# Origins and Evolution of the Archaebacterial and Eukaryotic DNA Replication Apparatus

by

David R. Edgell

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

at

Dalhousie University
Halifax, Nova Scotia
July, 1997

0-612-24737-6

**DALHOUSIE UNIVERSITY**

**FACULTY OF GRADUATE STUDIES**

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Origins and evolution of the archaebacterial and eukaryotic DNA replication apparatus"

by David Rohan Edgell

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

Dated: September 3, 1997

External Examiner

Research Supervisor

Examining Committee

[redacted]

DALHOUSIE UNIVERSITY

DATE: Sept 10th/97

AUTHOR: David R. Edgell

TITLE: Origins and Evolution of the Archaebacterial and Eukaryotic DNA Replication apparatus

DEPARTMENT OR SCHOOL: Biochemistry

DEGREE: PhD CONVOCATION: October YEAR: 1997

[redacted]

Signature of Author

# Table of contents

## Figures and Tables

# Abstract

DNA replication is a fundamental process of living things. This thesis examines the origins and evolution of DNA replication proteins found in the three primary domains of life, the eubacteria, the archaebacteria, and the eukaryotes.

First, DNA-dependent DNA polymerases of archaebacteria and eukaryotes were studied. A PCR-based approach was used to amplify and sequence various family B DNA polymerases from thermoacidophilic archaebacteria and early-diverging eukaryotes. Phylogenetic analysis of these and other sequences indicated that the DNA polymerases of archaebacteria and eukaryotes have evolved by a series of independent gene duplications, but the order of the duplication events remains unclear. Unexpectedly, one eukaryotic DNA polymerase, ε, appears more related to archaebacterial DNA polymerases than to any other eukaryotic polymerase.

Second, the replication proteins of eubacteria and eukaryotes that perform analogous functions at the replication fork were examined by computer-based methods to resolve issues of evolution by duplication and homology. It was found that many replication proteins of eukaryotes are members of gene families, whereas eubacterial replication proteins are not. Eukaryotic replication proteins likely evolved by gene duplication after the split of the eukaryotic and eubacterial lineages. Archaebacterial genomes also encode many proteins that are members of gene families and that are homologous to eukaryotic replication proteins.

The question of homology of eubacterial and eukaryotic replication proteins was addressed by comparisons of amino acid alignments of proteins performing analogous functions. There is no evidence from amino acid alignments that eubacterial and eukaryotic replication proteins are homologs. Thus, there is little evidence to support the notion that DNA replication proteins evolved from a single set of replication proteins present in the last common ancestor of eubacteria, archaebacteria, and eukaryotes.

## Abbreviations used

| | |
|---|---|
| BLAST | basic local alignment search tool |
| dNTP | deoxynucleoside triphosphate |
| kb | kilo-base pair |
| Mb | mega-base pari |
| ML | maximum likelihood |
| nts | nucleotides |
| nsy diff | non-synonymous differences |
| nsy sub rate | non-synonymous substitutions per site |
| ORF | open reading frame |
| PCR | polymerase chain reaction |
| rRNA | ribosomal RNA |
| SSU | small subunit |
| syn diff | synonymous differences |
| syn sub rate | number of synonymous substitutions per site |

# Acknowledgments

I cannot claim sole credit for many of the ideas presented in this thesis. Thoughts elaborated in this thesis are the end result of productive discussion over the last five years with various people, not the least of whom is Ford Doolittle. Ford has provided a rare and unique atmosphere for me to learn science, not only through experimental means but also through many of the informal discussion sessions spent in the lab "library", at lab meetings, or at one of his many (perhaps too many?) potluck dinners. It is these discussions, be it about science or some quirk of floor management policy and toilet planning, that I will miss the most. Ford's support has extened beyond providing my salary; he has showed great patience and encouragement when I was frustrated with benchwork and when I took time off to persue other aspects of my life. Of course, thanks is also due to all the members of the Doolittle lab, both past and present, that have been present during my time here - you know who you are. I would also like to thank Dr. Melanie Dobson for technical advice and for interesting and helpful discussions on various subjects.

As for my friends outside of the lab, I wish only that we continue our friendship for of all things in my life, I value that the most. To Sean, Bonnie, Johnathon, Bob, Sandie, Nathalie, and all my other climbing partners, I thank you for all the times you have been on the other end of the rope. To Ashley and Heather, I thank you for calming conversation in the face of utter panic and despair over this thesis. To my parents, I thank you for the financial and (much needed) emotional support you have provided over the last five years. And, finally, to Bruce. What more can I say, man? It's done, at long last it's done.

## Introduction

### Archaebacteria and the nature of the last common ancestor of life

In the late 1970's, Woese and Fox published two seminal findings that stemmed from comparisons of oligonucleotide catalogs of 16S and 18S ribosomal RNAs of prokaryotes and eukaryotes (Woese and Fox, 1977b; Woese and Fox, 1978). They were (1) that the eukaryotic nuclear lineage was not specifically related to any known prokaryotic lineage, and (2) that prokaryotes could be divided into two unrelated groups, one comprising well-studied bacteria such as *Escherichia coli, Bacillus subtilis* and cyanobacteria, which they called eubacteria, and the other group comprising a collection of organisms found in extreme habitats, which they called archaebacteria.

The first finding came as a surprise to many biologists. Previous to Woese and Fox's publication, the prevailing view of the biological world was that of a prokaryote/eukaryote transition. Prokaryotes were defined as those organisms that lacked many of the ultrastructural and cellular features found in eukaryotes (Stanier and van Niel, 1962; Stanier, 1970). Specifically, these were the absence of a nucleus, the lack of cell division by mitosis, the lack of membrane-bound respiratory organelles, and the lack of an elaborate cytoskeletal system. As such, the prokaryote state of cellular organization was seen as simpler and earlier than the eukaryote state of cellular organization. Prokaryotes were believed to have evolved from a yet simpler pre-prokaryote state of organization, while eukaryotes were believed to have evolved from a particular group of prokaryotes (most likely cyanobacteria) by a gradual increase in internal complexity, eventually resulting in the appearance of the nucleus and cytoskeleton. According to this view of cellular evolution, rRNA sequences from the eukaryotic nucleus would have shown a specific affinity to a known prokaryotic group, while the rRNA of organelles would have shown affinities to yet other prokaryotic group(s). However, Woese and Fox's analyses indicated that the eukaryotic nuclear lineage did not evolve from within a prokaryote group, but instead represented a separate line of descent altogether (Woese and Fox, 1977b; Fox et al., 1980).

That eukaryotes might represent a separate line of descent from a common ancestor was in agreement with the radical differences in genome structure between the two groups. Doolittle (Doolittle, 1978) argued that eukaryotic genome organization, typified by genes interrupted by non-coding sequences, was too different from that of prokaryotes to imagine how such (seemingly) informationally irrelevant sequences could be added to existing prokaryotic structural genes without deleterious effects. Rather, he envisioned the eukaryotic genomic state as a primitive one, from which the genomes of prokaryotes evolved by succumbing to pressure to eliminate non-coding DNA from the genome (most commonly referred to as streamlining). Darnell (Darnell, 1978) argued along similar lines, but was also struck by the haphazard organization of the eukaryotic genome as opposed to efficiently organizied prokaryotic genomes (operons, tightly regulated gene expression, linked transcription and translation) and saw the eukaryotic genome arising independently from a common ancestor.

Woese and Fox's second finding, that prokaryotes could be divided into two groups of organisms, each as distinct as both were from eukaryotes on the basis of oligonucleotide catalogs of rRNAs, was perhaps more of a surprise than their first. Woese and Fox named these two groups eubacteria and archaebacteria to replace the taxonomic division of the living world into prokaryote or eukaryote and to emphasize the distinctiveness of archaebacteria from other bacterial groups. The name archaebacteria was chosen purposely to connote a sense of antiquity because of the belief that the type of habitats that these organisms were isolated from, high temperature, low pH, extremely high levels of salinity, were representative of the types of habitats to which bacteria of the Archean age would have been exposed (Woese and Fox, 1978).

In the same year, Woese and Fox published another paper entitled "The concept of cellular evolution" (Woese and Fox, 1977a). This paper focused on the key question of whether the eukaryotic nuclear lineage represented a separate line of descent from the prokaryotic lineage (both eubacterial and archaebacterial) and if so, what was the state of cellular organization of the common ancestor of eukaryotes and prokaryotes. Woese and Fox argued that the eukaryotic lineage was distinct from that of the eubacterial and archaebacterial lineages and concluded that

> "Eucaryotes did arise from procaryotes, but only in the sense that the procaryote is an organizational, not a phylogenetic distinction. In analogous fashion, procaryotes arose from simpler entities. The latter are properly called *progenotes,* because they are still in the process of evolving the relationship between genotype and phenotype. It is at the progenote state, not the procaryote stage, that the line of descent leading to the eucaryotic cytoplasm diverged from the bacterial lines of descent."

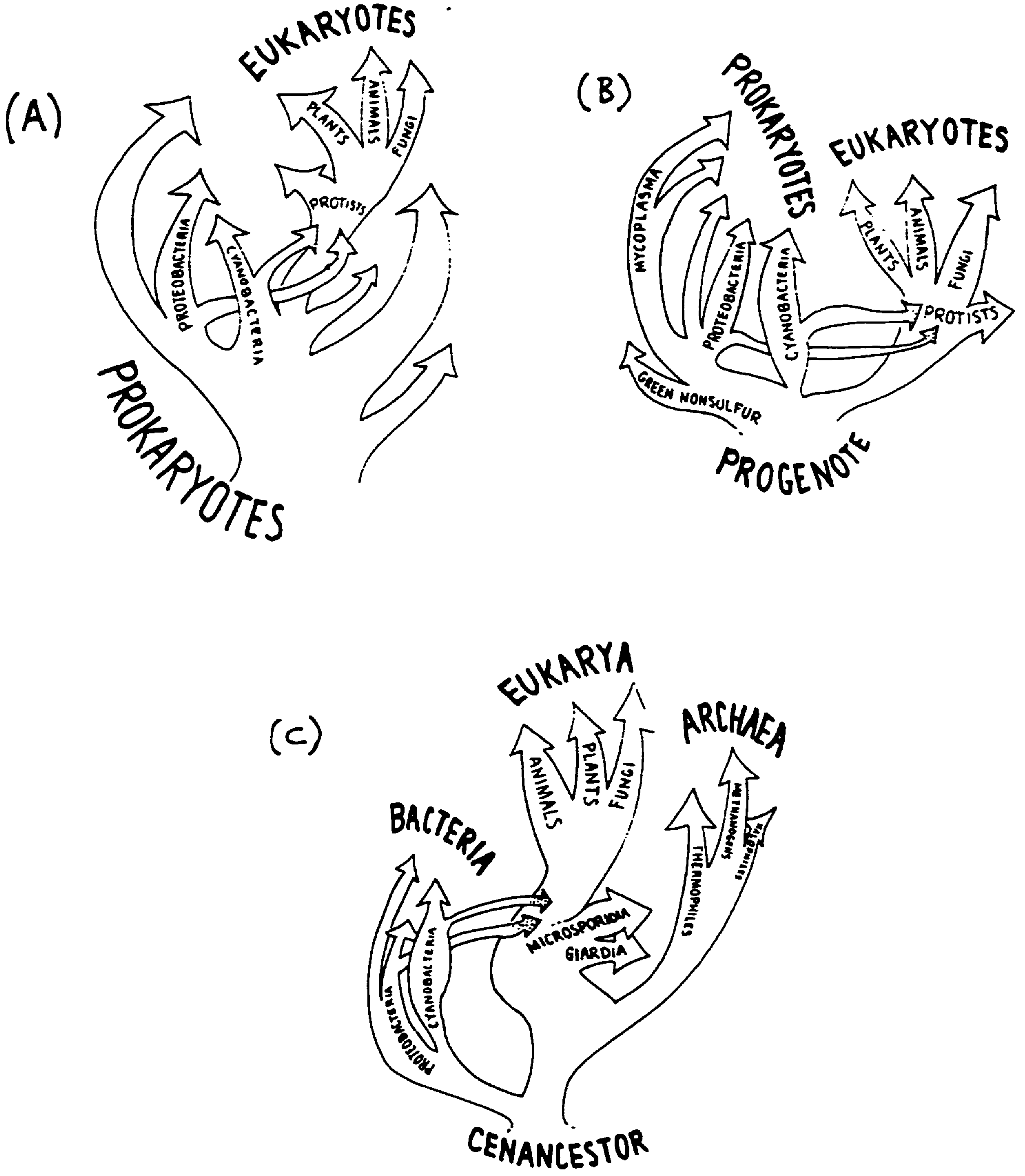
Fox and Woese believed that the ancestor of prokaryotes and eukaryotes was a progenote because of a number of key differences in the translation machineries: different patterns of post-translational modification of rRNAs, signature sequences in rRNAs specific to eukaryotes or prokaryotes, size of ribosomes (60S versus 80S), and differing antibiotic sensitivities. These differences were too great to be explained by either (i) a gradual change in the eukaryotic machinery if it evolved from a prokaryotic one (ie. eukaryotes evolved from within prokaryotes), or (ii) by the divergence time (measured by mutation rate) separating the two lineages from a common ancestor. Rather, Woese and Fox saw modifications and improvements in components of the translation apparatus as occurring independently in the prokaryote and eukaryote lineage after each diverged from the progenote. Woese and Fox also believed that, due to the presence of an inefficient translation apparatus, other cellular processes would also have been under Darwinian selection in the progenote. They argued that

> "It is difficult to overestimate the effect on the nature and evolution of the cell that an appreciable translation error rate (a noisy genetic transmission channel) would have. The primary constraint would be on the size and properties of the proteins that could be evolved. This in turn would delimit the specificity of *all* of the cell's interactions."

Problems due to inefficiencies in cellular processes of storage and processing of genetic information (genome organization, control pathways, DNA repair mechanisms and certain enzymes involved in DNA replication) would have been met and solved independently in the prokaryotic and eukaryotic lineages.

The discovery of the archaebacteria as a separate line of descent from eubacteria and eukaryotes only strengthened Woese and Fox's belief that the common ancestor of all life was a progenote (Woese and Fox, 1977b; Woese and Fox, 1978). Archaebacteria possessed key differences in components of the

**Figure 1** Historical relationships of organisms (from Doolittle and Brown, 1994). (A) Evolutionary view between 1970 and 1977. The eukaryotic nuclear lineage arose from within the already characterized prokaryotes (eubacteria). (B) Implications of rRNA oligonucleotide catalogs of Woese and colleagues. The prokaryotic (eubacterial) and eukaryotic nuclear lineage arose independently from a primitive last common ancestor, the progenote, still experiencing progressive Darwinian selection to improve various cellular processes. The archaebacteria (Woese and Fox, 1977b) would represent a third lineage on this diagram, also evolving independently from the progenote. (C) Implication from rooting of the universal tree of life using paralogous protein families (Gogarten et al., 1989; Iwabe et al., 1989) and renaming of the three domains (Woese et al., 1990).


(A)
EUKARYOTES
PLANTS
ANIMALS
FUNGI
PROTISTS
PROTEOBACTERIA
CYANOBACTERIA
PROKARYOTES
(B)
PROKARYOTES
EUKARYOTES
MYCOPLASMA
PROTEOBACTERIA
CYANOBACTERIA
PLANTS
ANIMALS
FUNGI
PROTISTS
GREEN NONSULFUR
PROGENOTE
(C)
EUKARYA
ANIMALS
PLANTS
FUNGI
ARCHAEA
BACTERIA
METHANOGENS
HALOPHILES
THERMOPHILES
MICROSPORIDIA
GIARDIA
CYANOBACTERIA
PROTEOBACTERIA
CENANCESTOR

translation apparatus that Woese and Fox thought were as distinct as differences between the eukaryotic and eubacterial translation machinery. In addition, archaebacteria did not appear to use peptidoglycan in construction of their cell walls, one of the defining structures of eubacterial cell walls. Archaebacteria also possessed unique ether linked lipids, not known as constituents of any eubacterial or eukaryotic cell membrane. However, archaebacteria did resemble eukaryotes in certain traits: a high level of modification of tRNAs, the initiator tRNA carrying a nonformylated methionine, and sensitivity to certain antibiotics specific for eukaryotes. In spite of these similarities, Woese and Fox still maintained that the deepest division in life was a tripartite one with the three groups arising from the progenote.

The concept of the progenote dominated the molecular evolutionary literature for much of the 1980s (see for example Doolittle, 1980; Doolittle, 1989; Doolittle and Brown, 1994). Two events brought a gradual change in the thinking about the relationship of the three domains to each other and on the nature of the last common ancestor. First, stimulated by the sequencing of archaebacterial genes, was that archaebacteria, eubacteria, and eukaryotes possessed homologous proteins that performed identical roles in biochemical pathways. It became evident from sequencing of archaebacterial genes that in terms of genomic organization, archaebacteria resembled eubacteria in many respects, and that this genome organization was probably ancestral to both groups.

The most compelling studies of archaebacterial proteins with functionally identical homologs in eubacteria and eukaryotes were those on DNA-dependent RNA polymerases by Zillig and co-workers (see for example Huet et al., 1983; Gropp et al., 1986). Archaebacteria, like eubacteria, possess only a single RNA polymerase, but it is more similar in subunit composition and biochemistry to the three eukaryotic RNA polymerases. As amino acid sequences appeared for archaebacterial RNA polymerase subunits, it became obvious that the RNA polymerases of eubacteria, archaebacteria, and eukaryotes were homologs. The observed degree of amino acid similarity was difficult to reconcile with Woese and Fox's prediction of independent evolution of components of the genetic machinery in eubacterial, archaebacterial and eukaryotic lineages after their divergence from the progenote. Amino acid sequences of archaebacterial proteins involved in translation (ribosomal proteins and elongation factors) also showed strong sequence similarity and a common function to proteins of eubacteria and eukaryotes (Amils et al., 1993; Rameriz et al., 1993).

If the cellular ancestor of eubacteria, archaebacteria, and eukaryotes was a progenote with inefficient translation and transcription, one might expect proteins involved in these processes in extant organisms to be different in function and sequence. However, the findings of Zillig and others posed a major problem for the progenote concept: the presence of functionally analogous proteins with high levels of sequence similarity in archaebacteria, eubacteria, and eukaryotes strongly implied that at least some (if not all) aspects of the processes in which these proteins functioned were already fixed in the progenote, and not under Darwinian selection for increased efficiency.

The second challenge to thinking on the nature of the last common cellular ancestor of eubacteria, archaebacteria and eukaryotes was the rooting of the universal tree of life which showed that archaebacteria and eukaryotes were sister groups, each sharing a more recent common ancestor than either did with eubacteria. Phylogenies of all life based on 16S and 18S rRNA cannot determine the branching order of eubacteria, eukaryotes, and archaebacteria because there is no outgroup; in three taxon trees there is no logical method of deciding which group is ancestral to the others unless other evidence (such as fossil records) exists. In 1989, two groups independently used the same method of finding an outgroup for all life; they based their phylogenies on paralogous protein families (Gogarten et al., 1989; Iwabe et al., 1989). The genes coding for the protein elongation factors EF-Tu/1$\alpha$ and EF-G/2 are present in the genomes of all extant organisms, are homologs of each other, and must have arisen by gene duplication from a single ancestral gene prior to the separation of the major lineages of life. By sequencing both genes from a diverse sampling of eubacteria, archaebacteria, and eukaryotes, it is possible to construct phylogenies and use one gene family (EF-Tu/1$\alpha$ for instance) to root the other gene family (EF-G/2). Phylogenies of both the elongation factor gene family and that of another ancestrally duplicated gene family, the proton-translocating ATPases, constructed in this manner showed that archaebacteria and eukaryotes shared a more recent common ancestor than either did with eubacteria (Gogarten et al., 1989; Iwabe et al., 1989).

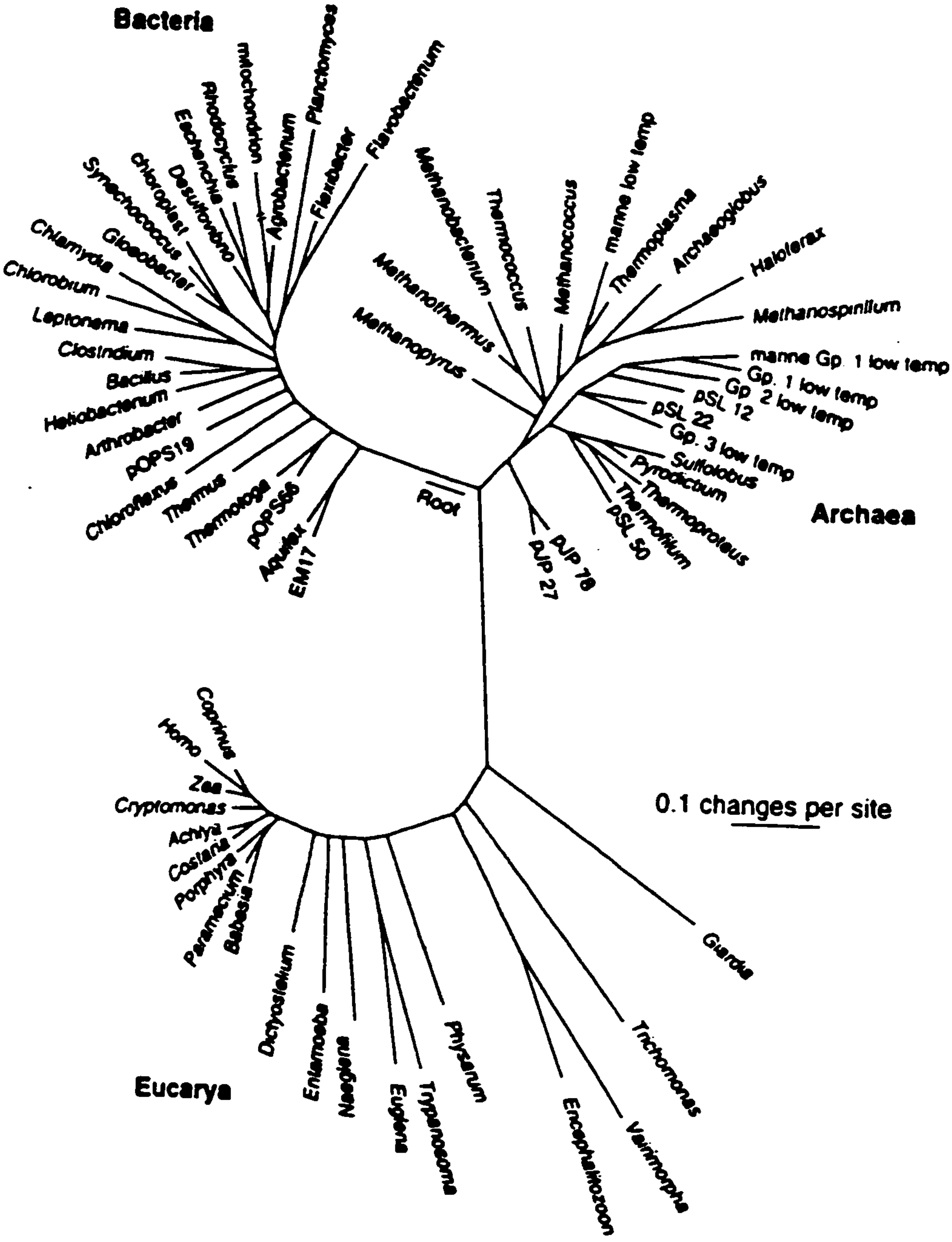
The rooting of the tree of life implied that the common ancestor of archaebacteria and eukaryotes was a structurally and biochemically advanced cell, and most likely prokaryote-like in grade of cellular organization. Combined with the growing body of evidence for homologous proteins performing identical steps in cellular pathways and for a common genome organization of eubacteria and archaebacteria, the notion of the last common ancestor of life as a progenote became less reasonable. However, the concept of a progenote should not be abandoned for it is a hypothetical, but logically necessary, ancestral entity in which the basic cellular machinery was still subject to Darwinian selection for increased efficiency and efficacy. But this stage of cellular evolution surely preceded that from which modern cells evolved.

## Archaebacterial phylogeny

The original rRNA oligonucleotide catalogs of Woese, Fox and colleagues indicated that archaebacteria could be divided into two groups, one consisting of extreme halophiles and methanogens, and the other consisting of thermoacidophiles (Woese and Fox, 1977b; Fox et al., 1980). The preliminary division of archaebacteria into two kingdoms was later confirmed by molecular phylogenies based on complete 16S rRNA sequences and various protein datasets (figure 2). These studies demonstrated that any one particular archaebacterial sequence was more closely related to other archaebacterial sequences than to any eubacterial or eukaryotic sequence; archaebacteria appeared to be a monophyletic assemblage of two distinct groups diverging from a single common ancestor. Woese (Woese et al., 1990) named the two archaebacterial kingdoms *Euryarchaeota* (encompassing the halophile/methanogen clade) and *Crenarchaeota* (encompassing the thermoacidophiles).

Notable exceptions to studies supporting archaebacterial monophyly were those of Lake and co-workers based on ribosome structure in electron micrographs (Lake et al., 1984; Lake et al., 1986). Lake believed the archaebacteria to be paraphyletic and divided them into three groups: the photocytes (halophiles plus all eubacteria), the methanogens (which Lake called archaebacteria), and the eocytes (the thermoacidophiles). Because of the lack of similarity of the eocyte ribosome to that of other prokaryotic groups, eocytes were seen as the closest prokaryotic relative of eukaryotes. Heavily criticized by advocates of archaebacterial monophyly and of the three domain

**Figure 2** Phylogeny of eubacteria (bacteria), archaebacteria (archaea), and eukaryotes (eucarya) based on SSU rRNA sequences (from Pace, 1997). The tree is unrooted, but phylogenies based on paralogous protein families would place the root between the archaebacterial and eubacterial branch.


Bacteria
mitochondrion
Planctomyces
Flavobacterium
Flexibacter
Agrobacterium
Rhodocyclus
Escherichia
Desulfovibrio
chloroplast
Synechococcus
Gloeobacter
Chlamydia
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Root
Methanobacterium
Methanothermus
Methanopyrus
Thermococcus
Methanococcus
marine low temp
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
marine Gp. 1 low temp
Gp. 1 low temp
Gp. 2 low temp
pSL 12
pSL 22
Gp. 3 low temp
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
Archaea
0.1 changes per site
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum
Encephalitozoon
Vairimorpha
Trichomonas
Giardia
Eucarya

concept of Woese, the ribosome structure data lost significance when ribosome structures were found in archaebacteria purported by Lake and co-workers to be lacking those structures (Stoffler and Stoffler-Meilicke, 1986). However, Lake and Rivera's 1992 description of a shared single insertion in the elongation factor-1$\alpha$ gene of crenarchaeotes and eukaryotes, but not euryarchaeotes, provided perhaps the best single line of evidence against the monophyly of archaebacteria. Again, this evidence has been criticized for alignment inconsistencies and *ad hoc* selection of taxa (Baldauf et al., 1996).

