 3019-3023.

Musialowski, M. S., Flett, F., Scott, G. B., Hobbs, G., Smith, C. P. and Oliver, S. G. (1994) Functional evidence that the principal DNA replication origin of the *Streptomyces coelicolor* chromosome is close to the *dnaA-gyrB* region. *J. Bacteriol.* **176**, 5123-5125.

Newlon, C. S. (1993) Two jobs for the origin replication complex. *Science* **262**, 1830-1.

Nugent, J. M. and Palmer, J. D. (1991) RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. *Cell* **66**, 473-481.

O'Hara, E. B., Chekanova, J. A., Ingle, C. A., Kushner, Z. R., Peters, E. and Kurshner, S. R. (1995) Polyadenylylation helps regulate mRNA decay in *Escherichia coli. Proc. Natl. Acad. Sci. USA* **92**, 1807-1811.

Oakley, B. R., Oakley, C. E., Yoon, Y. and Jung, M. K. (1990) Gamma-tubulin is a component of the spindle pole body that is essential for microtubule function in *Aspergillus nidulans. Cell* **61**, 1289-1301.

Oakley, C. E. and Oakley, B. R. (1989) Identification of gamma-tubulin, a new member of the tubulin superfamily encoded by *mip*A gene of *Aspergillus nidulans. Nature* **338**, 662-664.

Ogasawara, N. and Yoshikawa, H. (1992) Genes and their organization in the replication origin region of the bacterial chromosome. *Mol. Microbiol.* **6**, 629-634.

Old, I. G., Margarita, D. and Girons, I. S. (1993) Unique genetic arrangement in the *dnaA* re' ion of the *Borrelia burgdorferi* linear chromosome: nucleotide sequence of the *dna* . ene. *FEMS Micro. Lett.* **111**, 109-111.

Olsen, G. J., Woese, C. R. and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **1765**, 1-6.

Osawa, S. and Jukes, T. H. (1989) Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* **28**, 271-278.

Osawa, S., Jukes, T. H., Watanabe, K. and Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229-264.

Ota, I. M. and Varshavsky, A. (1993) A yeast protein similar to bacterial two-component regulators. *Science* **262**, 566-569.

Ouzonis, C. and Sander, C. (1992) TFIIB, an evolutionary link between the transcription machineries of archaebacteria and eukaryotes. *Cell* **71**, 189-190.

Paces, J., Urbankova, V. and Urbanek, P. (1992) Cloning and characterization of a repetitive DNA sequence specific for *Trichomonas vaginalis. Mol. Biochem. Parasitol.* **54**, 247-256.

Page, F. C. and Blanton, R. L. (1985) The Heterolobosea (Sarcodina:Rhizopoda), a new class uniting the Schizopyrenida and the Acrasidae (Acrasida). *Protisologica* **21**, 121-132.

Palmer, J. R. and Reeve, J. N. (1993) Structure and function of methanogen genes. In *The Biochemistry Of Archaea*. M. Kates, D. J. Kushner and A. T. Matheson, Ed., Amsterdam, Elsevier. 497-534.

Potter, S., Durovic, P. and Dennis, P. P. (1995) Ribosomal RNA precursor processing by a eukaryotic U3 small nucleolar RNA-like molecule in an archaeon. *Science* **268**, 1056-1060.

Prescott, D. M. (1994) The DNA of ciliated protozoa. *Microbiol. Rev.* **58**, 233-267.

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

Pühler, G., Pitzer, F., Zwickl, P. and Baumeister, W. (1994) Proteasomes: multisubunit proteinases common to *Thermoplasma* and eukaryotes. *System. Appl. Microbiol.* **16**, 734-741.

Qin, S., Nakajima, B., Nomura, M. and Arfin, S. M. (1991) Cloning and characterization of a *Saccharomyces cerevisiae* gene encoding a new member of the ubiquitin-conjugating protein family. *J. Biol. Chem.* **266**, 15549-15554.

Ramírez, C., Köpke, A. K. E., Yang, D.-C., Boeckh, T. and Matheson, A. T. (1993) The structure, function and evolution of Archaeal ribosomes. In *The Biochemistry Of Archaea*. M. Kates, D. J. Kushner and A. T. Matheson, Ed., Amsterdam, Elsevier. 439-466.

Reiter, W.-D., Hüdepohl, U. and Zillig, W. (1990) Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection *in vitro*. *Proc. Natl. Acad. Sci. USA* **87**, 9509-9513.

Roger, A. J., Keeling, P. J. and Doolittle, W. F. (1994) Introns, the broken transposons. *Soc. Gen. Physiol. Ser.* **49**, 27-37.