The phylogenetic coherence of archaebacteria was challenged again by analyses of the elongation factor protein dataset as support for the sisterhood of crenarchaeotes (eocytes) and eukaryotes was found (Baldauf et al., 1996; Hashimoto and Hasegawa, 1996). In addition, many other protein datasets (for instance glutamine synthetase and heat shock proteins) do not show the "expected" archaebacterial phylogeny; in most of these cases, some (but not all) archaebacterial sequences branch with some (but not necessarily the same) eubacterial sequences, while other archaebacterial sequences branch with eukaryotic or other eubacterial sequences (see for example Hilario and Gogarten, 1993; Brown et al., 1994; Golding and Gupta, 1995). To reconcile these molecular phylogenies with those supporting archaebacterial monophyly, lateral-transfer of protein coding genes between archaebacterial and eubacterial lineages has been invoked (Hilario and Gogarten, 1993; Tiboni et al., 1993; Brown et al., 1994; Pesole et al., 1995). A more radical solution to these conflicting phylogenies, involving the cellular fusion of an archaebacterium and an eubacterium, has been suggested by various authors (Sogin, 1991; Zillig, 1991; Gupta and Golding, 1993; Gupta and Golding, 1995). In these chimeric scenarios, metabolic genes (such as glutamine synthetase and heat shock proteins) were “donated” by the eubacterial counterpart, while genes involved in transcription and translation were provided by the archaebacterial counterpart.

## Phylogeny of relevant eukaryote groups

Much of what is known about DNA replication in eukaryotes is based on studies of mutants of *S. cerevisiae*, cell lines of *H. sapiens*, or mammalian viruses (Kornberg and Baker, 1992). As indicated by rRNA phylogenies (figure 2), these organisms are by no means representative of the genetic and phylogenetic depth of eukaryotes (Sogin et al., 1989; Cavalier-Smith, 1993). Almost all of the current "model" eukaryotes are found in what is commonly referred to as the crown; basically animals, fungi, plants, and a few protist groups. However, much of the genetic, biochemical, environmental, and morphological diversity is found among the protists. Cavalier-Smith (Cavalier-Smith, 1993) has argued that changes in cell structure and organization accompanying the evolution of various protist groups are far more significant than most of the changes observed in the evolution of organisms found in the crown. The evolution of eukaryotic DNA replication from that of a prokaryotic ancestor is best studied then by examining the replication proteins of representatives of the earliest eukaryotic cells, not late evolving crown groups.

The first eukaryotes were most likely morphologically and structurally simple compared to extant representatives of many eukaryotic groups. Extant representatives of early diverging eukaryotic lineages (as determined by molecular phylogenies of rRNA and various protein coding genes) lack many of the "typical" cellular structures found in crown eukaryotes: they do not have mitochondria, peroxisomes, plastids, or Golgi dictyosomes. Cavalier-Smith (Cavalier-Smith, 1987a) collectively called these protists Archezoa. He included the Metamonads (ie. *Giardia lamblia*), the Archamoebae (ie. *Pelomyxa*), the Microsporidia (ie. *Nosema locustae*), and the Parabasalids (ie. *Trichomonas vaginalis*) in this group. The parabasalians were later removed from the archezoa because of the presence of a well developed Golgi apparatus (Cavalier-Smith, 1987b); all other archezoa lack "well-developed" Golgi apparatus. There is also recent evidence to suggest that *T. vaginalis* once possessed a mitochondrion that evolved into the energy-generating organelle characteristic of parabasalids, the hydrogenosome (Bui et al., 1996; Germot et al., 1996; Horner et al., 1996; Roger et al., 1996).

rRNA and elongation factor phylogenies of eukaryotes support the original concept of the archezoa as early-diverging eukaryotes robustly placing *G. lamblia* and other metamonads at the base of eukaryotes (figure 2; Cavalier-Smith, 1993; Cavalier-Smith and Chao, 1986; Keeling and Doolittle, 1996a; Baldauf et al., 1996). However, phylogenies of α- and β-tubulin place representatives of the microsporidia as a sister group to fungi; this is in stark contrast to rRNA phylogenies (Keeling and Doolittle, 1996b). The SSU rRNA sequence of *Phreatamoeba balamuthi* (an archamoebae) does not group with those of archezoa as expected but rather with those of other protist groups, near the radiation leading to animals, plants, and fungi (Hinkle et al., 1994). Although *T. vaginalis* and other parabasalians were formally removed from the archezoa by Cavalier-Smith, both rRNA and elongation factor phylogenies consistently place parabasalians near the base of eukaryotes (Cavalier-Smith, 1993; Roger, 1996). These two groups, the metamonads and parabasalians, thus seem to be logical choices for studying the DNA replication proteins of early-diverging eukaryotes.

## What is known about archaebacterial DNA replication

Much of what we know about the molecular biology and biochemistry of archaebacteria is based on studies of the translation and transcription systems. Although other cellular pathways such as cell division and central metabolism have recently become subjects of intense study, relatively little is known about the biochemistry of DNA replication in archaebacteria compared to what is known about eubacterial and eukaryotic replication (Kornberg and Baker, 1992).

The only detailed studies on the biochemistry of archaebacterial DNA replication centered around cataloging the sensitivities of archaebacteria to various known inhibitors of eukaryotic and eubacterial DNA replication. Aphidicolin, a specific inhibitor of eukaryotic DNA replication, was found to inhibit incorporation of radioactively labeled precursors into DNA in growing cultures of halophilic and methanogenic archaebacteria, but not to inhibit incorporation of radioactive precursors into RNA or protein (Forterre et al., 1984; Schnizel and Burger, 1984; Zabel et al., 1985). *In vitro,* aphidicolin specifically interferes with the ability of eukaryotic replicative DNA polymerases to bind dNTPs (Sheaff et al., 1991); a similar mechanism of inhibition of archaebacterial DNA polymerases is suggested by *in vivo* and *in vitro* studies (Zabel et al., 1987). The isolation and characterization of aphidicolin-sensitive DNA polymerases from halophilic, methanogenic and some thermophilic archaebacteria suggested that archaebacteria use a eukaryote-like aphidicolin-sensitive DNA polymerase for replication (reviewed in Forterre and Elie, 1993). However, not all archaebacteria show a sensitivity to aphidicolin, and aphidicolin-resistant DNA polymerases have been purified from archaebacteria which also possess an aphidicolin-sensitive DNA polymerase activity (see for example Elie et al., 1989; Hamal et al., 1990).

| **Polymerase family** | **Similar to** | **Other DNA polymerases in family** |
|---|---|---|
| A | *E. coli* DNA polymerase I (polA) | all eubacterial polA homologs, some bacteriophage and plasmid-encoded polymerases, the mitochondrial replicative DNA polymerase |
| B | *E. coli* DNA polymerase II (polB) | some bacteriophage polymerases, all archaebacterial DNA polymerases sequenced to date, three nuclear replicative polymerases (α, δ, ε) of eukaryotes, some eukaryotic plasmid and viral-encoded polymerases. |
| C | *E.coli* DNA polymerase III (polC) | all eubacterial homologs, no plasmid or phage encoded homologs, no known archaebacterial or eukaryotic homologs |
| X | Eukaryotic terminal transferases | terminal transferases of yeast and animals, DNA polymerase β of animals and fungi. |

**Table 1** Classification of DNA-dependent DNA polymerases into families based on sequence similarity to one of the three DNA polymerases of *E. coli* and eukaryotic terminal transferases. After Braithwaite and Ito, 1993.

Sequencing of the genes corresponding to aphidicolin-sensitive and -resistant DNA polymerases revealed that these DNA polymerases are all homologs of the three eukaryotic nuclear replicative DNA polymerases (α, δ and ε), and of DNA polymerase II (*pol*B) of *E. coli* (Forterre and Elie, 1993).

The numerous DNA-dependent DNA polymerases isolated from eubacteria and eukaryotes are classified into families based on amino acid sequence similarity to one of the three *E. coli* DNA polymerases (Braithwaite and Ito, 1993; table 1). Family A DNA polymerases include *E. coli* DNA polymerase I (PolI), all eubacterial PolI homologs, some eubacterial phage and mitochondrial (often called γ polymerase) DNA polymerases. Family B DNA polymerases include *E. coli* DNA polymerase II (PolII), some eubacterial phage DNA polymerases, the eukaryotic nuclear replicative DNA polymerase (α, δ, and ε), eukaryotic viral and plasmid-borne enzymes, and all archaebacterial DNA polymerases sequenced to date. Family C includes only eubacterial DNA polymerase III (PolIII) homologs: there are no known phage, viral, archaebacterial, or eukaryotic family C DNA polymerases. An additional eukaryotic nuclear encoded DNA polymerase, β, which functions in repair is assigned to family X. Members of family X share little amino acid sequence similarity with DNA polymerases but instead exhibit amino acid sequence similarity to terminal transferases.

Other than sequences of DNA-dependent DNA polymerases, little was known about other replication associated proteins of archaebacteria until the complete genome sequence of *Methanococcus jannaschii* became available (Bult et al., 1996); the implications of this genome sequence for understanding the evolution of eubacterial, archaebacterial, and eukaryotic replication proteins are the focus of chapter IV of this thesis. In spite of the availability of the complete genome sequence, nothing is known about archaebacterial chromosomal replication origins, the regulation and control of initiation of replication, segregation of chromosomes after completion of replication, the resolution and termination of (presumably) the two replication forks, and whether DNA replication is tightly coupled to cell division, as it is in some eubacterial systems.

## A word about homology

Genome sequencing projects rely on database search algorithms to identify open reading frames (ORFs) with similarity to ORFs of known function. Often, ORFs from the genome sequence are assigned as homologs of proteins from other organisms with the assumption that function will be the same. However, in the absence of functional data on ORFs identified by genome sequencing projects, it is premature to assume that the function(s) of the ORF will be identical to those of similar proteins in databases.

Homology has a strict meaning for evolutionary biologists: descent from a common ancestor (Reeck et al., 1986). Two or more proteins are homologous if they evolved by descent from a common ancestral protein. Evolutionary biologists would not assume that a common function of two or more proteins is sufficient evidence for homology because proteins can convergently (and independently) arrive at the same mechanistic, structural, or biochemical solution to a particular biological problem. Often, amino acid sequence similarity is the only criterion that genome sequencing projects can use for judging homology. Comparisons of amino acid sequence are usually expressed in terms of similarity and identity: similarity refers to conserved amino acid substitutions while identity refers to the same amino acid in the homologous position of two or more proteins. Proteins which share significant amino acid identity (usually 20-25% with allowance for gaps) are considered to be homologs (Doolittle, 1986). Two or more proteins with less than this level of identity (which is considered no better than a random alignment of two amino acid sequences) *might* be homologs and *may* have evolved from a common ancestral sequence, but have diverged too much in sequence to allow reconstruction of their history. In these cases, other types of evidence can be used. For instance, many proteins with similar biochemistries and cellular functions have been crystallized from diverse organisms. Comparisons of the secondary and tertiary structural elements may reveal similarity even in the absence of significant amino acid identity; this can be interpreted as divergent evolution.

Individual researchers and genome sequencing projects are commonly finding examples of proteins that have more than one homolog, either in the same genome or other organisms. Although all these proteins are correctly called homologs, the usefulness of this term becomes limiting in describing the evolutionary history of these proteins because it does not accurately describe the relationship between multiple homologs. In such situations, evolutionary biologists use two additional terms to refer to the historical relationships of proteins: orthologous and paralogous (Fitch, 1970).

Orthologous proteins (or orthologs) are encoded by genes which are related by speciation events, while paralogous proteins (or paralogs) are encoded by proteins related by gene duplication events. For instance, both *Homo sapiens* and *Saccharomyces cerevisiae* possess the genes for α- and β-tubulin. Since these two proteins share significant amino acid similarity and are homologs, they must have evolved by a gene duplication event. The α-tubulin genes from *H. sapiens* and *S. cerevisiae* are more accurately called orthologs since they are more similar in sequence to each other and other α-tubulins than either is to any β-tubulin sequence. However, the α- and β-tubulin genes from *H. sapiens* are paralogs because they are related by a gene duplication event which pre-dated the speciation event that gave rise to the organismal lineages leading to *H. sapiens* and *S. cerevisiae.*

## The role of gene duplication in the evolution of novel protein functions

The importance of evolution of novel protein functions by gene duplication was most eloquently expressed by Ohno in 1970, although the role of duplications in evolution had been recognized earlier by various authors (see Ohno, 1970 and references therein). Ohno's (Ohno, 1973) model for gene duplication proposed that

> "The mechanism of gene duplication provides a temporary escape from the relentless pressure of natural selection to a duplicated copy of a functional gene locus. While being ignored by natural selection, a duplicated and thus redundant copy is free to accumulate all manner of randomly sustained mutations. As a result, it may become a degenerate, nonsense DNA base sequence. Occasionally, however, it may acquire a new active site sequence, therefore a new function and emerge triumphant as a new gene locus."

An awkward prediction of the above model is that one duplicate copy must be freed from functional constraints such that it is able to accumulate random nucleotide substitutions. It is not obvious how one copy, presumably duplicated with the necessary up- and downstream sequences for efficient expression, could escape expression as such expression immediately places the gene under purifying selection to eliminate nonsense or deleterious (nonsynonymous) mutations. However, it is the accumulation of nonsynonymous substitutions (resulting in change of amino acid) that is often the indicator of positive selection for a novel protein function relative to that of the original gene.

Hughes (Hughes, 1994; but see also Goodman et al., 1975; Jensen and Byng, 1981) has proposed another model to accommodate many of the difficulties associated with Ohno's original proposal. The original gene, prior to gene duplication, "shares" its protein product between two different functions because the protein is bifunctional. After the gene duplication event, each duplicate copy can specialize to perform one of the functions of the original parental gene. Changes in the expression pattern of one or both of the duplicates could lead to such specialization and to the fixation of nonsynonymous amino acid substitutions specific to the function of each duplicate copy.

The genomes of eukaryotes are full of examples of homologous proteins performing different functions, and there can be little doubt that duplication has been one of the molecular mechanisms behind the creation of these protein functions. Many of these examples result from detailed studies on animal, plant or fungal systems. Yet, as many protein-coding genes in the eukaryote nuclear genome evolved from the (much smaller) genomic content of an archaebacterial-like ancestor, can gene duplication alone explain the evolution of these "eukaryotic-specific" proteins, some of which have no identifiable homologs in archaebacterial or eubacterial genome sequences?

## What this thesis is about

This thesis describes my efforts to describe the evolution of protein components of the DNA replication apparatus of eubacteria, archaebacteria and eukaryotes. Two quite different experimental approaches were employed: the first using a PCR-based approach to amplify and sequence DNA-dependent DNA polymerases from archaebacteria and eukaryotes (chapters I and II), and the second utilizing computer-based methods to catalog similarities and differences in protein components of DNA replication systems of extant organisms (chapters III and IV). The material presented in this thesis is quite diverse in subject matter as it deals not only with functional aspects of DNA replication proteins, but also attempts to trace the evolution of replication proteins back to the last common ancestor of cellular life, the cenancestor. Such an effort requires the synthesis of computer-based experimental work (chapter IV) and relevant information obtained from the literature.

## Materials and Methods

Strains *Escherichia coli* DH5α, NM522 and INVαF (Invitrogen) were used for all cloning and DNA manipulations. *E. coli* strain JM101 *Eco*K+ (Bio101) was used for propagation of M13 phages and for preparation of single-stranded DNA (ssDNA). *E. coli* LE392 and XL-1 Blue were used for screening of genomic and cDNA libraries. Strains were grown at 37°C in either LB or 2xYT solid or liquid media. Liquid media was supplemented with 10 mM $MgSO_4$ and 0.2% maltose if cells were to be used for library screening. 100 μg/ml ampicillin was included in both solid and liquid media for selection of plasmids. 5-Bromo-4-chloro-3-indolyl-ß-D-galactosidase (X-gal) at 20 μg/ml and isopropylthiogalactosidase (IPTG) at 0.1 mM were included in solid media to screen for the presence of plasmid inserts.

Genomic DNAs Genomic DNA from *Trichomonas vaginalis* strain NIH-C1 (ATCC#30001) was a gift of Miklos Müller (Rockefeller University). Genomic DNA was isolated from *Giardia lamblia* strain WB (ATCC#30957) grown in 15 ml glass culture tubes at 37°C in Keister's modified media supplemented with 250 μg/ml streptomycin and 165 μg/ml penicillin. When confluent growth was achieved, cells were pelleted into lysis buffer consisting of 0.5% SDS, 300 μg/ml proteinase K, 0.1 M NaCl and 1 mM EDTA and incubated at 50°C for 1 hour. This mixture was then extracted with an equal volume of Tris-buffered phenol (pH 8.0) and with phenol/chloroform/isoamyl alcohol (25:24:1 ratio). DNA was precipitated by addition of 2 volumes of ethanol. *G. lamblia* is very rich in carbohydrates, which can interfere with all subsequent molecular biological applications. To remove carbohydrates from DNA preparations, the ethanol precipitated DNA was resuspended in $sddH_2O$, NaCl and cetyltrimethylammonium bromide (CTAB; Sigma) added to final concentrations of 0.7 M and 1% respectively. This mixture was incubated at 65°C for 30 minutes and extracted twice with an equal volume of chloroform. CTAB complexes with carbohydrates and forms an insoluble layer between the organic and aqueous layers. The aqueous layer was removed and DNA preciptated by the addition of 2 volumes ethanol, 0.1 volume sodium acetate (pH 5.0).

Genomic DNA from *Sulfolobus acidocaldarius, S. solfataricus* strain MT4, and *S. shibatae* were gifts of Dr. Hans-Peter Klenk (The Institute for Genomic Research). Genomic DNA from *S. solfataricus* strain P1 and P2 were gifts of Margaret Schenk (Dalhousie University). Genomic DNA from *S. solfataricus* P2 (DSM#1617), *S. acidocaldarius* (DSM#639), and *S. shibatae* (ATCC#51178) was isolated from logarthimic cultures ($OD_{600}$ of 0.8-1.0) grown in 25 ml Brock's Media (see ATCC media#88) at 75°C in 50 ml Falcon tubes. Cells were pelleted, resuspended in 10 ml $sddH_2O$, and N-lauryl sarcosine was added to a final concentration of 0.8%. The resulting mixture was incubated at room temperature for 30 minutes to achieve lysis. The aqueous layer was extracted 2 x with phenol/chloroform. DNA was precipitated by the addition of 0.1 volumes of 7 M ammonium acetate and 100% ethanol. The pelleted DNA was washed 2 x with 70% ethanol.

<u>Libraries and screening procedures</u> The *G. lamblia* library in λgt11 was a gift of Dr. T. Nash (NIH, Bethesda). Aliquots of the library were incubated at 37°C for 20 minutes with 250 µl of *E. coli* LE392 in 10 mM $MgSO_4$. The adsorbed phage and cells were mixed with 7.5 mls 0.7% agarose, plated on 150 mm diameter NZCYM plates, and incubated at 37°C overnight. Duplicate plaques were lifted with nylon filters (DuPont). Filters were denatured 2 x 1 minute

in 1.5 M NaCl/0.5 M NaOH, neutralized 2 x 5 minutes in 1.5 M NaCl/0.5 M Tris-HCl (pH 7.5), and then washed 2 x with 2XSSC/0.2 M Tris-HCl (pH 7.5). Dried filters were then UV crosslinked. Filters were pre-washed 2 x 30 minutes with 0.1XSSC, 1.0%SDS. For overnight hybridization at 65°C, either Blotto (5 ml 20%SDS, 20ml 20XSSC, 0.5g non-fat dry milk in 75 mls $dH_20$) or Denhardt's solution (Sambrook et al., 1989) was used in as small a volume as possible that allowed independent movement of all filters. Stringency washes were 2 x 20 minutes in 2XSSC, and 1 x 20 minute in 1XSSC, 1.0%SDS. Filters were then exposed to film at -70°C for 2-5 days.

*T. vaginalis* genomic and cDNA libraries in λZAP (Stratagene) were gifts of Drs. Miklos Müller (Rockefeller University) and Patricia Johnson (UCLA). The screening proceedure for these libraries was essentially the same as above except that *E. coli* XL-1 blue was used as the host strain. *In vivo* excision of putative positive clones was performed as per manufacturers' instructions (Stratagene).

Southern hybridizations For Southern hybridizations of protist DNAs, 5 μg of genomic DNA was digested with various restriction enzymes (New England Biolabs). Digests were run overnight in 0.7% agarose gels to achieve separation of fragments. Gels were stained with ethidium bromide and observed under UV light. To denature genomic DNA, gels were soaked in 0.25 M HCl for 30 minutes, rinsed with $dH_2O$, washed 2 x 20 minutes in 1.5 M NaCl/0.5 M NaOH, and then neutralized 2 x 20 minutes in 1.5 M NaCl/0.5 M Tris-HCl (pH 7.5). Gels were then transferred overnight to nylon membranes by capillary blot method with 20XSSC transfer buffer. After transfer, membranes were rinsed in 2XSSC, allowed to dry, and UV crosslinked. Probes (usually 10 ng) were labelled with [$\alpha$-$^{32}$P]dATP as per manufacturers' instructions (Boehringer Manheim). Nylon membranes were prehybridized in Denhardt's solution for 1-3 hours at 65°C at which point fresh hybridization solution along with labelled probe was added. Probes were allowed to hybridize overnight at 65°C and then washed under various stringencies depending on which probe was used.

PCR Primers, both degenerate and exact match, are listed in appendix 1. PCR conditions varied depending on the primer combination and template but typically were carried out in 10 mM Tris-HCl, 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% Triton X-100, 0.2 mg/ml BSA, 2 U of Taq polymerase (Gibco-BRL; 0.2 ul/50ul reaction), and 5% acetamide (see Reysenbach et al., 1992). Primers were usually at 200 nm final concentrations and genomic DNA at 10-100ng. Denaturation was at 92°C for 1-2 minutes, annealing temperature was dependent on primer combinations (but between 45-50°C), and extension was at 72°C for 1-5 minutes depending on expected length of target sequence. All PCR reactions were covered with an equal volume of mineral oil. Reaction volumes were typically 50 µl for initial reactons, but 100 µl reactions were used when products were to be gel purified. Reactions (5 of 50 µl) were visualized on agarose gels against known standards (1 kb ladder, Gibco-BRL). Single primer and no DNA controls were also run at the same time to check for contamination and single primer artefacts. If bands of the correct molecular weight were observed, bands from at least 2 independent PCR reactions were gel purified (Bio-Rad) and ligated into a T-tailed vector (pCR2.1, Invitrogen). Ligations were either electroporated into DH5α or heat-shock transformed into INVαF.

To check for plasmid inserts of the correct size, white colonies (on LB amp/X-gal plates) were toothpicked directly into a 10 ul PCR reaction (as above) containing direct match primers (universal forward and reverse sequencing primers) to the pCR2.1 vector and also onto a master LB amp plate. Reactions were denatured at 92°C for 2 minutes, annealed at 45°C for 1 minute, and extended at 72°C for 1 minute. Usually, 20-30 cycles was sufficient to visualize bands from the entire 10 µl reaction on 0.7% agarose gels. Clones corresponding to PCR reactions with bands of the correct size (minus approximately 200 nts for the polylinker) were picked for sequencing analysis.

Inverse PCR Additional coding sequence outside of the original PCR product was obtained for DNA polymerase ε of *T. vaginalis* by inverse PCR (Ochman et al., 1988). The sequence of the initial PCR product indicated that the coding region contained a *Hind*III site; two sets of direct-match primers were designed, one which would amplify the 5' region of the ORF, and the other that would amplify the 3' region of the ORF (see figure 2.2). *T. vaginalis* genomic DNA was cut with *Hind*III and self-ligated. PCR reaction conditions were as for normal PCR except that Tris pH 8.8 was included. Annealing was at 50°C for 1 minute and extension was at 72°C for 4 minutes. Products were visualized on agarose gels, gel purified, and cloned as described above.

Sequencing PCR clones of the expected molecular weight were manually sequenced with Sequenase 2.0 (USB) or T7 Sequencing Kit (Pharmacia) to identify the insert. Plasmid DNA was prepared by a variety of methods. Comerically available mini-columns from Promega or Macherey-Nagel were used. More commonly, plasmid DNA was prepared by the Speedprep method. Briefly, 1.5 mls of logarithmically growing cells in LB amp media (NOT saturated cultures, O.D. 600 > 1.0) were resuspended in 100 µl lysis buffer consisting of 25 ml 2 M Tris (pH 8.0), 12.5 ml 0.5 M EDTA, 4.0 mls Triton X-100, and 10.5 g LiCl per 100 ml. Phenol/chloroform (100 µl) was added and the resulting solution vortexed for 30 seconds. After centriguation, 100 µl of the aqueous solution was removed and DNA precipitated by the addition of 200 µl ethanol. The pellet was resuspended in 10 µl TE + 0.1 µl 10 mg/ml RNase A. Typically 5 µl was used for 1 sequencing reaction. All double-stranded templates were denatured by the addtion of 0.1 vol. 1 M NaOH/0.1 vol. 2 mM EDTA and incubated at 37°C for 30 minutes. Denatured templates were precipitated by the addition of 2 volumes of ethanol and 0.1 volumes sodium acetate and centrifuged for 20 minutes, washed once with 70% ethanol, dried and resuspended in TE or sddH$_2$0. If inserts were longer than could be read by a single reaction, direct match sequencing primers were designed and used to walk along the clone. Some clones were sequenced by automated methods on either ABI or LiCor machines at NRC-Halifax.

Other sequencing strategies

(1) Single-strand exonuclease deletions Because one of the inverse PCR products was 2.1 kb, primer walking was not chosen as the method of sequencing. Rather, the product was cut with *Eco*R1, which gave two smaller products of 1.4 and 0.8 kb. These were cloned into the M13 *Eco*K vector (Bio101), transformed into JM101 *Eco*K+, and unidirectional deletions made (Bio101 M13 Single Step Nested Deletion Kit; see also Shen and Waye, 1989). Single-stranded DNA was then isolated for sequencing. The inserts were cloned in both orientations and deletions made on both strands. Insert orientation was determined by mixing 20 µl ssDNA preparation from two different clones, incubating at 60°C for 30 minutes in the presence of 2 µl 5 M NaCl, and analyzing the products on agarose gels. Inserts in the same

orientation will not hybridize and migrate instead at the molecular weight expected of ssDNA products. Inserts in opposite orientations will hybridize and will migrate more slowly than ssDNA; these products are easily differentiated from non-hybridized products.