Roger, A. J., Smith, M. W., Doolittle, R. F. and Doolittle, W. F. (1996) Evidence for the Heterolobosea from phylogenetic analysis of genes encoding glyceraldehyde-3-phosphate dehydrogenase. *J. Eukaryot. Microbiol.* in press.

Rohrwild, M., O. Coux, H. C. Huang, R. P. Moerschell, S. J. Yoo, J. H. Seol, C. H. Chung, and Goldberg, A. L. (1996) HslV-HslU: A novel ATP-dependent protease complex in *Escherichia coli* related to the eukaryotic proteasome. *Proc. Natl. Acad. Sci. USA* **93**, 5808-5813.

Rothärmel, T. and Wagner, G. (1995) Isolation and characterization of a calmodulin-like protein from *Halobacterium salinarium*. *J. Bacteriol.* **177**, 864-866.

Rowlands, T., Baumann, P. and Jackson, S. P. (1994) The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. *Science* **264**, 1326-1329.

Rozario, C., Morin, L., Roger, A. J., Smith, M. W. & Müller, M. 1996. Primary structure and phylogenetic relationships of glyceraldehyde-3-phosphate dehydrogenase genes of free-living and parasitic diplomonad flagelates. *J. Euk. Microbiol.* in press.

Rudolph, J. and Oesterhelt, D. (1995) Chemotaxis and phototaxis require a CheA histidine kinase in the archaeon *Halobacterium salinarium*. *EMBO J.* **14**, 667-673.

Sakai, Y., Goh, T. K. and Tani, Y. (1993) High-frequency transformation of a methylotrophic yeast, *Candida boidinii*, with autonomously replicating plasmids which are also functional in *Saccharomyces cerevisiae*. *J. Bacteriol.* **175**, 3556-3562.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) Molecular cloning: a laboratory manual. 2nd Ed., Cold Springs Harbor Laboratory Press, Cold Springs Harbor, NY.

Sanchez, M. A., Peattie, D. A., Wirth, D. and Orozco, E. (1995) Cloning, genomic organization and transcription of the *Entamoeba histolytica* alpha-tubulin-encoding gene. *Gene* **146**, 239-244.

Sanchez, M., Valencia, A., Ferrandiz, M. J., Sander, C. and Vicente, M. (1994) Correlation between the structure and biochemical activities of FtsA, an essential cell division protein of the actin family. *EMBO J.* **1354**, 4919-4925.

Scheffner, M., Nuber, U. and Huibregtse, J. M. (1995) Protein ubiquitination involving an E1-E2-E3 enzyme ubiquitin thioester cascade. *Nature* **373**, 81-83.

Schlesinger, D. H. and Goldstein, G. (1975) Molecular conservation of 74 amino acid sequence of ubiquitin between catttle and man. *Nature* **253**, 423-424.

Schmidt, M., Svendsen, I. and Feierabend, J. (1995) Analysis of the primary structure of the chloroplast isozyme of triosephosphate isomerase from rye leaves by protein and cDNA sequencing indicates a eukaryotic origin of its gene. *Biochim. Biophys. Acta.* **1261**, 257-264.

Schneider, S. U., Leible, M. B. and Yang, X.-P. (1989) Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase-oxygenase of two species of *Acetabularia* and the occurrence of unusual codon useage. *Mol. Gen. Genet.* **218**, 445-452.

Schull, C. and Beier, H. (1994) Three *Tetrahymena* tRNA(Gln) isoacceptors as tools for studying unorthodox codon recognition and codon context effects during protein synthesis in vitro. *Nucleic Acids Res* **22**, 1278-1280.

Scolnick, E., Tompkins, R., Caskey, T. & Nirenberg, M. 1968. Release factors differing in specificity for terminator codons. *Proc. Natl. Acad. Sci. USA* **61**:768-774.

Sharp, P. M. and Li, W.-S. (1987) Ubiquitin genes as a paradigm of concerted evolution of tandem repeats. *J. Mol. Evol.* **25**, 58-64.

Shimmin, L. C., Newton, C. H., Ramírez, C., Yee, J., Downing, W. L., Louie, A., Matheson, A. T. and Dennis, P. P. (1989) Organization of genes encoding the L11, L1, L10, and L12 equivalent ribosomal proteins in eubacteria, archaebacteria, and eukaryotes. *Can. J. Microbiol.* **35**, 164-170.

Sobel, S. G. and Synder, M. (1995) A highly divergent gamma-tubulin gene is essential for cell growth and proper microtubule organization in *Saccharomyces cerevisiae*. *J. Cell. Biol.* **131**, 1775-1788.