(2) Sequencing by directed cloning of restriction fragments Screening of the *G. lamblia* genomic library for genomic fragments of DNA polymerase α resulted in the cloning of 2.2 and 1.4 kb *Bam*H1 fragments. These *Bam*H1 fragments were gel purified away from the cloning vector and separately digested to completion with two 4-cutter enzymes, *Nla*III and *Sau*3A1. The digestions were run on 7.5 % non-denaturing polyacrylamide gels and individual bands were purified and cloned into M13 cut with compatible enzymes (*Bam*HI in the case of *Sau*3A1, and *Sph*I in the case of *Nla*III). This procedure is more efficient than a random shot-gun method as the entire larger fragment can be sequenced by cloning of individual *Sau*3A1 or *Nla*III fragments. In addition, the *Sau*3A1 or *Nla*III fragments will overlap so that redundant coverage of the genomic fragment of interest is obtained. Gaps between fragments were filled by designing direct-match sequencing primers and used in sequencing reactions containing the 2.1 or 1.4 kb *Bam*HI fragments cloned in pBluescript SK- (Stratagene).

Database searches and contig assembly Putative PCR products were identified using the BLAST search algorithm (Altschul et al., 1990) at NCBI by either email or by World Wide Web (http://www.ncbi.nlm.nih.gov/BLAST/). The program BLASTX was primarily used as it translates query DNA sequences into all six possible reading frames and searches the protein database. BLASTN searches were not used as the level of sequence conservation of DNA polymerases and other replication-associated proteins is too low to detect with nucleotide level searches. In addition, the SEQ and XNU options, which filter repetitive sequences from searches, were used if PCR products were from organisms with biased A/T or G/C content (ie. *T. vaginalis*). P-values on the order of $10^{-6}$ or lower were considered as putative hits to similar genes in other organisms. Occasionally, the FASTA algorithm was used (http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html) to search for putative identities of PCR products and known sequences in databases. However, I found there was very little difference in the sensitivity of FASTA and BLAST algorithms to detect similarities between sequences of various PCR products and database sequences.

All contig assembly was done using the program LASERGENE (DNA Star) on MacIntosh computers.

Alignments Amino acid sequences were obtained from GenBank or other publicly available databases unless noted. Multiple alignments of amino acid sequences of the catalytic subunits of DNA-dependent DNA polymerases and DNA replication-associated proteins were done using the PILEUP option of GCG with default values or with CLUSTALW (Thompson et al. 1994), also with default values. For multiple alignments of the eukaryotic DNA polymerases, orthologous sequences were aligned first, due to the difficulty of aligning these proteins outside of conserved functional domains. Sets of orthologous alignments (those of DNA polymerases α, δ, ε, and Rev3) were then edited by hand to eliminate regions of low conservation and poor alignment and combined into a single large alignment. The same procedure was used to align archaebacterial family B DNA polymerases; this alignment was then combined with the eukaryotic-specific alignment. The numbering of amino acids in 3'-5' exonuclease and polymerization domains was as previously described (Wong et al., 1988): When available, information on catalytic residues critical for exonuclease or polymerization functions from site-directed mutagenesis studies was used to aid in the alignment of DNA polymerase sequences.

Phylogeny All datasets analyzed were coded as amino acid characters unless otherwise indicated. For parsimony analysis, PAUP 3.1.1 was used (Swafford, 1993). Initial searches to find the shortest trees were 100 random replicates with TBR branch swapping and 1 tree held at each step. 100 bootstrap replicates (simple stepwise addition) were performed to determine the branching confidence. Values of less than 50% were determined by analyzing the bootstrap partition printout. Alternative topologies were first created using MACLADE 3.0 (Maddison and Maddison, 1992) and then imported into PAUP. 100 random addition replicates were performed on these datasets to find the shortest tree.

For distance analysis, PHYLIP 3.57c (Felsenstein, 1996) was used on a Sun Sparc20 workstation. A PAM-corrected distance matrix was obtained with PROTDIST and used to calculate a tree by the neighbor-joining method (the NEIGHBOR option of PHYLIP). SEQBOOT was used to generate 100 random trees for bootstrap analysis. Templeton tests of alternative topologies were done using the PROTPARS option of PHYLIP with user-defined trees. The standard error for parsimony trees was determined by dividing the number of steps by the standard deviation.

Due to time constraints of exhaustive maximum likelihood (ML) searches with greater than 12 taxa, partially constrained trees based on optimal trees found by both parsimony and distance analyses were used. When parsimony and distance trees did not agree, nodes with greater than 75% in both parsimony and distance trees were constrained. Exhaustive searches on these constrained datasets was performed with PROTML (Adachi and Hasegawa, 1992) on a SunSparc20 workstation with the Jones, Thorton, Taylor substitution matrix. ML bootstrap values were obtained by the program MOL2CON (Arlin Stoltzfus, unpublished) which creates a consensus of the 100 most likely trees.

Calculation of substitution rates Nucleotide substitution rates were calculated with the program MEGA (Kumar et al., 1993). Nucleotide alignments of DNA polymerase and β-galactosidase genes from *S. solfataricus* P2 and *S. shibatae* were created by first aligning the amino acid sequences by hand, which were used as templates to align the nucleotide sequences. Gaps inserted into the amino alignments were scored as missing data. Nonsynonymous and synonymous substitutions per site were calculated according to the method of Nei and Gojobori (1986).

Pairwise comparions of amino acid sequences To determine the significance of pairwise alignments of replication proteins performing analogous functions in *E. coli* and *S. cerevisiae,* I used the FASTA 2.0 package (Pearson, 1990). The program PRSS shuffles a test sequence a specified number of times (I used 1000 shuffles), and aligns the shuffled test sequences with the second unshuffled sequence using the Smith-Waterman alignment algorithm. The 1000 shuffled alignment scores of the two sequences are compared to the scores of the unshuffled alignment, and a P-value calculated for the probability that the unshuffled score will fall within the range of shuffled alignment scores.

## Chapter I. Thermoacidophilic archaebacteria possess three family B DNA polymerases

### Introduction

Much of the eukaryotic nuclear genome likely derives from the genome of an archaebacterial-like ancestor (Gogarten et al., 1989; Iwabe et al. 1989; Brown and Doolittle, 1994; Baldauf et al., 1996; Lawson et al., 1996; but see also Gupta and Golding, 1995). The eukaryotic genetic apparatus (transcription and translation) and the proteins involved in these processes are more similar in amino acid sequence and biochemistry to archaebacterial homologs than they are to eubacterial homologs (for review see Zillig et al., 1993; Amils et al., 1993). However, little attention has been paid to the DNA replication apparatus of archaebacteria and the proteins involved.

Previous studies on archaeal DNA replication have focused primarily on the biochemistry of purified DNA polymerases in an attempt to classify them as eukaryotic- or eubacterial-like on the basis of enzymatic properties and sensitivity or resistance to various inhibitors (reviewed in Forterre and Elie, 1993). By using aphidicolin sensitivity as an indicator (see introduction), *Sulfolobus acidocaldarius* and *Sulfolobus solfataricus* were found to possess a eukaryotic-like DNA polymerase activity as well as an unclassified aphidicolin-resistant DNA polymerase activity (Elie et al., 1989; Klimczak et al., 1985; Rossi et al., 1985). Biochemical characterization and sequencing of the gene corresponding to the aphidicolin-sensitive activity from *S. solfataricus* strain MT4 confirmed that this DNA polymerase was a family B homolog more similar to eukaryotic than to eubacterial homologs (Pisani et al., 1992). This paralog is called *S. solfataricus* MT4 B1.

Prangishvili and Klenk attempted to clone and sequence the gene for the aphidicolin-resistant DNA polymerase activity from *S. solfataricus* P2 by designing a degenerate oligonucleotide against domain I of eukaryotic and archaebacterial family B homologs. The DNA polymerase they sequenced, and which I call *S. solfataricus* P2 B2, was significantly different on the amino-acid level from the *S. solfataricus* MT4 B1 aphidicolin-sensitive DNA polymerase (Prangishvili and Klenk, 1994). Since the biochemical activities of the cloned *S. solfataricus* P2 B2 DNA polymerase were not studied, it is unclear if this DNA polymerase actually corresponds to the aphidicolin-resistant activity.

*S. solfataricus* and *Pyrodictium occultum* (both crenarchaeotes) remain the only archaebacteria from which two family B DNA polymerases have been described (Pisani et al., 1992; Prangishvili and Klenk, 1994; Uemori et al., 1995). Euryarchaeotes appear to have only a single family B DNA polymerase as suggested by the complete genome sequence of *Methanococcus jannaschii* (Bult et al. 1996). This chapter focuses on the phylogenetic relationship of the multiple crenarchaeote paralogs to other archaebacterial DNA polymerases. As well, this chapter discusses the phylogenetic and functional implications of a newly discovered family B DNA polymerase from the genomes of *S. solfataricus* P2 (Sensen et al., 1996) and *Sulfolobus shibatae.*

## Results

### *Sulfolobus solfataricus* P2 has three family B DNA polymerases

In the course of sequencing the genome of *S. solfataricus* P2, two open reading frames (ORFs) that were highly similar to family B DNA polymerases were found by BLAST and FASTA database searches (Sensen et al., 1996). One of the ORFs was 880 of 882 residues identical on the amino acid level to the *S. solfataricus* MT4 B1 polymerase. This ORF is designated *S. solfataricus* P2 B1. The second ORF was not specifically close at the amino acid level to any *S. solfataricus* (or other archaebacterial) DNA polymerase sequenced to date and represented an as-yet-undescribed family B DNA polymerase. This ORF is designated *S. solfataricus* P2 B3. *S. solfataricus* P2 is the first archaebacterium reported in which three family B DNA polymerases have been found.

The catalytic subunits of family B DNA polymerases are difficult to align due to short, highly conserved exonuclease and polymerase domains separated by long stretches of low or no amino acid conservation. The new *S. solfataricus* P2 B3 DNA polymerase amino acid sequence could be aligned with other archaeal and eukaryotic family B homologs except in exonuclease domain III and polymerase domain VI; these domains were excluded from phylogenetic analysis. Four additional domains not previously identified in other analyses of archaebacterial DNA polymerases, and designated A through D, were included in the alignment (figure 1.1).

As noted previously, the *S. solfataricus* P2 B2 sequence contains a number of unusual amino acid substitutions in polymerase and exonuclease domains (Prangishvili and Klenk, 1994); this is also true for the *S. solfataricus* P2 B3 sequence. Neither DNA polymerase has the consensus Asp-Ile-Glu (DIE) motif found in the 3'-5' exonuclease domain I of other archaeal and eukaryotic family B DNA polymerases (figure 1.1). These residues are critical

**Figure 1.1** Amino acid alignment of archaeal family B DNA polymerases. Numbering of sequences is from N- to C-terminal and corresponds to the amino acid position at the start of each conserved domain. Exonuclease domain II and polymerase domain IV overlap and a space has been inserted in the alignment to indicate the start of exonuclease domain II. Signature sequences that support a specific relationship of the *P. occultum* B3/*S. solfataricus* P2 B3/*S. shibatae* B3 with euryarchaeote DNA polymerases are boxed. Signature sequences that support a relationship of the *S. solfataricus* P2 B1 and *P. occultum* B1 DNA polymerases with euryarchaeotes are boxed and shaded. Amino acid residues that have been identified as functionally important by mutational studies are in bold. Not all of the alignment was used for phylogenetic analysis (see results) and the regions used are indicated by ~. Gaps introduced in the alignment are indicated by a period (.).

|  |  | Exonuclease |  |  |  |  | Exonuclease Domain II |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | Domain I |  | A |  | Polymerase Domain IV |  |  | B |  | C |
| S.solfataricus MT4 B1 | [225] | IKRVAIDIEVY | [246] | QKAEFPIISI | [292] | EYELLGRFFDILLEY | .P.IVLTFNGDDFDLPYIYFR | [326] | .ALKLGY | [347] | KYLAGLHIDL |
| S.solfataricus P2 B1 | [225] | IKRVAIDIEVY | [246] | QKAEFPIISI | [292] | EYELLGRFFDILLEY | .P.IVLTFNGDDFDLPYIYFR | [326] | .ALKLGY | [347] | KYLAGLHIDL |
| S.acidocaldarius B1 | [223] | IKRVSLDIEVY | [244] | ERAEFPIISV | [290] | EKKLLARLFEIIREY | .P.MLLTFNGDDFDIPYIYFR | [324] | .ALRLNF | [344] | KFLAGIHIDL |
| S.solfataricus P2 B2 | [16] | LEKLERIIERL | [51] | DELVDPYNLV | [151] | ..............Y | FYYMRKRLNVVN.ETPTVLSQ | [172] | TLYRLGI | [208] | SLKGKV.FEV |
| S.solfataricus P2 B3 | [166] | LRTIGVDFQIY | [183] | NPRKDPIVVM | [211] | DLKIIRRFVDYILNY | DPDIIFVYDSDLLPWKYITER | [247] | .ASSLGV | [273] | SISGRLNVDL |
| S.shibatae B3 | [166] | LRAIGIDFQIY | [183] | NPRKDPIVVL | [211] | DLKIIRKFVDYILNY | DPDIIYVFDVDVFHWKYITER | [247] | .ANSLGV | [273] | SISGRLNVDL |
| P.occultum B1 | [261] | PRRLAVDIEVF | [282] | STASYPVISV | [331] | ERALILEAFRLISNY | .P.VLLTFNGDNFDLPYLYNR | [365] | .AVKLGI | [385] | TLEYGFHIDL |
| P.occultum B3 | [181] | MRLVAFDIEVY | [198] | NPARDPVIIV | [226] | DRRVLREFVEYVRAF | DPDIIVGYNSNHFDWPYLMER | [262] | .ARRLGI | [288] | SVQGRLNVDL |
| Pyrococcus fur. | [135] | LKILAFDIETL | [151] | EFGKGPIIMI | [187] | EREMIKRFLRIIREK | DPDIIVTYNGDSFDFPYLAKR | [223] | .AEKLGI | [251] | EVKGRIHFDL |
| Pyrococcus sp. | [135] | LKLLAFDIETL | [151] | EFAKGPIIMI | [187] | EREMIKRFLKVIREK | DPDVIITYNGDSFDLPYLVKR | [223] | .AEKLGI | [251] | EIKGRIHFDL |
| Pyrococcus KOD1 | [135] | LKMLAFDIQTL | [151] | EFAEGPILMI | [187] | EREMIKRFLRVVKEK | DPDVLITYNGDNFDFAYLKKR | [223] | .CEKLGI | [251] | EVKGRIHFDL |
| Thermococcus lit. | [135] | LKLLAFDIETF | [151] | EFGKGEIIMI | [187] | EREMIKRFVQVVKEK | DPDVIITYNGDNFDLPYLIKR | [223] | .AEKLGV | [253] | EIKGRIHFDL |
| Thermococcus 9oN-7 | [135] | LTMLAFDIETL | [151] | EFGTGPILMI | [187] | EKEMIKRFLRVVREK | DPDVLITYNGDNFDFAYLKKR | [223] | .CEELGI | [251] | EVKGRIHFDL |
| Methanococcus vol. | [175] | LNCIAFDMELY | [191] | NAKKDPIIMV | [231] | EKELIQKTIEIL..K | QYDVIYTYNGDNFDFPYLKKR | [265] | .ANIYEI | [298] | KIPGIIHIDL |
| Methanococcus jan. | [154] | LKSVAFDMEVY | [171] | NPERDPILMA | [207] | EKELIKKIIETL..K | EYDVIYTYNGDNFDFPYLKAR | [241] | .AKIYGI | [269] | YIPGRVHIDL |

|  |  | Exonuclease |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  |  | Domain III |  | Polymerase Domain II |  | D |  | Polymerase Domain VI |
| S.solfataricus MT4 B1 | [406] | LIEYNFRDAEITLQL | [496] | GAVVIDPPAGIFFNITVLDFASLYPSIIRT.WNLSYETV | [545] | KDETGEVLHIVCMDRP | [561] | GITAVITGLLRDFR..VKIYKKKA |
| S.solfataricus P2 B1 | [406] | LIEYNFRDAEITLQL | [496] | GAVVIDPPAGIFFNITVLDFASLYPSIIRT.WNLSYETV | [545] | KDETGEVLHIVCMDRP | [561] | GITAVITGLLRDFR..VKIYKKKA |
| S.acidocaldarius B1 | [403] | LIEYNLRDAEITLKL | [493] | GAVVIDPPAGVYFNVVVLDFASLYPSIIKN.WNISYETI | [542] | EDETGEKLHYVCMDKP | [558] | GITAVYQGLIRDFR..VKVYKKKA |
| S.solfataricus P2 B2 | [226] | LIEWS.......... | [326] | GGLILFPQPGCYDNVYQVDFSSMYPSLIVK.HNISAETV | [366] | CDDIKTELHSICLKEK | [382] | GIIPEALQWLIERKSELK...... |
| S.solfataricus P2 B3 | [331] | IREYSIENARSIYLL | [406] | KNIIIQPKVGIYTDVYVLDISSVY.SLVIRKFNIAPDTL | [451] | CYSSPISNYKFKREPS | [467] | GLYKTFLDELSNVR.......... |
| S.shibatae B3 | [331] | VKQYSLENAKSIYLL | [406] | KKTVIEPKIGIYSDVYVLDISSVY.LSVIRKFNISPDTL | [451] | CYVSTISNYKFKKEPS | [467] | GLYKTFLEELSNIQ.......... |
| P.occultum B1 | [444] | LVRYNVRDADLTLRL | [534] | GALVLDPPSGIYFNIVVLDFASLYPSIIKR.WNLSYETV | [583] | .EV.PDVGHKVCMSIP | [597] | GLTSQIVGLLRDYR..VKIYKKKA |
| P.occultum B3 | [348] | LERYALDDVRATYGL | [421] | GAVVLKPLKGVHENVVVLDFSSMYPSIMIK.YNVGPDTI | [471] | CYVAPEVGHRFRRSPP | [487] | GFFKTVLENLLKLRRQVKEKMKEF |
| Pyrococcus fur. | [308] | VAKYSMEDAKATYEL | [387] | GGFVKEPEKGLWENIVYLDFRALYPSIIIT.HNVSPDTL | [432] | YDIAPQVGHKFCKDIP | [448] | GFIPSLLGHLLEERQKIKTKMKET |
| Pyrococcus sp. | [308] | VAKYSMEDAKVTYEL | [387] | GGYVKEPEKGLWEGLVSLDFRSLYPSIIIT.HNVSPDTL | [432] | YDVAPEVGHKFCKDFP | [448] | GFIPSLLKRLLDERQEIKRKMKAS |
| Pyrococcus KOD1 | [308] | VARYSMEDAKVTYEL | [386] | GGYVKEPERGLWENIVYLDFRSLYPSIIIT.HNVSPDTL | [431] | YDVAPQVGHRFCKDFP | [447] | GFIPSLLGDLLEERQKIKKKMKAT |
| Thermococcus lit. | [310] | LAQYSMEDARATYEL | [389] | GGYVKEPEKGLWENIIYLDFRSLYPSIIVT.HNVSPDTL | [434] | YDVAPIVGYRFCKDFP | [450] | GFIPSILGDLIAMRQDIKKKMKST |
| Thermococcus 9oN-7 | [308] | VARYSMEDAKVTYEL | [386] | GGYVKEPERGLWDNIVYLDFRSLYPSIIIT.HNVSPDTL | [431] | YDVAPEVGHKFCKDFP | [447] | GFIPSLLGDLLEERQKIKRKMKAT |
| Methanococcus vol. | [353] | LLRYAYEDALYTYKM | [432] | GGYVREPLKGIQEDIVSLDFMSLYPSILIS.HNISPETV | [477] | .ENM.EL......... | [482] | GIIPKTLNELLSRRKHIKMLLKDK |
| Methanococcus jan. | [324] | LIEYSLQDAKYTYKI | [403] | GGYVKEPEKGMFEDIISMDFRSLYPSIIIS.YNISPDTL | [442] | ........CECCKDVS | [463] | GLIPKTLRNLIERRINIKRRMKKM |

|  |  |  |  | Polymerase |  | Polymerase |  | Polymerase |
|---|---|---|---|---|---|---|---|---|
|  |  | Polymerase Domain III |  | Domain I |  | Domain VII |  | Domain V |
| S.solfataricus MT4 B1 | [599] | QRAMKVFINATYGVFGAETFPLYAPRVAESVTALGRY | [648] | GLTVLYGDTDSLFL | [692] | FVAFSGLKKNYFGVY | [707] | QDGKVDIKGMLVKKRNTPEFVK |
| S.solfataricus P2 B1 | [599] | QRAMKVFINATYGVFGAETFPLYAPAVAESVTALGLR | [648] | GLTVLYGDTDSLFK | [692] | FVAFSGLKKNYFGVY | [707] | QDGKVDIKGMLVKKRNTPEFVK |
| S.acidocaldarius B1 | [596] | QRAMKVFINATYGVFGAENFPLYAPAVAESVTAIGRY | [645] | NLKVIYGDTDSLFL | [689] | YVAYSGLKKNYFGVY | [704] | PDGKTEIKGMLAKKRNTPEFIK |
| S.solfataricus P2 B2 | [406] | AEAIKWILVASFGYLGYRNSLFGKIEAYEMVTYLARK | [455] | GLRVLHGIIDSLVV | [482] | KET..GLRKRYNWII | [515] | MNGEMIAKGL..IRENMPNIVK |
| S.solfataricus P2 B3 | [485] | IKVIEELISSFNDYVHWVNARWYSREIA.SAFDEFSN | [534] | GLDVILANDLLIFV | [577] | KSLLVLDNNRYAGLS | [592] | EGDKIDIARKGEEDMNLCELAR |
| S.shibatae B3 | [485] | SKVIEELMSSFYDYIHWINSRWYSREIA.SAVDELSY | [534] | GFEVILANDFLVFV | [577] | RSLLILGNDRYAGLL | [592] | EGDKIDIARIGKEDRDLCELVR |
| P.occultum B1 | [635] | QAAMKVYINASYGVFGAESFPFYAPPVAESVTAIGRY | [684] | GLRVLYGDTDSLFI | [728] | FVTFSGLKKNYIGAY | [743] | EDGSIDVKGMVAKKRNTPEFLK |
| P.occultum B3 | [524] | QKALKVLANASYGYMGWSHARWYCKRCAEAVTAWGRN | [573] | GLKVIYGDTDSLFV | [616] | KVFFTEAKKRYVGLL | [631] | EDGRIDIVGFEAVRGDWCELAK |
| Pyrococcus fur. | [484] | QKAIKLLANSFYGYYGYAKARWYCKECAESVTAWGRK | [534] | GFKVLYIDTDGLYA | [586] | RGFFV.TKKRYAVID | [600] | EEGKVITRGLEIVRRDWSEIAK |
| Pyrococcus sp. | [484] | QRAIKILANSYYGYYGYAKARWYCKECAESVTAWGRE | [534] | GFKVLYIDTDGLYA | [586] | RGFFV.TKKKYALID | [600] | EEGKIITRGLEIVRRDWSEIAK |
| Pyrococcus KOD1 | [483] | QRAIKILANSYYGYYGYARARWYCKECAESVTAWGRE | [533] | GFKVIYSDTDGFFA | [585] | RGFFV.TKKKYAVID | [599] | EEGKITTRGLEIVRRDWSEIAK |
| Thermococcus lit. | [486] | QRAIKLLANSYYGYMGYPKARWYSKECAESVTAWGRH | [536] | GFKVLYADTDGFYA | [588] | RGFFV.TKKRYAVID | [602] | EEGRITTRGLEVVRRDWSEIAK |
| Thermococcus 9oN-7 | [483] | QRAIKILANSFYGYYGYAKARWYCKECAESVTAWGRE | [533] | GFKVLYADTDGLHA | [585] | RGFFV.TKKKYAVID | [599] | EEGKITTRGLEIVRRDWSEIAK |
| Methanococcus vol. | [522] | QKSIKVLANSHYGYLAFPMARWYSDKCAEMVTGLGRK | [571] | GFKVIYADTDGFYA | [645] | RGLFV.TKKKYALIE | [659] | DDGHIVVKGLEVVRRDWSNIAK |
| Methanococcus jan. | [503] | QKSLKILANSVYGYLAFPRARFYSRECAEIVTYLGRV | [522] | GFKVLYIDTDGFYA | [605] | RGIFV.TKKRYALVT | [627] | NGRVTVVKGLEFVRRDWSNIAK |

for exonuclease activity, since introduction of Asp➤Glu or Glu➤Ala substitutions in exonuclease domain I of the *Thermococcus litoralis* family B DNA polymerase abolishes exonuclease activity (Kong et al., 1993). Mechanistic studies with other DNA polymerases indicate that these acidic amino acids play crucial roles in exonuclease activity and are responsible for coordination of divalent metal ions (Beese and Steitz, 1991; De Vega et al., 1996). The absence of this exonuclease domain has been noted before in other family B DNA polymerases, notably all of the eukaryotic α DNA polymerases, but these homologs still retain polymerase functions (Braithwatie and Ito, 1993; Kornberg and Baker, 1992). It is possible that the 3'-5' exonuclease activity in *S. solfataricus* P2 is performed by the *S. solfataricus* P2 B1 paralog which does possess a consensus exonuclease domain I sequence.

Both the B2 and B3 sequences from *S. solfataricus* P2 exhibit a number of nonconserved substitutions in two metal-binding polymerase domains (I and II). The amino acid motif Asp-Thr-Asp (DTD), which is present in polymerase domain I of all other archaeal enzymes, is replaced by Ile-Ile-Asp and Asn-Asp-Leu in *S. solfataricus* P2 B2 and B3 respectively (figure 1.1). Copeland and Wang found that mutation of the Asp-Thr-Asp motif to Asn-Thr-Asp, Asp-Ser-Asp, or Asp-Thr-Asn in human DNA polymerase α drastically reduced polymerase activity (Copeland and Wang, 1993). Dong and Wang also found that Lys-950 of human DNA polymerase α is essential for the binding of deoxynucleoside triphosphates (Dong and Wang, 1995). The *S. solfataricus* P2 B2 sequence possesses this amino acid in the homologous position, but it is replaced by a Glu residue in the *S. solfataricus* P2 B2 and *S. shibatae* B3 sequences (see below).

### *Sulfolobus shibatae* possess a rapidly evolving ortholog of the *S. solfataricus* P2 B3 DNA polymerase

*S. solfataricus* P2 is the first archaebacterium reported to have three family B DNA polymerases. However, the extremely divergent amino acid sequence of the B3 DNA polymerase raises the question of whether this gene actually codes for a functional DNA polymerase. In an attempt to address this issue, I have cloned and sequenced an ortholog of the *S. solfataricus* P2 B3 DNA polymerase from a closely related member of the archaebacterial order *Sulfolobales*.

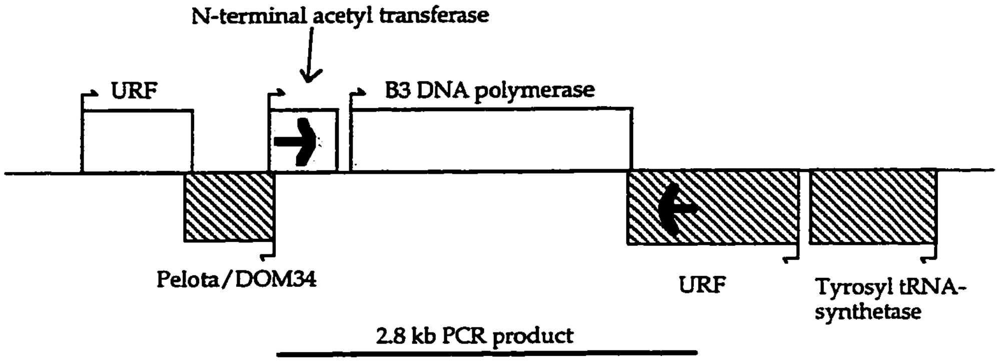
Using the *S. solfataricus* P2 genome sequence (Sensen et al., 1996), I designed nondegenerate PCR primers flanking the B3 ORF and attempted to amplify this genomic region from representatives of the *Sulfolobales* (figure 1.2 a). Bands of the expected size (2.8 kb) were consistently amplified from *S. solfataricus* P1, *S. solfataricus* MT4, and *S. shibatae* but not from *Sulfolobus acidocaldarius* (not shown). Of the organisms from which I could obtain amplification, *S. shibatae* is the most distant from *S. solfataricus* P2 on the basis of 16S rRNA phylogeny (Fuchs et al., 1996). I cloned and sequenced the *S. shibatae* PCR product and found it to be identical to the genomic region from *S. solfataricus* P2 surrounding the B3 DNA polymerase in gene identity and order (figure 1.2 b). However, the predicted amino acid sequence of the *S. shibatae* B3 DNA polymerase is only 79% identical to the *S. solfataricus* P2 B3 sequence (158 of 764 amino acids are different).

This number of amino acid substitutions, both conserved and nonconserved, was surprising. Protein-coding genes evolving neutrally and under stabilizing (purifying) selection for maintenance of a function will have high synonymous (no change of amino acid) substitution rates and low nonsynonymous (change of amino acid) rates (Li, 1997). Protein-

**Figure 1.2** (A) Schematic representation (not to scale) of the genomic region of *Sulfolobus solfataricus* P2 surrounding the B3 DNA polymerase ORF. Boxes indicate ORFs identified as potential coding regions. Names above or below boxes indicate ORFs that have significant matches in databases. URF indicates unidentified open reading frame. Direction of transcription is indicated by arrows. Solid arrows indicate the approximate location of direct match PCR-primers used to amplify the region from *S. shibatae.*

(B) Amino acid alignment of the orthologous B3 DNA polymerase from *S. shibatae* (S. shib-B3) with the *S. solfataricus* P2-B3 DNA polymerase (S. solf P2-B3). Conserved or non-conserved amino acid substitutions are highlighted in bold. Conserved functional regions are boxed. Polymerase domain IV and Exonuclease domain II overlap; amino acids corresponding to the polymerase domain are shaded.

**A**



**B**

```
                                Exonuclease
                  Exonuclease                                           Domain II
                   Domain I                               Polymerase Domain IV
S.solf P2 B3   MIKDFFILDFSYEIKGNTPLVYIWSVDDEGNSSVVIDNNFRPYFYIIYEGNENEIIENIKKNCEALQITKVKRKYLGNIV
S.shib B3      MIKDFFILDFSYEIKDNIPLIYIWSIDDEGNSCVVVERNFKPYFYVVYEGNGDEIIENIRKNCEVLLITKVKRKYLGNVV

S.solf P2 B3   ALLIQTSTPTQIKKCREKISELNNIKGIFDADIRYTMRYSLDFDLRPFTWFRAEVNEVKFDGFRTKKAYILDKILSHYEG
S.shib B3      ALLVQTFTPTQIKRCREKISRINGIKSIFDADIRFTMRYSIDFDLRPFTWFKAEVSEVKLEGFRAKKVYILDKILSHYEG

                                                                        Exonuclease
                  Exonuclease                                            Domain II
                   Domain I                               Polymerase Domain IV
S.solf P2 B3   NMPELRTIGVDFQIYSKYGSLNPRKDPIVVMSLWSKEGPMQFSLDEGIDDLKIIRRFVDYILNYDPDIIFVYDSDLLPWK
S.shib B3      KIPELRAIGIDFQIYSKYGSLNPRKDPIVVLSLWSKEGSMQFSLDESMDDLKIIRKFVDYILNYDPDIIYVFDVDVFHWK

S.solf P2 B3   YITERASSLGVKIDIGRKIGSEVSVGTYGHYSISGRLNVDLTGLLVNERSLGHVDLIDVSNYLGISPSRYSFKWYEISRY
S.shib B3      YITERANSLGVKIDIGRKIGSEVSQGTYGHYSISGRLNVDLVGLLMNERLTGHIDLIEVANYLGISPKRDSLNWYEISRY

                        Exonuclease
                         Domain III
S.solf P2 B3   WDNEKNRRIIREYSIENARSIYLLGNYLLSTYSELVKIVGLPLDKLSVASWGNRIETSLIRTATKSGELIPIRMDNPNRP
S.shib B3      WDDEKNRDLVKQYSLENAKSIYLLGNFLLSPYSELVKIIGLPLDKLSVASWGNRIEASLIRTAAKSEELIPIRMDNPNRS

                                                                             Polymerase
                        Polymerase Domain II                                 Domain VI
S.solf P2 B3   SKIKKNIIIQPKVGIYTDVYVLDISSVYSLVIRKFNIAPDTLVKEQCDDCYSSPISNYKFKREPSGLYKTFLDELSNVRD
S.shib B3      SKIKKTVI.EPKIGIYSDVYVLDISSVYLSVIRKFNISPDTLVKGQCDDCYVSTISNYKFKKEPSGLYKTFLEELSNIQD

                                                                   Polymerase
                    Polymerase Domain III                          Domain I
S.solf P2 B3   SNKIKVIEELISSFNDYVHWVNARWYSREIASAFDEFSNEIIRFIIDLIKSSGLDVILANDLLIFVTGGSRDKVNELITK
S.shib B3      TRKSKVIEELMSSFYDYIHWINSRWYSREIASAVDELSYEIGKLVIDLIKNSGFEVILANDFLVFVKGGSGDKLNELIFK

                           Polymerase       Polymerase
                           Domain VII       Domain V
S.solf P2 B3   INSLYNLDVKVKIFYKSLLVLDNNRYAGLSEGDKIDIARKGEEDMNLCELARNIKRKIIEEILISKDVKKAIKLVKSTVI
S.shib B3      INSLYDLNLKVRKIYRSLLILGNDRYAGLLEGDKIDIARIGKEDRDLCELVRNVKRKVVEEILISKDVKKAVKLVKSAVI

S.solf P2 B3   KLRRGEFDNEELITWAKIERDLNEYNNQLPFVTAARKAIQSGYLISKDSKIGYVIVKGLGPLNDRAEPFFLVKEKNRIDI
S.shib B3      KLRRGEFDIGELITWVHIEKDFSEYDKQLPFVVAARKAIQSGYLISKDSRIGYLIVKGHGSVHDRAEPFFFVKEKNRIDI

S.solf P2 B3   EYYVDQIFRETLKLLKPLGVNEESLKKTNITDILDLFGASKKK
S.shib B3      EYYVDQLLRESLKVLTPLGVSEESLKKTNITDILDMFGASKKK
```

coding genes that are not evolving at neutral rates will have a higher rate of nonsynonymous substitutions and a lower rate of synonymous substitutions per site. I calculated both synonymous ($K_S$) and nonsynonymous ($K_A$) nucleotide substitution rates for the B3 DNA polymerase and for the only other protein-coding gene sequenced from both organisms, β-galactosidase (table 1.1). The synonymous substitution rate of β-galactosidase is typical of protein-coding genes evolving neutrally (Li, 1997). The substitution rates of the B3 DNA polymerase gene suggest that functional constraints have been relaxed, allowing these genes to accumulate nonsynonymous substitutions. However, the rate of nonsynonymous substitutions is not high enough to suggest that the proteins are under positive selection for a novel function (Nei, 1987).

### Southern hybridization suggests that members of the *Sulfolobales* also possess an ortholog of the rapidly evolving *S. solfataricus* P2 B2 polymerase

As with the *S. solfataricus* P2 B3 paralog, the high number of unusual amino acid substitutions in the *S. solfataricus* P2 B2 paralog relative to other archaebacterial paralogs raises the question of whether or not this is a functional gene. It is possible that this paralog is the result of a very recent gene duplication within the *Sulfolobus* lineage; the divergent sequence could be the result of relaxation of any stabilizing selection for maintenance of function and the subsequent accumulation of random nucleotide substitutions. However, two observations argue against this possibility. First, the ORF is intact and continuous. If stabilizing selection has been relaxed or lost, then nonsense or missense codons would be expected to appear at a certain frequency; none is present. Second, consensus archaebacterial promoters (figure 1.3) and terminators (not shown) are found 5' to the

| | # nts | # nsy diff | # syn diff | nsy sub rate | syn sub rate | ratio nsy/syn |
|---|---|---|---|---|---|---|
| B3 DNAP | 2292 | 183.5 | 201.5 | 0.111 ± 0.0083 | 0.572 ± 0.0468 | 0.192 |
| β-galactosidase | 1467 | 41.0 | 171.0 | 0.037 ± 0.058 | 0.886 ± 0.0898 | 0.042 |

**Table 1.1** Nucleotide substitution rates for two protein-coding genes from *S. solfataricus* P2 and *S. shibatae*

```
S. solfataricus MT4 B1  taaaaCTTATAgcgtatttctcagaaaataatAtAtgttagaaaATG
 S. solfataricus P2 B1  taaaaCTTATAgcgtatttctcagaaaataatAtAtgttagaaaATG
 S. solfataricus P2 B2  aaagaCTTAATttaccagaggagagAtgtaacAcATG
 S. solfataricus P2 B3  aagaaTTTATAttataaatatctggattaattgttAa[42 bps]atATG
        P. occultum B1  gaactGTTATCggaaatatcctcatctaggagAcgcg[51 bps]gcATG
        P. occultum B3  atacgATTATGtaggggcgggtggtggtagAttctccagggcagagccagcccATG
      Pyrococcus fur.   aaggtTTTATActccaaactgagttagtagAtAtgtggggagcAtaATG
       Pyrococcus sp.   gcgttCTTAAAggCTTAAAtacgtgaatttagcgtaaAttAttgagggattaagtATG
     Thermococcus lit.  gggggTTTAAAaatttggcggaacttttaTTTAATttgaactccagtttatatctggtggtAtttATG

     archaebacterial         t   t
           consensus          tta a
                             c   a
```

**Figure 1.3** Putative promoter motifs in the 5' regions of archaeal family B DNA polymerases (22). Sequences which are capitalized and underlined correspond to the consensus archaeal BoxA motif. Nucleotides that are bold indicate a pyrimidine/purine pair and the probable site of transcription initiation. Start codons of the DNA polymerase ORFs are capitalized.

proposed start codon and Prangishvili and Klenk have mapped the transcription start site by pimer extension analysis (Prangishvili and Klenk, 1993). Although Southern hybridization with the *S. solfataricus* P2 B2 paralog as a probe to DNA of other *Sulfolobales* cannot address issues of functionality or gene expression, it can address whether the duplication event that gave rise to this paralog occurred within the *Sulfolobus* lineage, or before.

Genomic DNAs from various members of the *Sulfolobales* spanning the phylogenetic breadth of the order were used for Southern analysis. In addition, genomic DNAs from *Desulfurococcus mobilis* and *Hyperthermus butylicus*, representative of the next two closest crenarchaeote orders, were also used in Southern analysis. At low stringency hybridization conditions (50°C hybridization, and washes with 6XSSC), the *S. solfataricus* P2 B2 probe, generated by PCR using direct match primers, hybridized to all *Sulfolobales* DNA (figure 1.4). *Eco*RI and *Hind*III were specifically chosen because there are sites for each enzyme in the *S. solfataricus* P2 B2 GenBank sequence; each digest produced two hybridizing bands. Only a single strongly hybridizing band was observed for other *Sulfolobales* DNAs suggesting that this gene is single copy and too divergent in sequence to hybridize to paralogous DNA polymerase sequences. Two bands from *S. acidocaldarius Eco*RI and *Hind*III digests hybridized with high intensity to the probe as compared to other *Sulfolobales* DNA. This is likely due to unequal loading of genomic DNA. Non-specific hybridization was observed to genomic DNA from *D. mobilis* and *H. butylicus* and could be washed away by increasing the temperature of washes by 5°C from 50°C to 55°C (data not shown).

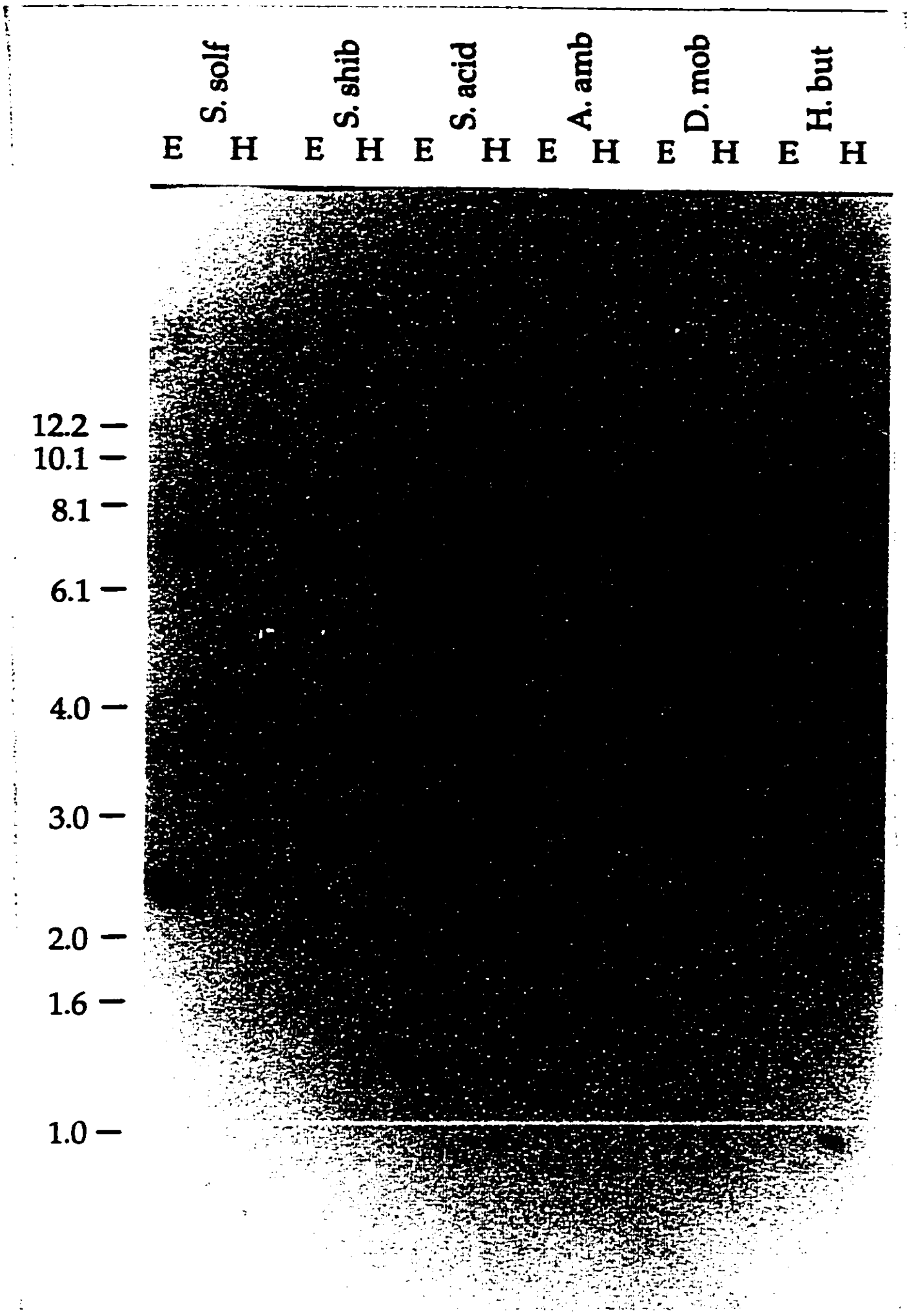


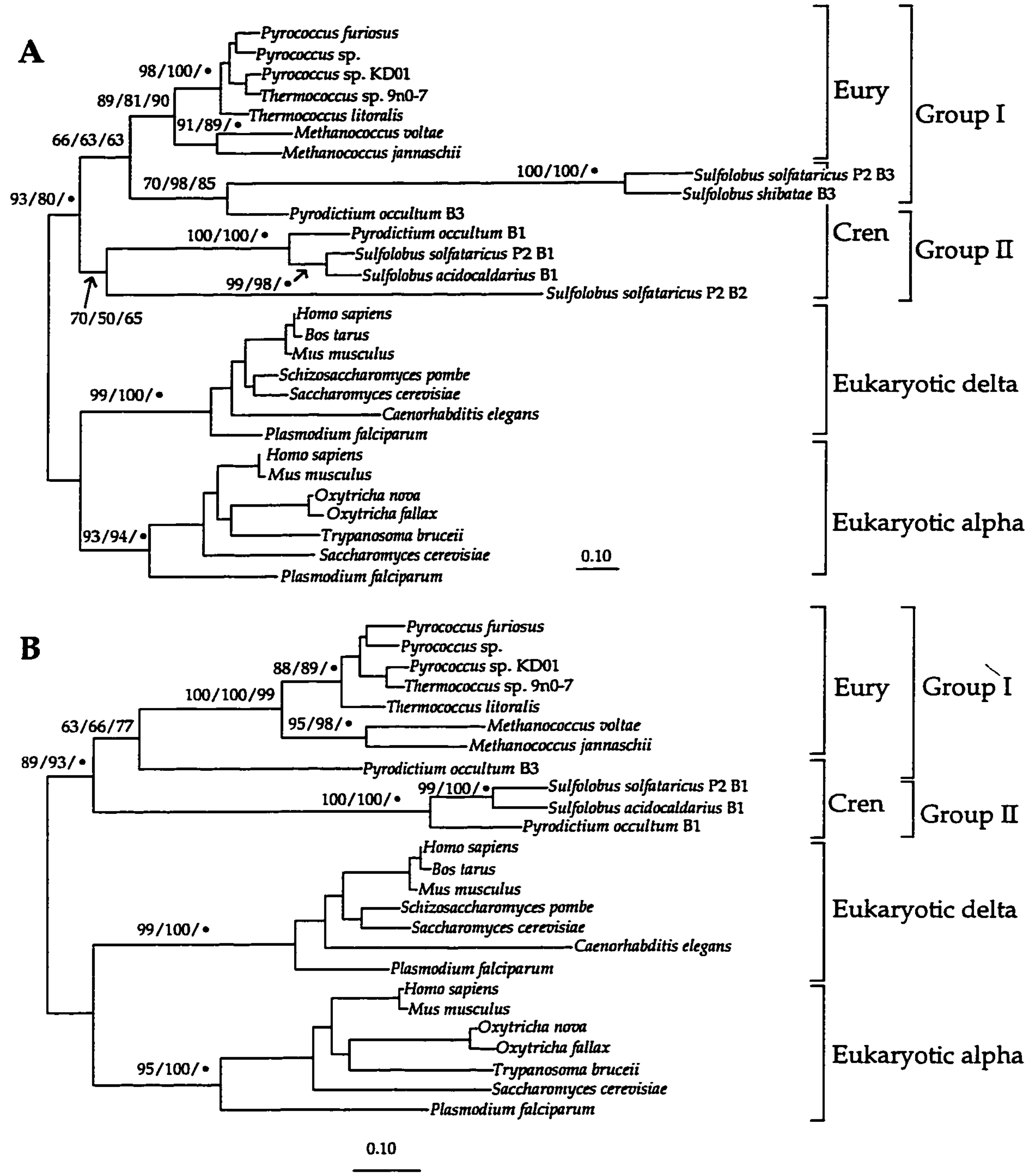
**Figure 1.4** Southern hybridization of *S. solfataricus* P2 B2 DNA polymerase against various crenarchaeote genomic DNAs exposed for 24 hours. S. solf, *S. solfataricus* P2; S. shib, *S. shibatae*; S. acid, *S. acidocaldarius*; A. amb, *Acidianus ambivalens*; D. mob, *Desulfurococcus mobilis*; H. but, *Hyperthermus butylicus*.

## Phylogenetic analysis of archaebacterial DNA polymerases is confounded by rapid rates of sequence evolution

To determine the relationships of the three *S. solfataricus* P2 family B DNA polymerases to other archaeal DNA polymerases, phylogenetic analysis was performed on a data set that contained all available archaebacterial family B DNA polymerase sequences. These include family B paralogs from euryarchaeotes, the *S. solfataricus* P2 B1, B2, and B3 paralogs, a B1 DNA polymerase from *S. acidocaldarius*, and the *S. shibatae* B3 paralog. The *S. solfataricus* MT4 B1 paralog was not included in phylogenetic analysis since it is 99.7% identical at the amino acid level to its ortholog from *S. solfataricus* P2. Two paralogs (designated A and B in Uemori et al., 1995) from the crenarchaeote *Pyrodictium occultum* were also included in analyses. *P. occultum* A and *S. solfataricus* P2 B1 are orthologs, but *P. occultum* B appears most related to *S. solfataricus* P2 B3 (see below). The two family B DNA polymerases from *P. occultum* have been renamed *P. occultum* B1 and B3 respectively to account for their relationship to *S. solfataricus* family B paralogs.

Regardless of the phylogenetic method used, the three *S. solfataricus* P2 DNA polymerases did not branch together, as would have been expected if they were related by recent gene duplications (figure 1.5 a, b). In both parsimony and distance analyses, the *S. solfataricus* P2 B3 DNA polymerase and *S. shibatae* B3 sequences grouped together with the *P. occultum* B3 (formerly B) DNA polymerase with moderate bootstrap support. The *S. solfatarricus* P2 B1 and *S. acidocaldarius* B1 paralogs form a separate group with *P. occultum* B1 (formerly A). High bootstrap support for this group was obtained. These results suggest that the *S. solfataricus* P2 B3, *S. shibatae* B3, and *P. occultum* B3 DNA polymerases are all orthologs as are the *S.*

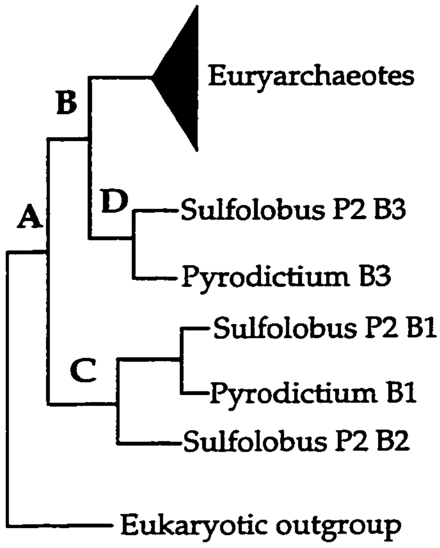
**Figure 1.5** Phylogenetic analysis of archaeal and eukaryotic family B DNA polymerases. (A) PROTDIST analysis with all taxa. An identical topology was found using parsimony (100 random replicates of stepwise addition, shortest tree was 748 steps, CI=0.741, HI=0.259). Based on the phylogeny obtained by parsimony and distance analyses, a partially constrained tree was used for a maximum likelihood search with PROTML. Nodes which were constrained in the maximum likelihood analysis are indicated by a bullet (•). Bootstrap values for nodes are indicated in the order parsimony/distance/maximum likelihood. Euryarchaeote and crenarchaeote are abbreviated to EURY and CREN respectively. (B) PROTDIST analysis with the rapidly evolving *S. solfataricus* P2 B2 and B3 sequences removed. An identical tree was found using parsimony and maximum likelihood methods. Bootstrap values are as above.


A
98/100/•
89/81/90
66/63/63
91/89/•
93/80/•
70/98/85
100/100/•
100/100/•
99/98/•
70/50/65
99/100/•
93/94/•
Pyrococcus furiosus
Pyrococcus sp.
Pyrococcus sp. KD01
Thermococcus sp. 9n0-7
Thermococcus litoralis
Methanococcus voltae
Methanococcus jannaschii
Sulfolobus solfataricus P2 B3
Sulfolobus shibatae B3
Pyrodictium occultum B3
Pyrodictium occultum B1
Sulfolobus solfataricus P2 B1
Sulfolobus acidocaldarius B1
Sulfolobus solfataricus P2 B2
Homo sapiens
Bos tarus
Mus musculus
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Caenorhabditis elegans
Plasmodium falciparum
Homo sapiens
Mus musculus
Oxytricha nova
Oxytricha fallax
Trypanosoma bruceii
Saccharomyces cerevisiae
Plasmodium falciparum
0.10
Eury
Group I
Cren
Group II
Eukaryotic delta
Eukaryotic alpha
B
88/89/•
100/100/99
63/66/77
95/98/•
89/93/•
99/100/•
100/100/•
99/100/•
95/100/•
Pyrococcus furiosus
Pyrococcus sp.
Pyrococcus sp. KD01
Thermococcus sp. 9n0-7
Thermococcus litoralis
Methanococcus voltae
Methanococcus jannaschii
Pyrodictium occultum B3
Sulfolobus solfataricus P2 B1
Sulfolobus acidocaldarius B1
Pyrodictium occultum B1
Homo sapiens
Bos tarus
Mus musculus
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Caenorhabditis elegans
Plasmodium falciparum
Homo sapiens
Mus musculus
Oxytricha nova
Oxytricha fallax
Trypanosoma bruceii
Saccharomyces cerevisiae
Plasmodium falciparum
0.10
Eury
Group I
Cren
Group II
Eukaryotic delta
Eukaryotic alpha

*solfataricus* P2 B1, *S. acidocaldarius* B1, and *P. occultum* B1 polymerases. It is not clear to which archaebacterial DNA polymerase the *S. solfataricus* P2 B2 sequence is related. Low bootstrap support was found for this sequence grouping with the *S. solfataricus* P2 B1, *S. acidocaldarius* B1, and *P. occultum* B1 DNA polymerases. If this placement is correct, a crenarchaeote-specific gene duplication event must have occurred to give rise to the *S. solfataricus* P2 paralog.