Sogin, M. L. (1989) Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. *Amer. Zool.* **29**, 487-499.

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

Soltys, B.J. and Gupta, R.S. (1994) Presence and cellular distribution of a 60-kDa protein related to mitochondrial Hsp60 in *Giardia lamblia*. *J. Protistol.* **80**, 580-588.

Sommer, T. and Jentsch, S. (1993) A protein translocation defect linked to ubiquitin conjugation at the endoplasmic reticulum. *Nature* **365**, 176-179.

Spudich, J. L. (1993) Color sensing in the archaea: a eukaryotic-like receptor coupled to a prokaryotic transducer. *J. Bacteriol.* **175**, 7755-7761.

Spudich, J. L. (1994) Protein-protien interaction converts a proton pump into a sensory receptor. *Cell* **79**, 747-750.

Stanier, R. Y. (1970) Some aspects of the biology of cells and their possible evolutionary significance. *Symp. Soc. Gen. Mircrobiol.* **20**, 1-38.

Stanier, R. Y. and van Neil, C. B. (1941) The main outlines of bacterial classification. *J. Bacteriol.* **48**, 437-466.

Stanier, R. Y. and van Neil, C. B. (1962) The concept of a bacterium. *Arch. Microbiol.* **42**, 17-35.

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

Strynadka, N. C. and James, M. N. (1989) Crystal structures of the helix-loop-helix calcium-binding proteins. *Annu. Rev. Biochem.* **58**, 951-998.

Suhan, M., Chen, S.-Y., Thompson, H. A., Hoover, T. A., Hill, A. and Williams, J. C. (1994) Cloning and characterization of an autonomous replicon from *Coxiella burnetii*. *J. Bacteriol.* **176**, 5233-5243.

Sullivan, M. L. and Vierstra, R. D. (1991) Cloning of a 16-kDa ubiquitin carrier protein from wheat and *Arabidopsis thaliana*. Identification of functional domains by *in vitro* mutagenesis. *J. Biol. Chem.* **266**, 23878-23885.

Sung, P., Prakash, S. and Prakash, L. (1990) Mutation of cysteine-88 in the *Saccharomyces cerevisiae* RAD6 protein abolishes its ubiquitin-conjugating activity and its various biological functions. *Proc. Natl. Acad. Sci. USA* **87**, 2695-2699.

Sutrave, P., Shafer, B. K., Strathern, J. N. and Hughes, S. H. (1994) Isolation, identification and characterization of the *FUN12* gene of *Saccharomyces cerevisiae*. *Gene* **146**, 209-213.

Swafford, D. L. (1993) *PAUP: Phylogenetic Analysis Using Parsimony*. Illinois Natural History Survey, Champaign. 3.1.1.

Swan, D. G., Hale, R. S., Dhillon, N. and Leadlay, P. F. (1987) A bacterial calcium-binding protein homologous to calmodulin. *Nature* **329**, 84-85.

Tan, Y., Bishoff, S. T. and Riley, M. A. (1993) Ubiquitins revisited: further examples of within- and between-locus concerted evolution. *Mol. Phylo. Evol.* **2**, 351-360.

Templeton, A. R. (1983) Phylogenetic inference from restriction endonuclease cleavege site maps with particular reference to the evolution of humans and the apes. *Evolution* **37**, 221-244.

Thomas, A., Goumans, H., Voorma, H. O. and Benne, R. (1980) The mechanism of action of eukaryotic initation factor 4C in protein synthesis. *Eur. J. Biochem.* **107**, 39-45.

Thomm, M., Hausner, W. and Hethke, C. (1994) Transcription factors and termination of transcription in *Methanococcus*. *Syst. App. Microbiol.* **16**, 648-655.

Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Breniere, S. F., Darde, M. L. and Ayala, F. J. (1991) Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proc. Natl. Acad. Sci. USA* **88**, 5129-5133.

Tittiger, C., Whyard, S. and Walker, V. K. (1993) A novel intron site in the triosephosphate isomerase gene from the mosquito *Culex tarsalis*. *Nature* **361**, 470-472.

Tourancheau, A. B., Tsao, N., Klobutcher, L. A., Pearlman, R. E. & Adoutte, A. 1995. Genetic code deviations in the ciliates: evidence for multiple and independent events. *EMBO J.* **14**:3262-3267.

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

van Neil, C. B. (1946) The classification and natural relationships of bacteria. *Cold Springs Harbor Symp. Quant. Biol.* **11**, 285-301.