Parsimony, distance, and maximum likelihood analyses split the archaebacterial DNA polymerases into two groups; group I includes euryarchaeote sequences and the *S. solfataricus* B2 B3-*S. shibatae* B3-*P. occultum* B3 clade, group II includes the remaining crenarchaeote sequences (figure 1.5 a, b). This tree topology is consistent with the last common ancestor of archaebacteria possessing at least two family B DNA polymerases. However, given the highly divergent amino acid sequences of the *S. solfataricus* P2 B2, B3 and *S. shibatae* B3 sequences compared to other archaebacterial DNA polymerases, it is possible that the tree topology found by all methods was artefactual. To test this possibility, the *S. shibatae* B3 sequence was eliminated from phylogenetic analyses. Removal of this sequence did not result in tree topologies different from those obtained when it was included (not shown); the *S. shibatae* B3 sequence was not included in any further phylogenetic analyses. The same rationale was applied to removing the *S. acidocaldarius* B1 sequence from further phylogenetic analyses. With the reduced data set, each of the three *S. solfataricus* P2 paralogs was separately removed from the analysis and the effect on tree topology measured by both parsimony and distance bootstrap analyses. Figure 1.6 indicates that removal of the *S. solfataricus* P2 B3 sequence from the analysis had the greatest effect since bootstrap values at the node supporting

Shortest tree



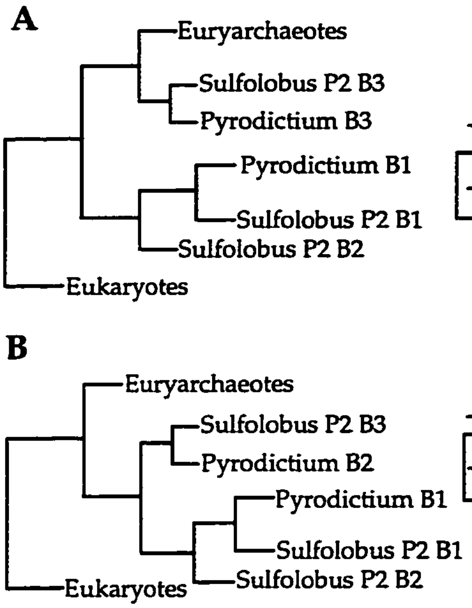
| Species excluded | Bootstrap at node A | B | C | D |
|---|---|---|---|---|
| None | 90/69 | 68/80 | 66/37 | 76/96 |
| Sulf P2 B1 | 92/71 | 64/83 | 67/47 | 61/100 |
| Sulf P2 B2 | 87/94 | 85/89 | ✗ | 67/100 |
| Sulf P2 B3 | 86/76 | 47/57 | 46/42 | ✗ |

**Figure 1.6** Effect of removing rapidly evolving taxa from phylogenetic analyses. The optimal tree found by parsimony and distance analyses is drawn schematically. Important nodes are indicated by letters A through D. A=archaeal unity, B=support for group I, C=support for group II, D=affinity of *S. solfataricus* P2 B3 and *P. occultum* B3. Confidence for each node after removal of rapidly evolving taxa was measured by 100 bootstrap replicates by both parsimony and distance methods.

group I (node B) and group II (node B) were reduced.

The long branch lengths of the *S. solfataricus* P2 B2 and B3 DNA polymerases were also of concern, since rapidly evolving sequences such as these are known to be positively misleading in phylogenetic reconstruction (Felenstein, 1978; Huslsenbeck, 1997). I was particularly interested in testing an alternative tree topology which would be consistent with the common ancestor of archaebacteria possessing a single family B DNA polymerase. This tree topology, which unites all crenarchaeote sequences to the exclusion of euryarchaeote sequences, was not significantly worse than the shortest tree topology since both the Kishino-Hasegawa and Templeton tests did not reject the alternative topology at the 5% significance level (figure 1.7 a).

I also removed the *S. solfataricus* P2 B2 and B3 sequences and performed parsimony, distance, and maximum likelihood analyses to find the best tree topology. If the branching order observed with all taxa is robust, removal of these two taxa should not result in a significantly different tree topology. Indeed, the best tree recovered by all methods was identical to the best tree recovered when all taxa were included, with the archaebacterial DNA polymerase split into two groups (figure 1.5 b). However, an alternative topology uniting all crenarchaeote sequences as a group and consistent with the last common ancestor of archaebacteria possessing a single family B DNA polymerase was not significantly worse than the best tree (figure 1.7 b). The lack of phylogenetic resolution cannot be attributed only to the rapidly evolving *S. solfataricus* P2 paralogs; other factors must also be contributing to the lack of phylogenetic resolution.



| ln L | Δ ln l/S.E. | No. of steps | S.D. | Sig. worse? |
|---|---|---|---|---|
| -3166.1 | 0.0 | 943 | 0.0 | BEST |
| -2447.2 | 0.0 | 725 | 0.0 | BEST |
| -3169.6 | -1.06 | 948 | 2.65 | NO |
| -2451.6 | -1.1 | 728 | 4.13 | NO |

**Figure 1.7** An alternative topology (tree B) consistent with the hypothesis that the common ancestor of archaea had only a single family B DNA polymerase is not significantly worse than the best tree (tree A). Abbreviations used are ln L, log likelihood; Δln L, difference in log likelihood; and S.E., standard error. Alternative topologies are considered significantly worse if the Kishino-Hasegawa and Templeton tests reject these topologies at the 5% significance level (Adachi and Hasegawa, 1992; Felsenstein, 1996). The boxed values are those obtained when the *S. solfataricus* P2 B2 and B3 sequences were removed from the analysis.

## Discussion

As only two DNA polymerase activities were found in cell extracts of *S. acidocaldarius* and *S. solfataricus* (Elie et al., 1989; Klimczak et al., 1985; Rossi et al., 1985), it is perhaps surprising that a gene coding for a third family B DNA polymerase was found. The amino acid sequence of this DNA polymerase is extremely divergent, raising the question of whether this is actually the product of a functional gene. I addressed this question by cloning and sequencing an ortholog of this gene from *S. shibatae*, a closely related member of the *Sulfolobales*, reasoning that if *S. shibatae* also possesses this divergent DNA polymerase, it is likely to have some function. Database search scores (not shown) and the alignment in figure 1.2 convincingly show that the genes encoding these proteins evolved from an archaebacterial family B DNA polymerase. In addition, putative BoxA motifs, which are known to be essential for transcription initiation, could be found in the 5′ noncoding regions of both the *S. solfataricus* P2 B3 and *S. shibatae* B3 paralogs (figure 1.1). Zillig and co-workers have shown through extensive mutational studies of a number of archaebacterial genes that the BoxA sequence is spaced between 25-29 nucleotides upstream of the start site (Hain et al., 1993). Also critical for initiation is the presence of an A/T rich sequence, approximately 2 to 11 nucleotides upstream of the start site. All of the *S. solfataricus* P2 DNA polymerase sequences possessed sequences that matched the consensus promoter motifs, but the BoxA sequence of the P2 B3 polymerase was positioned well beyond the average of 25-29 nucleotides from the start site; it is unclear if this sequence is in fact the BoxA motif. Examples of other archaebacterial genes with divergent BoxA motifs are known (Zillig et al., 1996), and it is possible that the *S. solfataricus* P2 B3 DNA polymerase also utilizes a divergent promoter sequence.

The number of amino acid differences between the orthologous *S. solfataricus* P2 B3 and *S. shibatae* B3 sequences is surprising given the close evolutionary relationship of these two organisms. A possible explanation for the low amino acid similarity between these two sequences is positive selection for a novel function(s). There are very few examples of positive selection based on molecular sequences and only a single possible example in archaebacteria, that of the superoxide dismutase genes of halophiles (Joshi and Dennis, 1993). Evidence for positive selection can be assessed by taking the ratio of nonsynonymous ($K_A$) to synonymous ($K_S$) substitutions per site (Li, 1997). Ratios of >1 are considered strong evidence for positive selection, since the rate of nonsynonymous substitutions exceeds that which can be explained by neutral evolution. A ratio of <1 is taken as evidence for stabilizing or purifying selection, since deleterious nonsynonymous substitutions do not become fixed in the population. The $K_A/K_S$ ratio for the B3 DNA polymerase genes is only 0.192 (table 1.1), well below what is considered evidence for positive selection, but higher than that of the β-galactosidase gene (0.042). It is clear from alignments that the *S. solfataricus* P2 and *S. shibatae* B3 DNA polymerases have undergone a high rate of nonsynonymous amino acid replacements after their divergence from a common ancestral sequence (figure 1.2). Amino acid substitutions in catalytic domains suggest that we cannot be certain that the encoded proteins retain all or any of the ancestral exonuclease or polymerization functions, but the fact that both ORFs remain uninterrupted by nonsense mutations indicates that the genes encoding these proteins remain under some sort of selection.

If phylogenetic analyses are correct, the duplication events that gave rise to the *S. solfataricus* P2 B3 and B3 paralogs must have occurred early in the evolution of crenarchaeotes and before the divergence of extant members of the *Sulfolobales*. The presence of hybridizing signals in *Sulfolobales* genomic DNA using the *S. solfataricus* P2 B2 paralog as a probe strongly supports this notion. It is still unclear, however, if the B2 and B3 paralogs are actually functional DNA polymerases or have been selected for a different cellular function. The cloning of the *S. shibatae* B3 and *S. solfataricus* P2 B3 paralogs presents an excellent opportunity to study the function and evolution of these proteins.

## Chapter II. Multiple independent gene duplications in the evolution of eukaryotic and archaebacterial family B DNA polymerases.

### Introduction

Three DNA-dependent DNA polymerases, α, δ, and ε, have been identified by genetic and biochemical studies as essential for DNA replication in the budding yeast, *Saccharomyces cerevisiae* (Morrison et al. 1990; Budd and Campbell, 1993; Stillman, 1994). Sequencing of the genes corresponding to the catalytic subunits of these DNA polymerases revealed significant amino acid similarity of each of the proteins to one another. In addition, all three DNA polymerases shared significant amino acid similarity with other eukaryotic cellular-encoded DNA polymerases, plasmid- and viral-encoded polymerases, and with DNA polymerase II (*pol*B) of *Escherichia coli;* all of these polymerases are family B DNA polymerases (Braithwaite and Ito, 1993). An additional cellular-encoded DNA polymerase, Rev3, was identified in *S. cerevisiae* through a genetic screen for strains displaying reduced frequencies of UV mutagenesis (Morrison et al., 1989). Although this DNA polymerase shares sequence similarity with other family B homologs, it is divergent in amino acid sequence and function from the nuclear replicative polymerases.

The nuclear replicative DNA polymerases of eukaryotes, α, ε and δ, likely evolved by gene duplications as they exhibit a high level of amino acid similarity. Likewise, the Rev3 DNA polymerase is more similar at the amino acid level to eukaryotic family B homologs than to other archaebacterial, eukaryotic, or eubacterial cellular-encoded polymerases. All of these eukaryotic family B DNA polymerases must be the result of gene duplication events that occurred early in, or before, the evolution of extant eukaryotes. However, most available sequences of eukaryotic family B DNA polymerases are confined to representatives of animals and fungi and a handful of medically relevant protists (*Trypanosoma* and *Plasmodium*). By sequencing orthologs of the three nuclear replicative DNA polymerases from extant representatives of early diverging eukaryotic lineages, it may be possible to determine when in eukaryote evolution the gene duplication(s) occurred. Phylogenetic analyses of archaebacterial and eukaryotic polymerases might also resolve whether or not the gene duplications that gave rise to the multiple family B homologs of crenarchaeotes (as discussed in chapter I) and eukaryotes occurred independently of one another. With these questions in mind, I chose two organisms, *Giardia lamblia* and *Trichomonas vaginalis,* from which to try and amplify orthologs of the three replicative family B DNA polymerases.

## Results

### Homologs of the three nuclear replicative family B DNA polymerases of animals and fungi are found in early diverging eukaryotes.

Based on multiple alignments of amino acid sequences of the catalytic subunits of eukaryotic family B DNA polymerases, I designed degenerate PCR primers to amplify homologs from early diverging eukaryotes (figure 2.1). It is impossible to design a single primer set to amplify all three family B paralogs from one organism. It is possible, however, to design primer combinations which can do this. Using these primer combinations (figure 2.1), I was able to amplify orthologs of DNA polymerase δ and ε from the parabasalid *Trichomonas vaginalis,* and an ortholog of DNA polymerase α from the diplomonad, *Giardia lamblia.* PCR-products identified as putative family B DNA polymerases by BLAST and FASTA database searches were used as probes in Southern hybridizations against genomic DNAs from

Length of *S. cerevisiae* paralogs: Alpha - 1490 aa
Delta - 1113 aa
Epsilon - 2256 aa

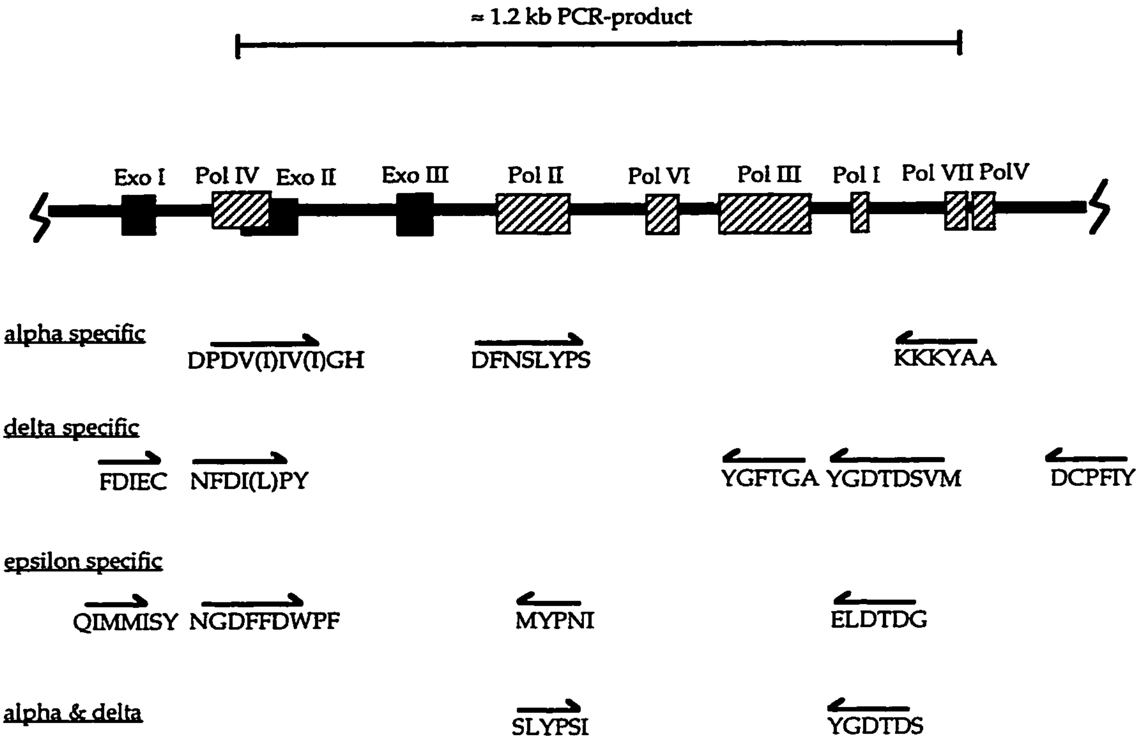


**Figure 2.1** Degenerate PCR primers used in attempts to amplify family B DNA polymerases from early-branching eukaryotes. A schematic drawing (not to scale) of the primary structure of the most conserved region of family B DNA polymerases is represented by a solid grey line. Conserved functional domains are indicated by solid (3′-5′ exonuclease) and hatched (polymerization) boxes in the order in which they appear N- to C-terminal. Domains are numbered according to the level of sequence conservation between different family B DNA polymerases (Wong et al., 1988). The amino acid sequences for primers specific to each ortholog are listed below functional domains to which they correspond. Direction of each primer is indicated by an arrow.

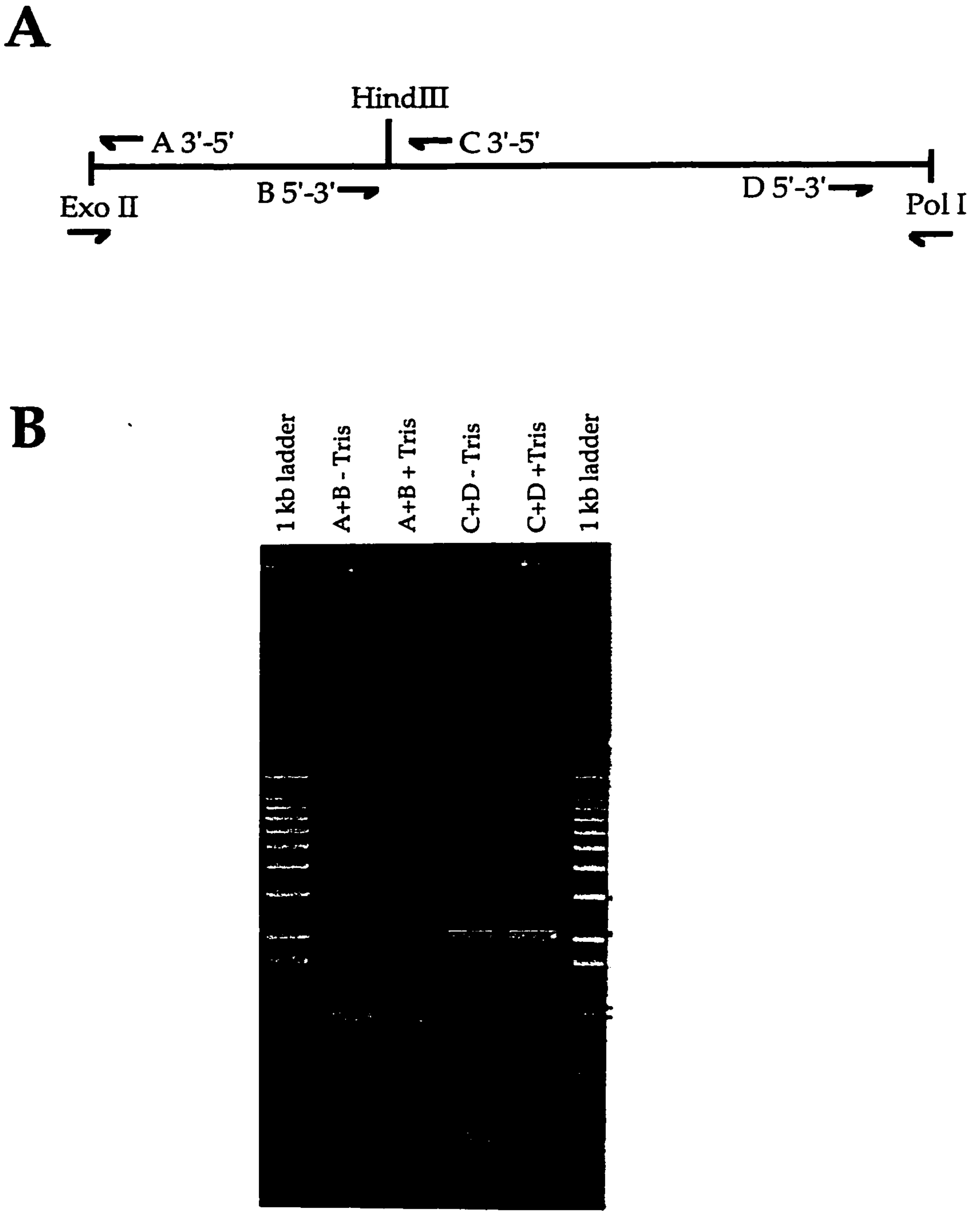
various protists. Each PCR-product hybridized to genomic DNA of the organism that initial PCR reactions were performed with, and each DNA polymerase appeared single copy (not shown). However, I was unable to amplify the three paralogs from a single organism. This result does not imply that early diverging eukaryotes do not possess all three paralogs as there could be a number of reasons why amplification was not successful (for instance, divergent target sequences or biased base composition of the genomic DNA).

It is also impossible to design PCR primers to amplify the entire coding region of eukaryotic family B DNA polymerases; they are too divergent in sequence and in *S. cerevisiae,* all paralogs are over 3 kb in coding sequence (the ε paralog is approximately 7 kb; Morrison et al., 1990). As well, the number of phylogenetically useful sites shared between eubacterial, archaebacterial, and eukaryotic homologs is less than 200 amino acids; between eukaryotic paralogs, the number of useful sites increases to around 300 amino acids. Additional coding sequence outside of the initial PCR products was obtained for DNA polymerase α of *G. lamblia* by screening a genomic DNA library in λgt11. One clone was picked for analysis after three rounds of screening using the initial PCR product as a probe. From the larger λ clone, two smaller *Bam*HI fragments of 2.1 and 1.4 kb were identified by Southern analysis (not shown) as containing sequences that hybridized to the initial PCR product. The two *Bam*HI fragments were cloned into pBluescript and sequenced as described (see materials and methods). These *Bam*HI subclones covered 3.6 kb of coding sequence, including the complete sequence of the initial PCR product. All phylogenetically useful sites were included in the 3.6 kb of sequence obtained.

Screening of both cDNA and genomic DNA libraries of *T. vaginalis* failed to recover clones carrying additional coding sequence of DNA

**Figure 2.2** (A) Strategy to obtain additional coding sequence of *T. vaginalis* DNA polymerase ε by IPCR. The initial PCR product (solid horizontal line), representing approximately 1.2 kb of coding sequence of DNA polymerase ε, was amplified from *T. vaginalis* genomic DNA with the degenerate oligonucleotides ExoIIe and PolIe (see appendix 1 and figure 2.1). The *Hind*III restriction site in the PCR product allowed two sets of IPCR primers to be designed, one set to amplify 5' sequence and the other to amplify 3' sequence (see appendix 1 for primer sequences).

(B) Results of IPCR run on 0.7% agarose gel with 1 kb ladder as a marker. Two different sized products were obtained with the two sets of primers: one approximately 0.9 kb and the other 2.1 kb. Lack of addition of higher pH Tris (pH 8.8) to reactions, as indicated by + and - symbols, did not seem to affect amplification.


A
HindIII
A 3'-5'
C 3'-5'
Exo II
B 5'-3'
D 5'-3'
Pol I
B
1 kb ladder
A+B - Tris
A+B + Tris
C+D - Tris
C+D +Tris
1 kb ladder

polymerase ε. Since the initial PCR product from *T. vaginalis* contained a single *Hind*III restriction site, two primer sets for use in inverse PCR (figure 2.2 a), each flanking the *Hind*III site, were designed to amplify additional coding sequence 5' and 3' to that of the PCR product. Inverse PCR reactions resulted in the amplification of 2.1 and 0.9 kb fragments (figure 2.2 b) and were cloned and sequenced as described. In all, 3.2 kb of coding sequence of DNA polymerase ε encompassing all of the phylogenetically useful sites from *T. vaginalis* was obtained.

Attempts to obtain additional coding sequence for DNA polymerase δ of *T. vaginalis* by inverse PCR, screening of both genomic and cDNA libraries, and by construction of a sized subgenomic library were unsuccessful even though the PCR product hybridized to *T. vaginalis* genomic DNA (figure 2.3).

The *T. vaginalis* and *G. lamblia* paralogs could be aligned with family B sequences from other eukaryotes, archaebacteria, and the single eubacterial homolog in conserved functional regions (figure 2.4). Although exonuclease domain III is conserved between orthologous sequences, it is difficult to align across all family B homologs and so was not included in phylogenetic analyses. Likewise, polymerase domain VI is not conserved in sequence and in archaebacterial homologs, multiple gaps must be introduced; this region was also excluded from analyses.

### Phylogenetic analyses suggest that multiple gene duplication events occurred early in eukaryote evolution

Phylogenetic analyses were performed on a dataset that included cellular-encoded family B DNA polymerases from eukaryotes, archaebacteria, and eubacteria. The *S. cerevisiae* Rev3 paralog was also included, as well as an EST sequence from *Mus musculus* that is likely an ortholog of the *S.*

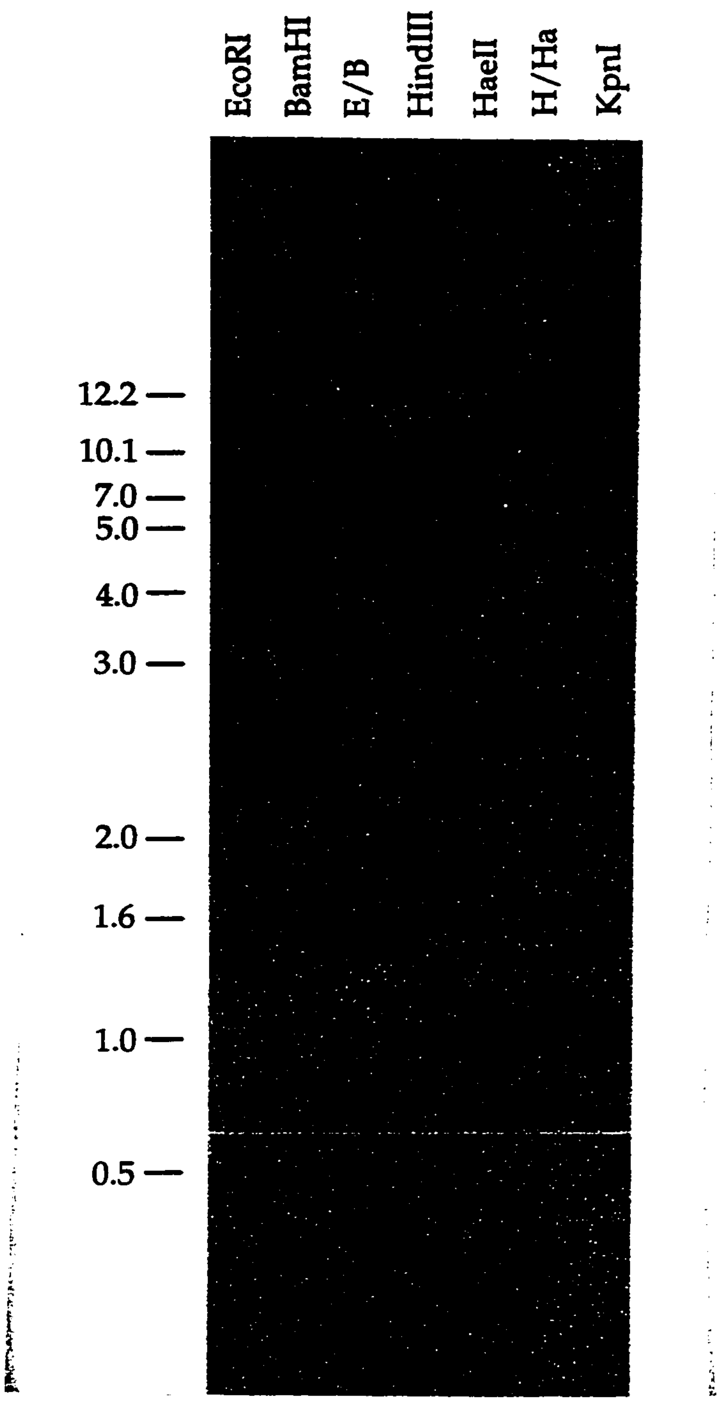
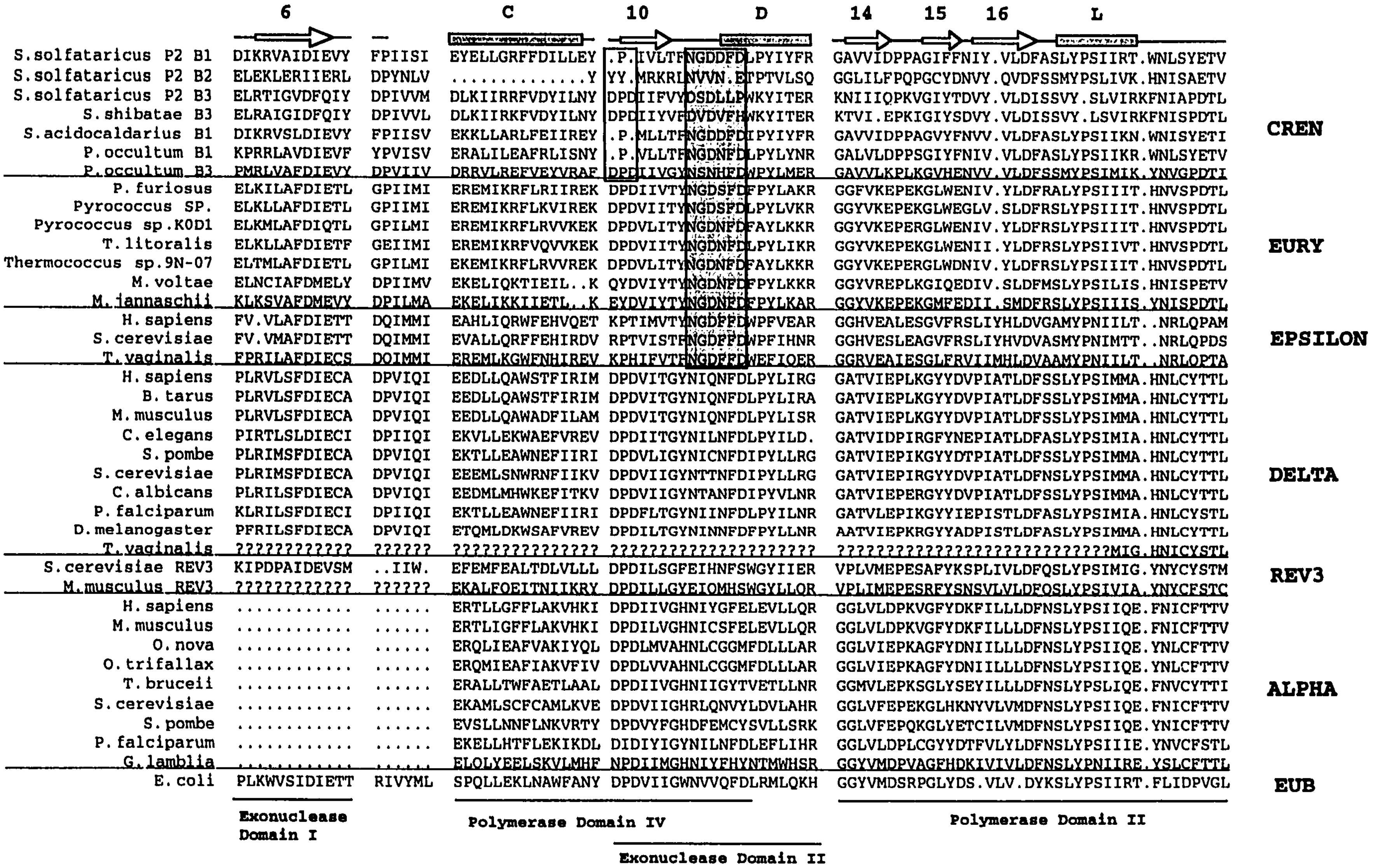


**Figure 2.3** Southern hybridization of the putative DNA polymerase δ PCR product to genomic DNA from *T. vaginalis* cut with various restriction enzymes and run on 0.7% agarose gel. Exposure shown is 15 hours. Ha=*Hae*III, H=*Hind*III.

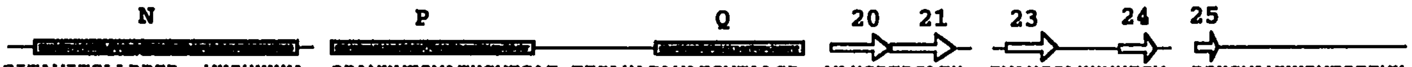
**Figure 2.4** Amino acid alignment of family B DNA polymerases of eubacteria, archaebacteria and eukaryotes. Numbering of conserved functional domains is as previously published (Wong et al., 1988). Secondary structure elements corresponding to the family B DNA polymerase of bacteriophage RB69 are indicated by an arrow (for β-sheets) or a hatched rectangle (for α-helices). Lines represent unstructured regions of the protein. Each structural element is assigned a letter or number corresponding to its position in the RB69 DNA polymerase amino acid sequence (Wang et al., 1997). Sheet 6, 109-117; 10, 211-216; 14, 395-399; 15, 403-405; 16, 407-412; 20, 626-621; 21, 622-626; 23, 700-703; 24, 707-710; 25, 726-728. Helix C, 194-208; D, 222-230; L, 417-424; N, 471-491; P, 547-571; Q, 581-597. The single amino acid deletions in exonuclease domain II of crenarchaeote polymerases that support a grouping of B1 orthologs are indicated by a box. Signature sequences that support a grouping of archaebacterial and eukaryotic ε polymerases are indicated by a stippled box. Gaps introduced in the alignment are indicated by a period (.). Missing data are indicated by a question mark (?).



```
                         6                            C                10        D         14       15    16          L

S.solfataricus P2 B1  DIKRVAIDIEVY  FPIISI  EYELLGRFFDILLEY .P.IVLTFNGDDFDLPYIYFR  GAVVIDPPAGIFFNIY.VLDFASLYPSIIRT.WNLSYETV
S.solfataricus P2 B2  ELEKLERIIERL  DPYNLV  ..............Y YY.MRKRLNVVN.ETPTVLSQ  GGLILFPQPGCYDNVY.QVDFSSMYPSLIVK.HNISAETV
S.solfataricus P2 B3  ELRTIGVDFQIY  DPIVVM  DLKIIRRFVDYILNY DPDIIFVYDSDLLPWKYITER  KNIIIQPKVGIYTDVY.VLDISSVY.SLVIRKFNIAPDTL
       S.shibatae B3  ELRAIGIDFQIY  DPIVVL  DLKIIRKFVDYILNY DPDIIYVFDVDVFHWKYITER  KTVI.EPKIGIYSDVY.VLDISSVY.LSVIRKFNISPDTL
  S.acidocaldarius B1 DIKRVSLDIEVY  FPIISV  EKKLLARLFEIIREY .P.MLLTFNGDDFDIPYIYFR  GAVVIDPPAGVYFNVV.VLDFASLYPSIIKN.WNISYETI   CREN
       P.occultum B1  KPRRLAVDIEVF  YPVISV  ERALILEAFRLISNY .P.VLLTFNGDNFDLPYLYNR  GALVLDPPSGIYFNIV.VLDFASLYPSIIKR.WNLSYETV
       P.occultum B3  PMRLVAFDIEVY  DPVIIV  DRRVLREFVEYVRAF DPDIIVGYNSNHFDWPYLMER  GAVVLKPLKGVHENVV.VLDFSSMYPSIMIK.YNVGPDTI
---------------------------------------------------------------------------------------------------------------------
         P.furiosus   ELKILAFDIETL  GPIIMI  EREMIKRFLRIIREK DPDIIVTYNGDSFDFPYLAKR  GGFVKEPEKGLWENIV.YLDFRALYPSIIIT.HNVSPDTL
       Pyrococcus SP. ELKLLAFDIETL  GPIIMI  EREMIKRFLKVIREK DPDVIITYNGDSFDLPYLVKR  GGYVKEPEKGLWEGLV.SLDFRSLYPSIIIT.HNVSPDTL
   Pyrococcus sp.KOD1 ELKMLAFDIQTL  GPILMI  EREMIKRFLRVVKEK DPDVLITYNGDNFDFAYLKKR  GGYVKEPERGLWENIV.YLDFRSLYPSIIIT.HNVSPDTL
         T.litoralis  ELKLLAFDIETF  GEIIMI  EREMIKRFVQVVKEK DPDVIITYNGDNFDLPYLIKR  GGYVKEPEKGLWENII.YLDFRSLYPSIIVT.HNVSPDTL   EURY
Thermococcus sp.9N-07 ELTMLAFDIETL  GPILMI  EKEMIKRFLRVVREK DPDVLITYNGDNFDFAYLKKR  GGYVKEPERGLWDNIV.YLDFRSLYPSIIIT.HNVSPDTL
            M.voltae  ELNCIAFDMELY  DPIIMV  EKELIQKTIEIL..K QYDVIYTYNGDNFDFPYLKKR  GGYVREPLKGIQEDIV.SLDFMSLYPSILIS.HNISPETV
         M.jannaschii KLKSVAFDMEVY  DPILMA  EKELIKKIIETL..K EYDVIYTYNGDNFDFPYLKAR  GGYVKEPEKGMFEDII.SMDFRSLYPSIIIS.YNISPDTL
---------------------------------------------------------------------------------------------------------------------
           H.sapiens  FV.VLAFDIETT  DQIMMI  EAHLIQRWFEHVQET KPTIMVTYNGDFFDWPFVEAR  GGHVEALESGVFRSLIYHLDVGAMYPNIILT..NRLQPAM
        S.cerevisiae  FV.VMAFDIETT  DQIMMI  EVALLQRFFEHIRDV RPTVISTFNGDFFDWPFIHNR  GGHVESLEAGVFRSLIYHVDVASMYPNIMTT..NRLQPDS   EPSILON
          T.vaginalis FPRILAFDIECS  DQIMMI  EREMLKGWFNHIREV KPHIFVTFNGDFFDWEFIQER  GGRVEAIESGLFRVIIMHLDVAAMYPNIILT..NRLQPTA
---------------------------------------------------------------------------------------------------------------------
           H.sapiens  PLRVLSFDIECA  DPVIQI  EEDLLQAWSTFIRIM DPDVITGYNIQNFDLPYLIRG  GATVIEPLKGYYDVPIATLDFSSLYPSIMMA.HNLCYTTL
             B.tarus  PLRVLSFDIECA  DPVIQI  EEDLLQAWSTFIRIM DPDVITGYNIQNFDLPYLIRA  GATVIEPLKGYYDVPIATLDFSSLYPSIMMA.HNLCYTTL
          M.musculus  PLRVLSFDIECA  DPVIQI  EEDLLQAWADFILAM DPDVITGYNIQNFDLPYLISR  GATVIEPLKGYYDVPIATLDFSSLYPSIMMA.HNLCYTTL
           C.elegans  PIRTLSLDIECI  DPIIQI  EKVLLEKWAEFVREV DPDIITGYNILNFDLPYILD.  GATVIDPIRGFYNEPIATLDFASLYPSIMIA.HNLCYTTL
             S.pombe  PLRIMSFDIECA  DPVIQI  EKTLLEAWNEFIIRI DPDVLIGYNICNFDIPYLLRG  GATVIEPIKGYYDTPIATLDFSSLYPSIMMA.HNLCYTTL
        S.cerevisiae  PLRIMSFDIECA  DPVIQI  EEEMLSNWRNFIIKV DPDVIIGYNTTNFDIPYLLRG  GATVIEPIRGYYDVPIATLDFNSLYPSIMMA.HNLCYTTL   DELTA
          C.albicans  PLRILSFDIECA  DPVIQI  EEDMLMHWKEFITKV DPDVIIGYNTANFDIPYVLNR  GATVIEPERGYYDVPIATLDFSSLYPSIMMA.HNLCYTTL
        P.falciparum  KLRILSFDIECI  DPIIQI  EKTLLEAWNEFIIRI DPDFLTGYNIINFDLPYILNR  GATVLEPIKGYYIEPISTLDFASLYPSIMIA.HNLCYSTL
      D.melanogaster  PFRILSFDIECA  DPVIQI  ETQMLDKWSAFVREV DPDILTGYNINNFDFPYLLNR  AATVIEPKRGYYADPISTLDFASLYPSIMMA.HNLCYTTL
          T.vaginalis ????????????  ??????  ??????????????? ?????????????????????  ????????????????????????????MIG.HNICYSTL
---------------------------------------------------------------------------------------------------------------------
   S.cerevisiae REV3 KIPDPAIDEVSM  ..IIW.  EFEMFEALTDLVLLL DPDILSGFEIHNFSWGYIIER  VPLVMEPESAFYKSPLIVLDFQSLYPSIMIG.YNYCYSTM   REV3
     M.musculus REV3  ????????????  ??????  EKALFQEITNIIKRY DPDILLGYEIQMHSWGYLLQR  VPLIMEPESRFYSNSVLVLDFQSLYPSIVIA.YNYCFSTC
---------------------------------------------------------------------------------------------------------------------
           H.sapiens  ............  ......  ERTLLGFFLAKVHKI DPDIIVGHNIYGFELEVLLQR  GGLVLDPKVGFYDKFILLLDFNSLYPSIIQE.FNICFTTV
          M.musculus  ............  ......  ERTLIGFFLAKVHKI DPDILVGHNICSFELEVLLQR  GGLVLDPKVGFYDKFILLLDFNSLYPSIIQE.FNICFTTV
              O.nova  ............  ......  ERQLIEAFVAKIYQL DPDLMVAHNLCGGMFDLLLAR  GGLVIEPKAGFYDNIILLLDFNSLYPSIIQE.YNLCFTTV
          O.trifallax ............  ......  ERQMIEAFIAKVFIV DPDLVVAHNLCGGMFDLLLAR  GGLVIEPKAGFYDNIILLLDFNSLYPSIIQE.YNLCFTTV
            T.bruceii ............  ......  ERALLTWFAETLAAL DPDIIVGHNIIGYTVETLLNR  GGMVLEPKSGLYSEYILLLDFNSLYPSLIQE.FNVCYTTI   ALPHA
        S.cerevisiae  ............  ......  EKAMLSCFCAMLKVE DPDVIIGHRLQNVYLDVLAHR  GGLVFEPEKGLHKNYVLVMDFNSLYPSIIQE.FNICFTTV
             S.pombe  ............  ......  EVSLLNNFLNKVRTY DPDVYFGHDFEMCYSVLLSRK  GGLVFEPQKGLYETCILVMDFNSLYPSIIQE.YNICFTTV
        P.falciparum  ............  ......  EKELLHTFLEKIKDL DIDIYIGYNILNFDLEFLIHR  GGLVLDPLCGYYDTFVLYLDFNSLYPSIIIE.YNVCFSTL
           G.lamblia  ............  ......  ELQLYEELSKVLMHF NPDIIMGHNIYFHYNTMWHSR  GGYVMDPVAGFHDKIVIVLDFNSLYPNIIRE.YSLCFTTL
---------------------------------------------------------------------------------------------------------------------
              E.coli  PLKWVSIDIETT  RIVYML  SPQLLEKLNAWFANY DPDVIIGWNVVQFDLRMLQKH  GGYVMDSRPGLYDS.VLV.DYKSLYPSIIRT.FLIDPVGL   EUB

                      Exonuclease             Polymerase Domain IV                          Polymerase Domain II
                      Domain I                                  Exonuclease Domain II
```



```
S.solfataricus P2 B1  GITAVITGLLRDFR..VKIYKKKA  QRAMKVFINATYGVFGAE.TFPLYAPAVAESVTALGR  VLYGDTDSLFK  FVAYSGLKKNYFGV  DIKGMLVKKRNTPEFVK
S.solfataricus P2 B2  GIIPEALQWLIERKSELK......  AEAIKWILVASFGYLGYR.NSLFGKIEAYEMVTYLAR  VLHGIIDSLVV  KET..GLRKRYNWI  IAKGL..IRENMPNIVK
S.solfataricus P2 B3  GLYKTFLDELSNVR..........  IKVIEELISSFNDYVHWV.NARWYSREIA.SAFDEFS  VILANDLLIFV  KSLLVLDNNRYAGL  DIARKGEEDMNLCELAR
       S.shibatae B3  GLYKTFLEELSNIQ..........  TFLEEELMSSFYDYIHWI.NSRWYSREIA.SAVDELS  VILANDFLVFV  RSLLILGNDRYAGL  DIARIGKEDRDLCELVR   CREN
  S.acidocaldarius B1  GITAVYQGLIRDFR..VKVYKKKA  QRAMKVFINATYGVFGAE.NFPLYAPAVAESVTAIGR  VIYGDTDSLFL  YVAYSGLKKNYFGT  EIKGMLAKKRNTPEFIK
       P.occultum B1  GLTSQIVGLLRDYR..VKIYKKKA  QAAMKVYINASYGVFGAE.SFPFYAPPVAESVTAIGR  VLYGDTDSLFI  FVTFSGLKKNYIGA  DVKGMVAKKRNTPEFLK
       P.occultum B3  GFFKTVLENLLKLRRQVKEKMKEF  QKALKVLANASYGYMGWS.HARWYCKRCAEAVTAWGR  VIYGDTDSLFV  KVFFTEAKKRYVGL  DIVGFEAVRGDWCELAK
-----------------------------------------------------------------------------------------------------------------------
          P.furiosus  GFIPSLLGHLLEERQKIKTKMKET  QKAIKLLANSFYGYYGYA.KARWYCKECAESVTAWGR  VLYIDTDGLYA  RGFFV.TKKRYAVI  ITRGLEIVRRDWSEIAK
       Pyrococcus sp.  GFIPSLLKRLLDERQEIKRKMKAS  QRAIKILANSYYGYYGYA.KARWYCKECAESVTAWGR  VLYIDTDGLYA  RGFFV.TKKKYALI  ITRGLEIVRRDWSEIAK
   Pyrococcus sp.KOD1  GFIPSLLGDLLEERQKIKKKMKAT  QRAIKILANSYYGYYGYA.RARWYCKECAESVTAWGR  VIYSDTDGFFA  RGFFV.TKKKYAVI  TTRGLEIVRRDWSEIAK   EURY
         T.litoralis  GFIPSILGDLIAMRQDIKKKMKST  QRAIKLLANSYYGYMGYP.KARWYSKECAESVTAWGR  VLYADTDGFYA  RGFFV.TKKRYAVI  TTRGLEVVRRDWSEIAK
 Thermoccus sp.9n-07  GFIPSLLGDLLEERQKIKRKMKAT  QRAIKILANSFYGYYGYA.KARWYCKECAESVTAWGR  VLYADTDGLHA  RGFFV.TKKKYAVI  TTRGLEIVRRDWSEIAK
            M.voltae  GIIPKTLNELLSRRKHIKMLLKDK  QKSIKVLANSHYGYLAFP.MARWYSDKCAEMVTGLGR  VIYADTDGFYA  RGLFV.TKKKYALI  VVKGLEVVRRDWSNIAK
         M.jannaschii  GLIPKTLRNLIERRINIKRRMKKM  QKSLKILANSVYGYLAFP.RARFYSRECAEIVTYLGR  VLYIDTDGFYA  RGIFV.TKKRYALV  TVKGLEFVRRDWSNIAK
-----------------------------------------------------------------------------------------------------------------------
           H.sapiens  GLHKVWKKKLSAAVEVGDAAEVKR  QLAHKCILNSFYGYVMRK.GARWYSMEMAGIVCFTGA  PLELDTDGIWC  VLP.KKLKKRYAVF  ELKGFEVKRRGELQLIK
        S.cerevisiae  GLAKTWKGNLSKIDPSDKHARDEA  QLAHKVILNSFYGYVMRK.GSRWYSMEMAGITCLTGA  PLELDTDGIWC  ILP.KGIKKRYAVF  ELKGFELKRRGELQLIK   EPSILON
         T.vaginalis  AKYYKQRLNNFCKKHYQKPKVEKE  QLAHKAILNSFYGYVMRD.VDRWRSMEMAGVVTNSGA  PLELDTDGIWC  ALP.KKLKKRYAVF  ELKGFESKR.GE.....
-----------------------------------------------------------------------------------------------------------------------
           H.sapiens  GLLPQILENLLSARKRAKAELAKE  QLALKVSANSVYGFTGAQ.VGKLPCLEISQSVTGFGR  VVYGDTDSVMC  FPYLLISKKRYAGL  DCKGLEAVRRDNCPLVA
             B.tarus  GLLPQILENLLSARKRAKAELAKE  QLALKVSANSVYGFTGAQ.VGKLPCLEISQSVTGFGR  VVYGDTDSVMC  FPYLLISKKRYMGL  DCKGLEAVRRDNCPLVA
          M.musculus  GLLPQILENLLSARKRAKAELAQE  QLALKVSANSVYGFTGAQ.VGKLPCLEISQSVTGFGR  VVYGDTDSVMC  FPYLLISKKRYAGL  DCKGLEAVRRDNCPLVA
           C.elegans  GLLPEILEDILAARKRAKNDMKNE  QLALKISANSVYGFTGAT.FGRKMIDMTKLEVERIYK  VIYGDTDSVMV  KFGLLINKKRYAGL  DCKGLETVRRDNCPLVA
             S.pombe  GLLPIILADLLNARKKAKADLKKE  QLALKVSANSVYGFTGAT.NGRLPCLAISSSVTSYGR  VIYGDTDSVMV  FPYLLISKKRYAGL  DSKGIETVRRDNCPLVS
        S.cerevisiae  GILPIILDELISARKRAKKDLRDE  QLALKISANSVYGFTGAT.VGKLPCLAISSSVTAYGR  VVYGDTDSVMV  FPYLLINKKRYAGL  DQKGLASVRRDSCPLVS   DELTA
          C.albicans  GILPTILDELLTARKKAKADLKKE  QLALKISANSVYGFTGAT.VGKLPCLAISSSVTAFGR  VIYGDTDSVMV  FPYLLINKKRYAGL  DTKGIETVRRDNCQLVQ
         P.falciparum  GILPLIVEELIEARKKVKLLIKNE  QLALKISANSVYGYTGASSGGQLPCLEVAVSITTLGR  VIYGDTDSVMV  CPYLLLNKKRYAGL  DCKGIETVRRDFCILIQ
       D.melanogaster  GILPLQLKRLVESRKEVKKLMAAP  QLALKISANSVYGFTGAQ.VGKLPCLEISGSVTAYGR  VIYGDTDSVMV  YPYLLINKKRYAGL  DCKGIETVRRDNSPLVA
         T.vaginalis  GVLPEILKELLAARKATKKLMEEA  QLALKISANSVYGFTGAT.VGKLPCLQISESVTAYGR  VI?????????  ??????????????  ?????????????????
-----------------------------------------------------------------------------------------------------------------------
   S.cerevisiae Rev3  STLSKMLTDILDVRVMIKKTMNEI  QLALKLLANVTYGYTSASFSGRMPCSDLADSIVQTGR  VVYGDTDSLFV  HPSILISKKRYVIF  DAKGIETVRRDGIPAQQ   REV3
     M.musculus Rev3  GVLPRMLEEILKTRLMVKQSMKSY  QLGLKLIANVTFGYTAANFSGRMPCIEVGDSIVHKAR  VVYGDTDSMFV  LPCVLQTKKRYVVF  DAKGIETVRRDSCPAVS
-----------------------------------------------------------------------------------------------------------------------
           H.sapiens  GILPREIRKLVERRKQVKQLMKQQ  QKALKLTANSMYGCLGFS.YSRFYAKPLAALVTYKGR  VIYGDTDSIMI  NTNLLLKKKKYAAL  ELKGLDIVRRDWCDLAK
          M.musculus  GILPREIRKLVERRKQVKQLMKQQ  QKALKLTANSMYGCLGFS.YSRFYAKPLAALVTYKGR  VIYGDTDSIMI  NTNLLLKKKKYAAL  ELKGLDIVRRDWCDLAK
              O.nova  AVLPMVLRDLVQKRKAVKDKMKTE  QKAIKLTANSMYGCLGFG.SSRFHAQAIAALITRTGR  VVYGDTDSIMI  NTGLLLKKKKYAAL  EMKGLDMVRRDWCPLSK
            O.fallax  AVLPNVLRDLVQKRKAVKEKMKNE  QKAIKLTANSMYGCLGFG.SSRFHAQAIAALITKTGR  VVYGDTDSIMI  NTGLLLKKKKYAAL  KYEGLDMVRRDWCPLSK
           T.bruceii  CILPKVIRGLVDSRREIKRMMKSE  QLALKLTANSMYGCLGFE.YSRFYAQPLAELVTRQGR  VIYGDTDSVMI  QTGLLLRKKKYAAL  EVKGLDMVRRDWCPLSQ   ALPHA
        S.cerevisiae  GVLPRLLANLVDRRREVKKVMKTE  QQALKLTANSMYGCLGYT.KSRFYARPLAVLITYKGR  VIYGDTDSVML  NTNLLHAKKKYAAL  DVKGLDMKRREFCTLAK
             S.pombe  GIFPRLIANLVERRRQIKGLLKDN  QQALKLTANSMYGCLGYT.KSRFYARPLAVLITYKGR  VIYGDTDSVML  NTNLLHAKKKYAAL  DVKGLDMKRREFCTLAK
         P.falciparum  GILPCILKSLVEKRSVIKKLISNE  SLSIKLISNSIYGCLGNT.NNRFYAKHIASYITSKGR  VIYGDTDSIMI  DTGLLLKKKKYACA  EMKGINFIKRDFSKISK
           G.lamblia  VILPKIIARLITHETGHKEKNQGP  QLAIKLCANAMYGSLGYQ.YGRFYCPHVAARIPARGR  VIYGDTDSIMI  KTNLLLQKKKYPAT  EYKGLDLVRRDWCVLFP
-----------------------------------------------------------------------------------------------------------------------
              E.coli  GFLDAWFS...REKHCL.PEIVTN  SQALKIIMNAFYGVLGTT.ACRFFDPRLASSITMRGH  VIYGDTDSTFV  RGADTGSKKRYAGL  VFKGLETVRTDWTPLAQ   EUB

                      ________________________  _____________________________________  ___________  ______________  _________________
                      Polymerase Domain VI      Polymerase Domain III                  Polymerase   Polymerase      Polymerase
                                                                                       Domain I     Domain VII      Domain V
```

*cerevisiae* Rev3 polymerase. Most eukaryotic sequences (16 of 24 included in this analysis) are from animals and fungi; in some instances, a single sequence was used in phylogenetic analyses where an ortholog had been sequenced from two closely related animals or fungi (ie. from rat and hamster). Likewise, some archaebacterial sequences are from closely related species (ie. *Pyrococcus furiosus* and *Pyrococcus abysii*). In these instances, only a single sequence was used. The final dataset consisted of 38 taxa and 162 amino acid positions.

The choice of outgroup sequence for eukaryotic and archaebacterial family B polymerases is problematic because there is only a single eubacterial sequence available, that of *E. coli.* Single outgroup sequences with long branch lengths relative to ingroup sequences can be extremely problematic for parsimony analyses, but less so for distance analyses (Swofford et al., 1996). Preliminary PROTDIST analyses with the *E. coli* sequence resulted in tree topologies that placed *E. coli* as a sister taxon to eukaryotic δ paralogs. Bootstrap support for this branching order was under 50% as measured by both parsimony and distance methods (not shown). In addition, the *E. coli* sequence is not specifically closer to one particular eukaryotic paralog than it is to any archaebacterial paralog based on amino acid identity (table 2.1). The branching of *E. coli* with eukaryotic δ paralogs is most likely artefactual and I did not include the *E. coli* sequence in any further analyses.