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

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

Warburton, P. E., Waye, J. S. and Willard, H. F. (1993) Nonrandom localization of recombination events in human alpha satelite repeat unit variants: implications for higher-order structural characteristics within centromeric heterchromatin. *Mol. Cell. Biol.* **13**, 6520-6529.

Wenzel, T. and Baumeister, W. (1993) *Thermoplasma acidophilum* proteasomes degrades partially unfolded and ubiquitin-associated proteins. *FEBS Lett.* **326**, 215-218.

Wettach, J., Gohl, H. P., Tschochner, H. and Thomm, M. (1995) Functional interaction of yeast and human TATA-binding protein with an archaeal polymerase and promoter. *Proc. Natl. Acad. Sci. USA* **92**, 472-476.

White, B. N. and Bayley, S. T. (1972) Methionine transfer RNAs from the extreme halophile, *Halobacterium halobium. Biochim. Biophys. Acta* **272**, 583-587.

Whittaker, R. H. (1969) New concepts of kingdoms of organisims. *Science* **163**, 150-163.

Woese, C. R. (1994) There must be a prokaryote somewhere: Microbiology's search for itself. *Microbiol. Rev.* **58**, 1-9.

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

Woese, C. R. and Fox, G. E. (1977b) The concept of cellular evolution. *J. Mol. Evol.* **10**, 1-6.

Wolf, S., Lottspeich, F. and Baumeister, W. (1993) Ubiquitin found in the archaebacterium *Thermoplasma acidophilum. FEBS Lett.* **326**, 42-44.

Wong, S., Elgort, M. G., Gottesdiener, K. and Campbell, D. A. (1992) Allelic polymorphism of the *Trypanosoma brucei* polyubiquitin gene. *Mol. Biochem. Parasitol.* **55**, 187-196.

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

Yee, T. W. and Smith, D. W. (1990) *Pseudomonas* chromosomal replication origins: a bacterial class distinct from *Eschericheria coli*-type origins. *Proc. Natl. Acad. Sci. USA* **87**, 1278-1282.

Yoshikawa, H. and Ogasawara, N. (1991) Structure and function of DnaA and the DnaA-box in eubacteria: evolutionary relationships of bacterial replication origins. *Mol. Microbiol.* **5**, 2589-2597.

Zakrzewska-Czerwinska, J. and Schrempf, H. (1992) Characterization of an autonomously replicating region from the *Streptomyces lividans* chromosome. *J. Bacteriol.* **174**, 2688-2693.

Zambryski, P. (1988) Basic processes underlying *Agrobacterium*-mediated DNA transfer to plant cells. *Annu. Rev. Genet.* **22**, 1-30.

Zheng, Y., Jung, M. K. and Oakley, B. R. (1991) Gamma-tubulin is present in *Drosophila melanogaster* and *Homo sapiens* and is associated with the centrosome. *Cell* **65**, 817-823.

Zhouravleva, G., Frolova, L., Le Goff, X., Le Guellec, R., Inge-Vechtomov, S., Kisselev, L. and Philippe, M. (1995) Termination of translation in eukaryotes is governed by two interacting polypeptide chain release factors, eRF1 and eRF3. *EMBO J.* **14**, 4065-4072.

Zillig, W. (1987) Eukaryotic traits in archaebacteria. *Ann. N. Y. Acad. Sci.* **503**, 78-81.

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., Hüdepohl, U., Hain, J., Lanzendörfer, M. and Holz, H. (1993) Transcription in Archaea. In *The Biochemistry Of Archaea*. M. Kates, D. J. Kushner and A. T. Matheson, Ed., Amsterdam, Elsevier. 367-391.

Zuckerkandl, E. and Pauling, L. (1965a) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*. New York, Academic Press.

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

Zwickl, P., Grziwa, A., Pühler, G., Dahlmann, B., Lottspeich, F. and Baumeister, W. (1992) Primary structure of the *Thermoplasma* proteasome and its implications for the structure, function, and evolution of the multicatalytic proteinase. *Biochemistry* **31**, 964-972.

Zwickl, P., Lottspeich, F., Dahlmann, B. and Baumeister, W. (1991) Cloning and sequencing of the gene encoding the large (α-) subunit of the proteasome from *Thermoplasma acidophilum*. *FEBS Lett.* **278**, :217-221.

Zyskind, J. W., Cleary, J. M., W S A Brusilow, Harding, N. E. and Smith, D. W. (1983) Chromosomal replication origin from the marine bacterium *Vibrio harveyi* functions in *Escherichia coli: oriC* consensus sequence. *Proc. Natl. Acad. Sci. USA* **80**, 1164-1168.