In both parsimony and distance analyses, the three eukaryotic replicative family B DNA polymerases, α, δ and ε, formed monophyletic groups with high bootstrap values (figure 2.5). The three DNA polymerases amplified from representatives of early diverging protist lineages all grouped with orthologous sequences from other eukaryotes. This result is to be expected as orthologous sequences from different organisms should show a

**Table 2.2** Percent identities of eubacterial, archaebacterial, and eukaryotic family B DNA polymerases. Eukaryotic paralogs are divided into groups and separated by lines, as are crenarchaeote and euryarchaeote polymerases. Number of amino acid differences between two sequences is below the diagonal break. Percent identity between two sequences is above the diagonal break.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 S. solf P2 B1 | - | 32.0 | 23.5 | 26.3 | 85.2 | 75.5 | 46.5 | 44.8 | 44.8 | 42.9 | 45.5 | 43.5 | 41.6 | 40.9 | 34.2 | 34.9 | 35.4 | 41.9 | 41.9 | 42.6 | 41.6 | 41.3 | 41.3 |
| 2 S. solf P2 B2 | 102 | - | 19.5 | 18.9 | 32.7 | 36.0 | 31.8 | 33.3 | 32.7 | 30.7 | 32.7 | 33.3 | 33.3 | 32.7 | 22.8 | 22.1 | 22.2 | 29.1 | 29.1 | 29.8 | 27.3 | 29.1 | 28.5 |
| 3 S. solf P2 B3 | 117 | 120 | - | 74.8 | 24.2 | 20.9 | 36.8 | 30.5 | 28.6 | 29.9 | 29.2 | 29.2 | 31.2 | 28.6 | 20.4 | 19.7 | 20.4 | 25.8 | 25.2 | 26.5 | 27.9 | 26.5 | 24.5 |
| 4 S. shib B3 | 112 | 120 | 39 | - | 25.7 | 21.7 | 35.1 | 30.7 | 29.4 | 30.7 | 30.1 | 30.1 | 32.0 | 30.1 | 19.9 | 21.9 | 21.2 | 28.6 | 27.9 | 29.2 | 28.8 | 29.9 | 27.3 |
| 5 S. acid B1 | 23 | 101 | 116 | 113 | - | 74.2 | 49.0 | 41.6 | 42.2 | 42.2 | 40.9 | 41.6 | 40.3 | 37.7 | 31.6 | 33.6 | 31.3 | 40.0 | 40.0 | 40.6 | 40.9 | 41.9 | 40.0 |
| 6 P. occultum B1 | 38 | 96 | 121 | 119 | 40 | - | 48.4 | 44.2 | 44.8 | 44.2 | 45.5 | 45.5 | 42.9 | 40.9 | 31.6 | 32.2 | 34.0 | 43.9 | 43.9 | 44.5 | 39.6 | 42.6 | 41.9 |
| 7 P. occultum B3 | 83 | 103 | 98 | 100 | 79 | 80 | - | 57.7 | 55.8 | 53.2 | 54.5 | 52.6 | 50.6 | 49.4 | 35.7 | 34.4 | 34.2 | 51.0 | 51.0 | 51.6 | 46.8 | 50.3 | 47.8 |
| 8 P. furiosus | 85 | 100 | 107 | 106 | 90 | 86 | 66 | - | 90.4 | 85.9 | 86.5 | 87.8 | 64.7 | 66.7 | 46.4 | 41.2 | 45.9 | 46.2 | 45.5 | 46.8 | 41.3 | 44.9 | 44.9 |
| 9 Pyrococcus sp. | 85 | 101 | 110 | 108 | 89 | 85 | 59 | 15 | - | 87.8 | 87.8 | 88.5 | 66.7 | 67.9 | 43.1 | 41.2 | 40.5 | 47.4 | 46.8 | 48.1 | 41.9 | 45.5 | 46.2 |
| 10 Pyro sp. KOD1 | 88 | 104 | 108 | 106 | 89 | 86 | 73 | 22 | 19 | - | 85.9 | 94.2 | 65.4 | 67.3 | 41.2 | 38.6 | 40.5 | 44.2 | 43.6 | 44.9 | 40.6 | 44.9 | 44.9 |
| 11 T. litoralis | 84 | 101 | 109 | 107 | 91 | 84 | 71 | 21 | 19 | 22 | - | 85.9 | 67.3 | 69.2 | 45.8 | 43.1 | 43.9 | 48.7 | 48.1 | 49.4 | 41.9 | 46.2 | 46.2 |
| 12 Thermo sp. 9n-07 | 87 | 100 | 109 | 107 | 90 | 84 | 74 | 19 | 18 | 9 | 22 | - | 64.7 | 66.0 | 43.1 | 40.5 | 42.6 | 45.5 | 44.9 | 46.2 | 40.6 | 45.5 | 46.2 |
| 13 M. voltae | 90 | 100 | 106 | 104 | 92 | 88 | 77 | 55 | 52 | 54 | 51 | 55 | - | 75.6 | 36.6 | 35.9 | 34.5 | 43.6 | 42.9 | 44.2 | 40.0 | 41.7 | 39.7 |
| 14 M. jannaschii | 91 | 101 | 110 | 107 | 96 | 91 | 79 | 52 | 50 | 51 | 48 | 53 | 38 | - | 41.2 | 39.9 | 38.5 | 43.6 | 42.9 | 44.2 | 39.4 | 42.3 | 42.3 |
| 15 H. sapiens E | 100 | 115 | 121 | 121 | 104 | 104 | 99 | 82 | 87 | 90 | 83 | 87 | 97 | 90 | - | 83.2 | 79.2 | 32.9 | 32.3 | 33.5 | 31.2 | 31.0 | 29.7 |
| 16 S. cerevisiae E | 99 | 116 | 122 | 118 | 101 | 103 | 101 | 90 | 90 | 94 | 87 | 91 | 98 | 92 | 26 | - | 72.5 | 33.5 | 32.9 | 34.2 | 31.2 | 32.3 | 31.0 |
| 17 T. vaginalis E | 95 | 112 | 117 | 115 | 101 | 97 | 98 | 80 | 88 | 88 | 83 | 85 | 97 | 91 | 31 | 41 | - | 34.0 | 33.3 | 34.7 | 32.2 | 32.0 | 32.7 |
| 18 H. sapiens D | 90 | 107 | 115 | 110 | 93 | 87 | 77 | 84 | 82 | 87 | 80 | 85 | 88 | 88 | 104 | 103 | 99 | - | 98.7 | 98.7 | 71.5 | 84.9 | 84.3 |
| 19 B. tarus D | 90 | 107 | 116 | 111 | 93 | 87 | 77 | 85 | 83 | 88 | 81 | 86 | 89 | 89 | 105 | 104 | 100 | 2 | - | 98.1 | 70.9 | 83.6 | 83.0 |
| 20 M. musculus D | 89 | 106 | 114 | 109 | 92 | 86 | 76 | 83 | 81 | 86 | 79 | 84 | 87 | 87 | 103 | 102 | 98 | 2 | 3 | - | 71.5 | 83.6 | 83.0 |
| 21 C. elegans D | 90 | 109 | 111 | 109 | 91 | 93 | 83 | 91 | 90 | 92 | 90 | 92 | 93 | 94 | 106 | 106 | 101 | 45 | 46 | 45 | - | 71.5 | 70.3 |
| 22 S. pombe D | 91 | 107 | 114 | 108 | 90 | 89 | 78 | 86 | 85 | 86 | 84 | 85 | 91 | 90 | 107 | 105 | 102 | 24 | 26 | 26 | 45 | - | 88.7 |
| 23 S. cerevisiae D | 91. | 108 | 117 | 112 | 93 | 90 | 82 | 86 | 84 | 86 | 84 | 84 | 94 | 90 | 109 | 107 | 101 | 25 | 27 | 27 | 47 | 18 | - |
| 24 C. albicans D | 89 | 104 | 114 | 108 | 90 | 88 | 79 | 83 | 81 | 82 | 81 | 81 | 91 | 89 | 103 | 104 | 98 | 25 | 26 | 25 | 42 | 19 | 17 |
| 25 P. falciparum D | 89 | 107 | 111 | 105 | 89 | 86 | 79 | 82 | 81 | 83 | 82 | 83 | 90 | 85 | 101 | 99 | 96 | 43 | 44 | 42 | 46 | 39 | 43 |
| 26 D. melanogaster D | 91 | 110 | 115 | 109 | 91 | 88 | 81 | 84 | 86 | 84 | 86 | 84 | 92 | 90 | 105 | 106 | 98 | 30 | 31 | 29 | 41 | 29 | 31 |
| 27 T. vaginalis D | 29 | 36 | 38 | 37 | 27 | 30 | 27 | 26 | 25 | 24 | 28 | 25 | 30 | 28 | 37 | 37 | 35 | 11 | 11 | 11 | 19 | 10 | 7 |
| 28 S. cer Rev3 | 96 | 118 | 119 | 110 | 98 | 95 | 88 | 95 | 97 | 96 | 96 | 96 | 102 | 99 | 108 | 107 | 104 | 83 | 83 | 82 | 84 | 81 | 87 |
| 29 M. mus Rev3 | 91 | 95 | 103 | 97 | 91 | 87 | 80 | 83 | 87 | 83 | 86 | 84 | 91 | 88 | 100 | 103 | 96 | 73 | 73 | 72 | 75 | 69 | 73 |
| 30 H. sapiens A | 83 | 87 | 99 | 98 | 82 | 77 | 71 | 72 | 73 | 74 | 73 | 74 | 74 | 73 | 90 | 92 | 86 | 70 | 71 | 69 | 68 | 70 | 70 |
| 31 M. musculus A | 83 | 87 | 100 | 99 | 81 | 76 | 72 | 72 | 73 | 73 | 74 | 73 | 75 | 74 | 90 | 93 | 86 | 71 | 72 | 70 | 69 | 68 | 71 |
| 32 O. nova A | 83 | 90 | 104 | 103 | 84 | 80 | 81 | 75 | 79 | 76 | 76 | 75 | 77 | 76 | 92 | 94 | 90 | 71 | 72 | 70 | 73 | 69 | 68 |
| 33 O. fallax A | 83 | 91 | 104 | 103 | 86 | 81 | 81 | 75 | 79 | 76 | 76 | 75 | 78 | 77 | 95 | 96 | 92 | 72 | 73 | 71 | 75 | 70 | 69 |
| 34 T. bruceii A | 86 | 91 | 103 | 97 | 83 | 78 | 75 | 75 | 76 | 76 | 75 | 77 | 78 | 77 | 93 | 95 | 87 | 69 | 70 | 68 | 66 | 68 | 68 |
| 35 S. cerevisiae A | 88 | 91 | 104 | 101 | 86 | 81 | 80 | 80 | 78 | 79 | 77 | 80 | 80 | 75 | 99 | 97 | 95 | 74 | 75 | 75 | 78 | 71 | 73 |
| 36 S. pombe A | 88 | 91 | 104 | 101 | 86 | 81 | 80 | 80 | 78 | 79 | 77 | 80 | 80 | 75 | 99 | 97 | 95 | 74 | 75 | 75 | 78 | 71 | 73 |
| 37 P. falciparum A | 86 | 91 | 104 | 102 | 83 | 78 | 81 | 76 | 75 | 73 | 74 | 73 | 77 | 78 | 94 | 97 | 90 | 76 | 77 | 75 | 76 | 70 | 76 |
| 38 G. lamblia A | 83 | 96 | 108 | 107 | 83 | 83 | 79 | 80 | 80 | 79 | 81 | 79 | 84 | 85 | 96 | 94 | 92 | 78 | 78 | 76 | 76 | 79 | 78 |
| 39 E. coli | 98 | 108 | 121 | 114 | 97 | 98 | 90 | 89 | 85 | 86 | 86 | 86 | 95 | 92 | 108 | 105 | 104 | 92 | 93 | 92 | 91 | 87 | 90 |

| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 34.9 | 35.4 | 41.9 | 41.9 | 42.6 | 41.6 | 41.3 | 41.3 | 41.8 | 42.6 | 41.3 | 40.8 | 36.8 | 32.6 | 38.5 | 38.5 | 38.5 | 38.5 | 36.3 | 34.8 | 34.8 | 36.3 | 38.5 | 35.9 | |
| 22.1 | 22.2 | 29.1 | 29.1 | 29.8 | 27.3 | 29.1 | 28.5 | 30.2 | 29.1 | 27.2 | 26.5 | 20.3 | 27.5 | 33.6 | 33.6 | 31.3 | 30.5 | 30.5 | 30.5 | 30.5 | 30.5 | 26.7 | 27.5 | |
| 19.7 | 20.4 | 25.8 | 25.2 | 26.5 | 27.9 | 26.5 | 24.5 | 25.5 | 28.4 | 25.8 | 20.8 | 21.7 | 23.7 | 26.7 | 25.9 | 23.0 | 23.0 | 23.7 | 23.0 | 23.0 | 23.0 | 20.0 | 20.9 | |
| 21.9 | 21.2 | 28.6 | 27.9 | 29.2 | 28.8 | 29.9 | 27.3 | 28.9 | 31.8 | 29.2 | 22.9 | 27.2 | 27.6 | 26.9 | 26.1 | 23.1 | 23.1 | 27.6 | 24.6 | 24.6 | 23.9 | 20.1 | 25.0 | |
| 33.6 | 31.3 | 40.0 | 40.0 | 40.6 | 40.9 | 41.9 | 40.0 | 41.2 | 42.6 | 41.3 | 44.9 | 35.5 | 32.6 | 39.3 | 40.0 | 37.8 | 36.3 | 38.5 | 36.3 | 36.3 | 38.5 | 38.5 | 36.6 | CREN |
| 32.2 | 34.0 | 43.9 | 43.9 | 44.5 | 39.6 | 42.6 | 41.9 | 42.5 | 44.5 | 43.2 | 38.8 | 37.5 | 35.6 | 43.0 | 43.7 | 40.7 | 40.0 | 42.2 | 40.0 | 40.0 | 42.2 | 38.5 | 35.9 | |
| 34.4 | 34.2 | 51.0 | 51.0 | 51.6 | 46.8 | 50.3 | 47.8 | 49.0 | 49.7 | 48.4 | 44.9 | 42.9 | 41.6 | 48.2 | 47.4 | 40.9 | 40.9 | 45.3 | 41.6 | 41.6 | 40.9 | 42.3 | 41.9 | |
| 41.2 | 45.9 | 46.2 | 45.5 | 46.8 | 41.3 | 44.9 | 44.9 | 46.1 | 47.4 | 46.2 | 46.9 | 37.9 | 39.0 | 47.1 | 47.1 | 44.9 | 44.9 | 44.9 | 41.2 | 41.2 | 44.1 | 41.2 | 42.2 | |
| 41.2 | 40.5 | 47.4 | 46.8 | 48.1 | 41.9 | 45.5 | 46.2 | 47.4 | 48.1 | 44.9 | 49.0 | 36.6 | 36.0 | 46.3 | 46.3 | 41.9 | 41.9 | 44.1 | 42.6 | 42.6 | 44.9 | 41.2 | 44.8 | |
| 38.6 | 40.5 | 44.2 | 43.6 | 44.9 | 40.6 | 44.9 | 44.9 | 46.8 | 46.8 | 46.2 | 51.0 | 37.3 | 39.0 | 45.6 | 46.3 | 44.1 | 44.1 | 44.1 | 41.9 | 41.9 | 46.3 | 41.9 | 44.2 | |
| 43.1 | 43.9 | 48.7 | 48.1 | 49.4 | 41.9 | 46.2 | 46.2 | 47.4 | 47.4 | 44.9 | 42.9 | 37.3 | 36.8 | 46.3 | 45.6 | 44.1 | 44.1 | 44.9 | 43.4 | 43.4 | 45.6 | 40.4 | 44.2 | EURY |
| 40.5 | 42.6 | 45.5 | 44.9 | 46.2 | 40.6 | 45.5 | 46.2 | 47.4 | 46.8 | 46.2 | 49.0 | 37.3 | 38.2 | 45.6 | 46.3 | 44.9 | 44.9 | 43.4 | 41.2 | 41.2 | 46.3 | 41.9 | 44.2 | |
| 35.9 | 34.5 | 43.6 | 42.9 | 44.2 | 40.0 | 41.7 | 39.7 | 40.9 | 42.3 | 41.0 | 38.8 | 33.3 | 33.1 | 45.6 | 44.9 | 43.4 | 42.6 | 42.6 | 41.2 | 41.2 | 43.4 | 38.2 | 38.3 | |
| 39.9 | 38.5 | 43.6 | 42.9 | 44.2 | 39.4 | 42.3 | 42.3 | 42.2 | 45.5 | 42.3 | 42.9 | 35.3 | 35.3 | 46.3 | 45.6 | 44.1 | 43.4 | 43.4 | 44.9 | 44.9 | 42.6 | 37.5 | 40.3 | |
| 33.2 | 79.2 | 32.9 | 32.3 | 33.5 | 31.2 | 31.0 | 29.7 | 32.7 | 34.8 | 32.3 | 22.9 | 28.9 | 26.5 | 33.8 | 33.8 | 32.4 | 30.1 | 31.6 | 27.2 | 27.2 | 30.9 | 29.4 | 29.4 | |
| - | 72.5 | 33.5 | 32.9 | 34.2 | 31.2 | 32.3 | 31.0 | 32.0 | 36.1 | 31.6 | 22.9 | 29.6 | 24.3 | 32.4 | 31.6 | 30.9 | 29.4 | 30.1 | 28.7 | 28.7 | 28.7 | 30.9 | 31.4 | EPSILON |
| 41 | - | 34.0 | 33.3 | 34.7 | 32.2 | 32.0 | 32.7 | 33.8 | 36.0 | 34.7 | 27.1 | 29.3 | 26.2 | 33.8 | 33.8 | 30.8 | 29.2 | 33.1 | 26.9 | 26.9 | 30.8 | 29.2 | 29.7 | |
| 103 | 99 | - | 98.7 | 98.7 | 71.5 | 84.9 | 84.3 | 84.1 | 73.0 | 81.1 | 77.6 | 46.8 | 47.1 | 49.3 | 48.6 | 48.6 | 47.8 | 50.0 | 46.4 | 46.4 | 44.9 | 43.5 | 41.4 | |
| 104 | 100 | 2 | - | 98.1 | 70.9 | 83.6 | 83.0 | 83.4 | 72.3 | 80.5 | 77.6 | 46.8 | 47.1 | 48.6 | 47.8 | 47.8 | 47.1 | 49.3 | 45.7 | 45.7 | 44.2 | 43.5 | 40.8 | |
| 102 | 98 | 2 | 3 | - | 71.5 | 83.6 | 83.0 | 84.1 | 73.6 | 81.8 | 77.6 | 47.4 | 47.8 | 50.0 | 49.3 | 49.3 | 48.6 | 50.7 | 45.7 | 45.7 | 45.7 | 44.9 | 41.4 | |
| 106 | 101 | 45 | 46 | 45 | - | 71.5 | 70.3 | 73.2 | 70.9 | 74.1 | 61.2 | 45.8 | 45.3 | 50.4 | 49.6 | 46.7 | 45.3 | 51.8 | 43.1 | 43.1 | 44.5 | 44.5 | 41.7 | |
| 105 | 102 | 24 | 26 | 26 | 45 | - | 88.7 | 87.9 | 75.5 | 81.8 | 79.6 | 48.1 | 50.0 | 49.3 | 50.7 | 50.0 | 49.3 | 50.7 | 48.6 | 48.6 | 49.3 | 42.8 | 44.6 | DELTA |
| 107 | 101 | 25 | 27 | 27 | 47 | 18 | - | 89.2 | 73.0 | 80.5 | 85.7 | 44.2 | 47.1 | 49.3 | 48.6 | 50.7 | 50.0 | 50.7 | 47.1 | 47.1 | 44.9 | 43.5 | 42.7 | |
| 104 | 98 | 25 | 26 | 25 | 42 | 19 | 17 | - | 75.2 | 85.4 | 85.7 | 46.8 | 47.1 | 48.5 | 47.8 | 47.8 | 47.1 | 50.0 | 45.6 | 45.6 | 45.6 | 43.4 | 45.2 | |
| 99 | 96 | 43 | 44 | 42 | 46 | 39 | 43 | 39 | - | 77.4 | 73.5 | 52.2 | 51.1 | 50.0 | 50.7 | 46.4 | 45.7 | 52.9 | 46.4 | 46.4 | 48.6 | 44.2 | 40.1 | |
| 106 | 98 | 30 | 31 | 29 | 41 | 29 | 31 | 23 | 36 | - | 85.7 | 48.1 | 50.0 | 48.6 | 49.3 | 47.8 | 47.1 | 51.4 | 42.8 | 42.8 | 44.9 | 42.8 | 40.1 | |
| 37 | 35 | 11 | 11 | 11 | 19 | 10 | 7 | 7 | 13 | 7 | - | 53.1 | 40.8 | 40.8 | 40.8 | 34.7 | 34.7 | 44.9 | 40.8 | 40.8 | 40.8 | 38.8 | 32.7 | |
| 107 | 104 | 83 | 83 | 82 | 84 | 81 | 87 | 82 | 75 | 81 | 23 | - | 69.8 | 41.3 | 42.0 | 41.3 | 40.6 | 42.0 | 38.4 | 38.4 | 39.9 | 40.6 | 35.1 | REV3 |
| 103 | 96 | 73 | 73 | 72 | 75 | 69 | 73 | 72 | 68 | 69 | 29 | 42 | - | 40.6 | 41.3 | 41.3 | 40.6 | 39.9 | 38.4 | 38.4 | 39.9 | 39.9 | 33.8 | |
| 92 | 86 | 70 | 71 | 69 | 68 | 70 | 70 | 70 | 69 | 71 | 29 | 81 | 82 | - | 97.8 | 75.4 | 73.9 | 76.8 | 68.1 | 68.1 | 61.6 | 62.3 | 46.3 | |
| 93 | 86 | 71 | 72 | 70 | 69 | 68 | 71 | 71 | 68 | 70 | 29 | 80 | 81 | 3 | - | 75.4 | 73.9 | 75.4 | 68.1 | 68.1 | 61.6 | 60.9 | 45.6 | |
| 94 | 90 | 71 | 72 | 70 | 73 | 69 | 68 | 71 | 74 | 72 | 32 | 81 | 81 | 34 | 34 | - | 96.4 | 71.7 | 63.8 | 63.8 | 57.2 | 58.7 | 39.7 | |
| 96 | 92 | 72 | 73 | 71 | 75 | 70 | 69 | 72 | 75 | 73 | 32 | 82 | 82 | 36 | 36 | 5 | - | 69.6 | 63.0 | 63.0 | 55.1 | 58.0 | 39.0 | ALPHA |
| 95 | 87 | 69 | 70 | 68 | 66 | 68 | 68 | 68 | 65 | 67 | 27 | 80 | 83 | 32 | 34 | 39 | 42 | - | 63.8 | 63.8 | 55.1 | 55.8 | 43.4 | |
| 97 | 95 | 74 | 75 | 75 | 78 | 71 | 73 | 74 | 74 | 79 | 29 | 85 | 85 | 44 | 44 | 50 | 51 | 50 | - | 100.0 | 52.2 | 51.4 | 43.4 | |
| 97 | 95 | 74 | 75 | 75 | 78 | 71 | 73 | 74 | 74 | 79 | 29 | 85 | 85 | 44 | 44 | 50 | 51 | 50 | 0 | - | 52.2 | 51.4 | 43.4 | |
| 97 | 90 | 76 | 77 | 75 | 76 | 70 | 76 | 74 | 71 | 76 | 29 | 83 | 83 | 53 | 53 | 59 | 62 | 62 | 66 | 66 | - | 51.4 | 42.6 | |
| 94 | 92 | 78 | 78 | 76 | 76 | 79 | 78 | 77 | 77 | 79 | 30 | 82 | 83 | 52 | 54 | 57 | 58 | 61 | 67 | 67 | 67 | - | 43.4 | |
| 105 | 104 | 92 | 93 | 92 | 91 | 87 | 90 | 85 | 94 | 94 | 33 | 100 | 90 | 73 | 74 | 82 | 83 | 77 | 77 | 77 | 78 | 77 | - | EUB |

| | Bootstrap support for | | | | |
|---|---|---|---|---|---|
| Taxa deleted | α/δ/ε/rev3 | all crens | eury/ε | δ/α/rev3 | cren/ε |
| none | 12/5 | 0/0 | 29/36 | 69/56 | 1/0 |
| S. solf P2 B2 | 8/7 | 2/0 | 40/42 | 83/69 | 1/2 |
| S. solf P2 B3<br>S. shib B3 | 11/13 | 4/0 | 33/48 | 68/78 | 0/0 |
| S. solf P2 B2<br>S. solf P2 B3<br>S. shib B3 | 8/5 | 8/9 | 45/70 | 58/89 | 2/3 |

**Table 2.2** Effect of removing rapidly evolving taxa from phylogenetic analyses as measured by bootstrap. Confidence for each particular branching topology was measured by 100 bootstrap replicates by both parsimony and distance methods. Values are indicated in the order parsimony/distance. "All crens" refers to monophyly of all crenarchaeote paralogs.

closer phylogenetic relationship than paralogous sequences from the same organism. Both methods also supported a closer relationship of the α and δ paralogs while neither method supported a α/δ/ε grouping (table 2.2). Within orthologous groups, the most significant difference between parsimony and distance analyses was the placement of the partial *T. vaginalis* δ sequence; distance analyses placed this sequence as a sister to *S. cerevisiae,* whereas parsimony analyses placed this sequence among the other protist sequences, but bootstrap support for either placement was weak (figure 2.5). Both parsimony and distance methods placed the *Plasmodium falciparum* δ and α sequences at the base of each paralogous eukaryotic group; this result is not expected based on phylogenies using SSU rRNA (Cavalier-Smith, 1993), elongation factors (Baldauf and Palmer, 1993), actins (Drouin et al., 1995), α-, and β-tubulin (Keeling and Doolittle, 1996).

The Rev3 paralogs of *S. cerevisiae* and *M. musculus* formed another monophyletic group with high bootstap values (figure 2.5). Both parsimony and distance methods suggest that the Rev3 paralogs are most closely related to δ-type paralogs; this branching order is also supported by high bootstrap values. In parsimony analyses, the branching of the Rev3 paralogs with either α or ε paralogs was not observed in any of the shortest trees found (not shown). If accurate, the observed pattern of branching suggests that, in addition to the duplication event that resulted in the α and δ paralogs, another duplication occurred within the δ-type lineage to give rise to Rev3-like polymerases. Both of these duplication events must have occurred before, or very early in, the origin of eukaryotes.

## Ancient paralogy confuses the phylogenetic relationships of archaebacterial and eukaryotic family B DNA polymerases

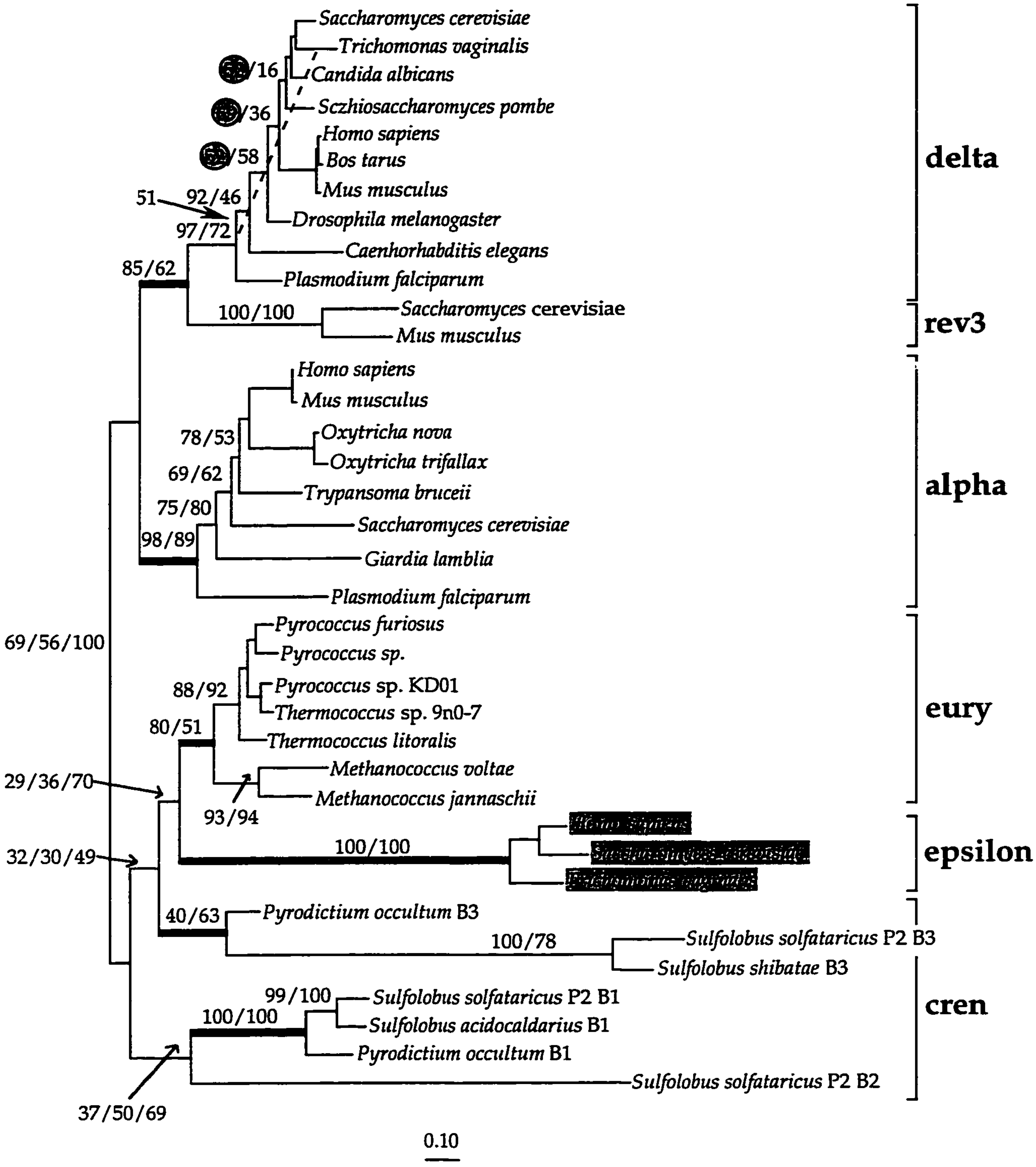
The finding of multiple family B DNA polymerases in two crenarchaeotes (Prangishvili and Klenk, 1994; Uemori et al., 1995; Edgell et al., 1997) and eukaryotes (Braithwaite and Ito, 1993), but as yet only a single family B DNA polymerase in euryarchaeotes, including the whole-genome sequence of *M. jannaschii* (Bult et al., 1996), raises a number of interesting questions concerning the evolution of archaebacterial and eukaryotic DNA polymerases. Foremost is whether or not the gene duplications that gave rise to the multiple crenarchaeote and eukaryotic paralogs occurred independently of one another after the split of archaebacteria and eukaryotes, or occurred in a common ancestor of eukaryotes and archaebacteria. If the latter case is true, are any of the crenarchaeote and eukaryote DNA polymerases orthologs? That is, if any of the crenarchaeote family B DNA polymerases performs analogous functions to those of the three eukaryotic paralogs, it would not be unreasonable to expect a specific phylogenetic relationship between particular crenarchaeote and eukaryotic polymerases.

Phylogenetic analyses do not recover a topology that would be consistent with gene duplications occurring after the split of eukaryotes and archaebacteria (figure 2.5); the four eukaryotic paralogs do not form a monophyletic group to the exclusion of archaebacterial sequences and one of the eukaryotic paralogs, ε, consistently branches with archaebacterial sequences as a sister group to euryarchaeotes (although bootstrap support for this branching pattern is low; table 2.2). There is no significant bootstrap support for the monophyly of all archaebacterial sequences (figure 2.5; table 2.2), but there is support for topologies that split the crenarchaeote sequences into two groups: one comprising the *S. solfataricus* P2 B1/*S. acidocaldarius*

B1/*P. occultum* B1 orthologs, and the other comprising the remaining crenarchaeote and euryarchaeote sequences. Grouping of the crenarchaeote B1 orthologs is also supported by two single amino acid deletions in polymerase domain IV (figure 2.4). The phylogenetic relationship of the *S. solfataricus* P2 B2 sequence to other archaebacterial and eukaryotic sequences is not well supported by bootstrap analysis. All methods support a grouping of the *S. solfataricus* P2 B2 sequence with the crenarchaeote B1 orthologs, but with bootstrap values that are not convincing (figure 2.5).

Based solely on amino acid identity, the ε-type polymerases are more similar to euryarchaeote polymerases than to any other eukaryotic or archaebacterial sequence (42.5% average identity of all euryarchaeote sequences to *H. sapiens* DNA polymerase ε; table 2.1). In addition, the alignment of eukaryotic and archaebacterial paralogs shows that in exonuclease domains II, ε and euryarchaeote polymerases are remarkably similar (figure 2.4). This amino acid identity, taken together with the phylogenies supporting a grouping of ε and euryarchaeote polymerases, suggests that these polymerases might be orthologs. However, a number of archaebacterial and eukaryotic paralogs have extremely long branch lengths relative to other paralogs. To measure the effect of these rapidly evolving taxa, I sequentially removed each paralog from the dataset and measured the effect on tree topology by bootstrap analysis (table 2.2). Removal of the *S. solfataricus* P2 B2 or *S. solfataricus* P2 B3/*S. shibatae* B3 paralogs had the greatest effect on the node supporting a sisterhood of ε and euryarchaeote paralogs, as bootstrap support increased when these taxa were removed. Bootstrap support for the node uniting the eukaryotic α/δ/Rev3 paralogs also increased, particularly when the *S. solfataricus* P2 B2 paralog was removed from analyses (table 2.2).

**Figure 2.5** Phylogenetic analysis of eukaryotic and archaebacterial family B DNA polymerases. The tree shown is from PROTDIST analysis with all eukaryotic and archaebacterial sequences. Bootstrap values are indicated above nodes in the order parsimony/ distance/ML. Nodes constrained during ML analysis are indicated by an oversize line. A similar topology was found by parsimony analysis (72 shortest trees, 1162 steps, CI=0.644, HI=0.356) but differing in terminal arrangements within eukaryotic orthologs, and in the placement of *Sulfolobus solfataricus* P2 B2 and the partial *T. vaginalis* δ sequences. All of the shortest trees in parsimony analysis placed the *T. vaginalis* sequence as an outgroup to all other δ-type sequences except those from *Caenhorabditis elegans* and *Plasmodium falciparum* with 51% bootstrap support (indicated by an arrow). A dashed line joins the *T. vaginalis* δ sequence with the node common to *C. elegans* and *P. falciparum*. Parsimony bootstrap values that are circled and shaded indicate that these values are for the shown topology, but with *T. vaginalis* branching at the base of the δ paralogs. Only 3% support was found in parsimony bootstrap for a grouping of *T. vaginalis* and *S. cerevisiae* versus 19% bootstrap support in distance methods.


Saccharomyces cerevisiae
Trichomonas vaginalis
/16
Candida albicans
/36
Sczhiosaccharomyces pombe
Homo sapiens
/58
Bos tarus
Mus musculus
51
92/46
Drosophila melanogaster
97/72
Caenhorhabditis elegans
85/62
Plasmodium falciparum
100/100
Saccharomyces cerevisiae
Mus musculus
Homo sapiens
Mus musculus
78/53
Oxytricha nova
Oxytricha trifallax
69/62
Trypansoma bruceii
75/80
Saccharomyces cerevisiae
98/89
Giardia lamblia
Plasmodium falciparum
69/56/100
Pyrococcus furiosus
Pyrococcus sp.
88/92
Pyrococcus sp. KD01
Thermococcus sp. 9n0-7
80/51
Thermococcus litoralis
29/36/70
Methanococcus voltae
Methanococcus jannaschii
93/94
Homo sapiens
100/100
Saccharomyces cerevisiae
32/30/49
Trichomonas vaginalis
40/63
Pyrodictium occultum B3
100/78
Sulfolobus solfataricus P2 B3
Sulfolobus shibatae B3
99/100
Sulfolobus solfataricus P2 B1
100/100
Sulfolobus acidocaldarius B1
Pyrodictium occultum B1
Sulfolobus solfataricus P2 B2
37/50/69
0.10
delta
rev3
alpha
eury
epsilon
cren

ML analysis also suggests that the ε-type and euryarchaeote polymerases might be orthologs as 70% support was found for this tree topology. In addition, all 100 log-likelihood trees found by ML analysis placed the ε paralogs as branching with archaebacterial polymerases to the exclusion of α, δ, and Rev3 paralogs (figure 2.5). As maximum likelihood methods, in general, perform better than parsimony or distance methods under conditions of rapid sequence evolution (Hasegawa and Fujiwara, 1993; Kuhner and Felenstein, 1994; Huelsenbeck, 1995; Swofford et al., 1996), the branching position of the eukaryotic ε paralogs within archaebacterial paralogs might well reflect the history of this gene family rather than be due to a treeing artifact.

## Discussion

### Duplication is a common theme in the evolution of family B DNA polymerases

The rooted universal phylogenies that place archaebacteria and eukaryotes as sister groups to the exclusion of eubacteria (Gogarten et al., 1989; Iwabe et al., 1989; Brown and Doolittle, 1995; Baldauf et al., 1996; Lawson et al., 1996) support a common origin for archaebacterial and eukaryotic DNA replication proteins (Edgell and Doolittle, 1997). It is not surprising then that archaebacterial family B DNA polymerases show greater sequence similarity (table 2.1) and phylogenetic affinity to the multiple eukaryotic homologs than to the *E. coli* family B DNA polymerase. What is surprising, however, is the finding that one of the eukaryote paralogs, ε, shows a tendency in phylogenetic analyses to branch with euryarchaeote homologs to the exclusion of other eukaryotic or crenarchaeote homologs. Although support for this branching order is weak and suspect because of high rates of sequence evolution of some paralogs, it is consistently recovered by parsimony, distance and maximum likelihood methods.

Three interpretations are possible. First, this implies that an ε-type polymerase is ancestral to archaebacteria and eukaryotes and has since been lost from the crenarchaeote lineage of archaebacteria, or has diverged in primary sequence to such an extent as to be unrecognizable as an ε-type polymerase. All other archaebacterial and eukaryotic paralogs are ultimately derived from this 'ancestral' polymerase. The second interpretation assumes that ε-type polymerases are not ancestral to archaebacteria and eukaryotes, that the α, δ, and Rev3 paralogs evolved by duplication at the origin of eukaryotes from an archaebacterial polymerase most resembling crenarchaeote B1 paralogs, and that the branching order and high amino acid identity of eukaryotic ε and euryarchaeote polymerases can be best explained by a lateral transfer event from euryarchaeotes to eukaryotes early in eukaryotic evolution. The third interpretation assumes that the recovered tree topologies are not significant, but that the true topology cannot be known because of rapid rates of sequence evolution of eukaryotic ε and crenarchaeote B2 and B3 paralogs.

Regardless of the origin of eukaryotic ε-like polymerases, both euryarchaeote and eukaryotic ε polymerases have independently undergone numerous changes in structure and function. For instance, the DNA polymerase ε proteins of *S. cerevisiae* and *H. sapiens* are both over 2200 amino acids in length (Morrison et al., 1990; Kesti et al., 1993), yet euryarchaeote polymerases are under 900 amino acids in length (see for example Uemori et al., 1993). These additional amino acids of DNA polymerase ε, present as a long carboxy-terminal extension relative to other family B homologs, are implicated in cell cycle regulation as deletion of this region interferes with a DNA replication checkpoint in S phase (Navas et al., 1995). Since archaebacteria do not possess a eukaryote-like cell cycle (or the elaborate checkpoint controls associated with one), it is likely that this region of DNA polymerase ε was acquired early in the evolution of eukaryotes. However, database searches with this region of the polymerase fail to match known sequences in public databases with any significance (not shown). In spite of studies demonstrating an absolute requirement for DNA polymerase ε in both replication (Morrison et al., 1990) and cell cycle regulation (Navas et al., 1995), the exact biochemical role of DNA polymerase ε at the eukaryotic replication fork is uncertain (Stillman, 1994). Likewise, the cellular function of the euryarchaeote family B DNA polymerase is unknown; there is no experimental evidence with which to address its function, but it is likely that the polymerase has *some* role in replication.

At least two other sets of duplication events, in addition to the event that gave rise to ε-like polymerases, must have occurred independently in the evolution of crenarchaeotes and eukaryotes. One set of duplications was that which gave rise to the multiple paralogs of crenarchaeotes, typified by the *S. solfataricus* P2 B2 and B3 polymerases (Prangishvili and Klenk, 1994; Edgell et al., 1997). Phylogenetic analyses cannot resolve the order of these duplications, nor accurately resolve the relationships among crenarchaeote family B DNA polymerases. Yet, it is clear from alignments (figures 1.2 and 2.4) that the crenarchaeote B2 and B3 paralogs evolved from an archaebacterial family B DNA polymerase and that both paralogs have experienced high rates of nonsynonymous amino acid substitutions at some point in their histories. Whether or not this reflects positive selection for a novel function(s) different from the parental gene is not clear, since the function of either the B2 and B3 paralogs is unknown. In the absence of such data, it is also difficult to ascribe a particular biochemical role for the multiple crenarchaeote paralogs analogous to those of the three eukaryotic replicative polymerases. It is entirely possible that the divergent family B paralogs of crenarchaeotes have some cellular role other than replication; they might function in DNA repair, as does the family B homolog in *E. coli* (see below) and the Rev3 paralog in *S. cerevisiae* (Morrison et al., 1989).

The α, δ, and Rev3 paralogs of eukaryotes must also have evolved by duplication since phylogenetic analyses strongly place these three paralogs as a monophyletic group (figure 2.5). This duplication event must have occurred independently of those that gave rise to the multiple paralogs of crenarchaeotes. All three of these eukaryotic paralogs share conserved amino acid positions, but are also sufficiently different in many positions suggesting that each paralog experienced high rates of nonsynonymous amino acid substitutions and probable selection for novel function(s). Biochemical evidence from mammalian cell lines and *S. cerevisiae* suggests that DNA polymerase α is responsible for the synthesis of leading and lagging strand primers (Tsurimoto et al., 1990; Waga and Stillman, 1994), while DNA polymerase δ is responsible for leading strand synthesis (Waga and Stillman, 1994). These changes in structure and function must have occurred early in the evolution of eukaryotes.

Genetic evidence suggests that the Rev3 paralog of *S. cerevisiae* functions exclusively in DNA repair (Morrison et al., 1989). If phylogenetic analyses indicating that a grouping of Rev3 and δ-type polymerases is correct, a change of function from replication-associated to repair must have occurred early in the evolution of this paralog. It is also possible that eukaryotic family B DNA polymerases ancestrally functioned in DNA repair and switched to function in replication at some point in the evolution of eukaryotes. In this case, the present-day function of the *S. cerevisiae* Rev3 paralog would more accurately represent the ancestral function of eukaryotic family B DNA polymerases than any other paralog. However, in the absence of experimental evidence pointing to a functional role of archaebacterial family B paralogs, it is easiest to assume that the ancestral function of eukaryotic family B polymerases was replication.

## The curious case of the eubacterial family B DNA polymerase

The replicative DNA polymerases of eubacteria (DNA polymerase III, a family C polymerase, and DNA polymerase I, a family A polymerase) and the replicative DNA polymerases of eukaryotes (all family B polymerases) do not share significant primary sequence similarity (Braithwaite and Ito., 1993; chapter IV this thesis). However, one homologous DNA-dependent DNA polymerase, a family B polymerase, is present in representatives of all three domains (Iwasaki et al., 1991; Braithwaite and Ito, 1993). In *E. coli*, this polymerase (*pol*B) does not have a replication function but instead functions in DNA repair (Bonner et al., 1990; Iwasaki et al., 1990). Interestingly, *pol*B can interact with the eubacterial processivity factor, Polβ (encoded by the *dna*N gene), and with the clamp loading γ complex (see chapter III; Hughes et al., 1991; Bonner et al., 1992). These proteins primarily associate with the eubacterial replicative polymerase, *pol*C, implying that association of these proteins with *pol*B could confer processive replication on *pol*B. Recent experimental evidence has demonstrated that *pol*B does replicate chromosomal and episomal (F') DNA in dividing cells, but only in the presence of an antimutator allele of *pol*C (Rangarajan et al., 1997). It is not clear if *pol*B is used in a replicative function in logarithmically growing wild-type cells.

The presence of this homologous family B DNA polymerase in eubacteria, archaebacteria and eukaryotes suggests that it must have been present in the genome of the cenancestor. Yet the cenancestral function is unclear; a change of function(s) must have occurred in either the eubacterial lineage (where the family B homolog now functions primarily as a repair polymerase) or the archaebacterial/eukaryotic lineage (where family B homologs now function as a replicative polymerases). However, a family B homolog is missing, or has diverged so much in primary sequence as to be unrecognizable by common database search algorithms, from the completely sequenced eubacterial genomes of *Haemophilus influenzae* (Fleischman et al., 1995), *Mycoplasma genitalium* (Fraser et al., 1995), *Mycoplasma pneumoniae* (Himmelreich et al., 1996), and *Synechocystis* sp. strain PCC6803 (Kaneko et al., 1996), and from the partially sequenced genomes of *Bacillus subtilus* (47% of 4.2Mb completed; http://www.pasteur.fr/Bio/SubtiList.html), *Mycobacterium tuberculosis* (% completion not known; http://www.sanger.ac.uk/Projects/M_tuberculosis/), *Neisseria gonorhoeae* (88.9% of 2.7 Mb completed; http://dna1.chem.uoknor.edu/gono.html), and *Streptococcus pyogenes* (88.3% of 1.98 Mb completed; http://dna1.chem.uoknor.edu/strep.html). Escarceller and colleagues found evidence for the presence of a family B homolog in proteobacteria other than *E. coli* by Southern hybridization using the *E. coli* gene as a probe and by western blots with antibodies raised against *pol*B (Escarceller et al., 1994). But, in the absence of sequence data to confirm these preliminary results, *E. coli* remains the only eubacterium to possesses a family B DNA polymerase. Either a family B DNA polymerase was present ancestrally and lost independently from eubacterial lineages except that leading to *E. coli*, or *E. coli* acquired this polymerase by lateral transfer from a non-eubacterial source.

The *E. coli* family B DNA polymerase sequence can be aligned with archaebacterial and eukaryotic homologs (figure 2.4), but is not specifically close to any particular archaebacterial or eukaryotic sequence as judged by amino acid identity (table 2.1). This lack of obvious sequence identity to any eukaryotic or archaebacterial sequence can be interpreted as evidence against horizontal transfer of this polymerase to *E. coli* from a non-eubacterial source. If this family B DNA polymerase is ancestral to eubacteria, the reason behind its systematic and independent deletion from multiple eubacterial genomes, except that of *E. coli,* remains unknown.

# Chapter III. Gene duplications in the evolution of other protein components of the archaebacterial and eukaryotic DNA replication apparatus

## Introduction

The finding of multiple independent gene duplications in the evolution of archaebacterial and eukaryotic family B DNA polymerases stimulated a search for other replication-associated proteins of eukaryotes and archaebacteria that also evolved by gene duplications. This search for gene families was entirely computer-based because (1) complete genome sequence from a variety of eubacteria, archaebacteria, and *S. cerevisiae* were available, and (2) PCR-based surveys often fail to isolate all paralogous sequences from a single organism.

Results of database searches with eukaryotic replication proteins could be grouped into two categories. The first category included searches showing that, in general, replication fork proteins performing analogous functions in eubacteria and eukaryotes do not share significant amino acid similarity (discussed in detail in chapter IV). However, database searches often revealed the presence of multiple archaebacterial and eukaryotic paralogs of these replication proteins. Proteins involved in the control of initiation of replication fall into this category. The second category included eubacterial and eukaryotic proteins with significant sequence similarity that perform analogous functions at the replication fork, but fo which eukaryotes possess multiple paralogs. Clamp loading proteins involved in the loading of the DNA polymerase and additional accessory factors onto the template are included in this category.

Clamp loading proteins

DNA-dependent DNA polymerases of eubacteria and eukaryotes (and presumably archaebacteria) cannot synthesize DNA *de novo* and thus require the actions of a number of accessory proteins to "load" or assemble the DNA polymerase onto the activated DNA template (Kornberg and Baker, 1992; Kelman and O'Donnell, 1995). After origin unwinding and synthesis of the primer, the next step in the initiation of both eubacterial and eukaryotic replication is the binding of a complex of "clamp loading" proteins (O'Donnell et al., 1993; Stillman, 1994). The clamp loading complex recognizes and preferentially binds to the 3' end of the primer, thus acting as a primer recognition factor for the replicative DNA polymerase(s) (O'Donnell et al., 1993). In addition, the binding of the clamp loading complex stimulates assembly of the processivity factor (often called the sliding clamp). This ring shaped protein encircles duplex DNA behind the primer-template junctions and functions to confer processive replication to the polymerase by tethering the polymerase to the template (Stukenberg et al., 1991; Naktinis et al., 1996). Binding of the sliding clamp protein to dsDNA and to the clamp loading protein stimulates the ATPase activity of the clamp loading proteins so that the DNA polymerase is "loaded" onto the active template.

The biochemistry of the eukaryotic and eubacterial clamp loading proteins is very similar; both sets of proteins are DNA-dependent ATPases, and both specifically recognize primer-template junctions (reviewed in Stillman, 1994). In *E. coli*, the clamp loading complex (composed of five proteins and also called $\gamma$ complex) was first isolated as part of a large molecular weight complex that included DNA polymerase III core subunits and accessory factors (reviewed in Kornberg and Baker, 1992). Subsequent purification of the accessory factors resulted in the identification of $\gamma$ complex proteins. Only two, encoded by the *hol*B and *dna*X genes, show similarity to eukaryotic proteins with analogous functions (see below). In eukaryotes, the clamp loading proteins (called replication factor-C, individual proteins are designated RFC1-5) were first identified by studying SV40 replication in mammalian cell lines (Tsurimoto and Stillman, 1989). As in *E. coli,* five proteins are responsible for the loading of the processivity factor and DNA polymerases onto the activated template (Cullman et al., 1995). All five proteins share significant amino acid similarity to each other, to the eubacterial DnaX gene products, and to ORFs found in the complete genome sequence of *Methanococcus jannaschii* (see below).

## Minichromosome maintenance (MCM) proteins

Eukaryotes replicate their nuclear genome only once per cell cycle. The precise control of replication initiation is not well understood, although a number of proteins involved in this process have been identified through mutational studies in *S. cerevisiae* (reviewed in Kearsey et al., 1996; Rowles and Blow, 1997). In the mechanistic control of initiation of replication, eukaryotes (at least *S. cerevisiae*) differ from eubacteria (at least *E. coli*) in that initiation requires many *trans*-acting, positively-acting protein factors (Diffley, 1997). By contrast, replication initiation in *E. coli* requires the modification (by methylation) of *cis*-acting DNA sequences and the actions of a protein, SeqA, that sequesters origin sequences and acts to negatively regulate initiation (Slater et al., 1995).

Blow and Laskey proposed in 1988 that the initiation of replication in S-phase was controlled by a protein(s) which they called licensing factor. Among the relevant characteristics of a licensing factor would be the ability to modify chromatin (either by directly binding to DNA or by enzymatic modification), that this modification would be absolutely required for the initiation of replication, that the modification would result in only a single round of initiation of replication per cell cycle, and that it be a diffusable factor excluded from the nucleus (Blow and Laskey, 1988). Mutant hunts in *S. cerevisiae* for cells unable to go through the G1/S-phase transition, or cells unable to support the replication of plasmids carrying ARS consensus sequences, resulted in the identification of a number of proteins with some of the expected characteristics of licensing factors (Yan et al., 1991; Gibson et al., 1990; Chen et al., 1992). For clarity, I will refer to these proteins as minichromosome maintenance (MCM) proteins (after Maine et al., 1984), although they have at some point had other designations (ie. MCM4 in *S. cerevisiae* was originally isolated as a cell division control mutant, CDC54).

The exact biochemical role of MCM proteins is not known, although MCM proteins of *S. cerevisiae* have been shown to bind chromatin at levels that exceed the number of active replication origins; moreover, this binding somehow limits initiation to once per cell cycle (Donovan et al., 1997). In all, six MCM proteins have been identified in *S. cerevisiae, S. pombe, Xenopus laevius* and mammalian systems as essential for the control of initiation of replication (Kearsey et al., 1996). Three archaebacterial homologs are present in the complete genome sequence of *M. jannaschii*; there are no known eubacterial MCM homologs.

## Results

### Database searches for eukaryotic DNA replication proteins that are members of a gene family

To search for protein components of the eukaryotic replication apparatus that are members of a gene family, I individually searched public

**Table 3.1** Results of database searches with eukaryotic replication-associated proteins. Biochemical functions and name(s) of replication associated proteins are listed for each protein used to search databases. Database matches similar to those of the query sequence are listed as orthologous sequences. Matches to proteins with unknown function(s), or function(s) not related to replication, are listed as paralogous sequences. Eubacterial (Eub) or archaebacterial homologs of multiple eukaryotic proteins performing the same biochemical function are presented as hits to each of the eukaryotic proteins. The abbreviation A or F represents more than two animal (A) or fungal (F) sequences that matched the query sequence. Human or yeast (for example) refers to a single animal (*Homo sapiens*) or fungal (*Saccharomyces cerevisiae*) sequence that matched the query sequence. Other organisms from which orthologous or paralogous sequences have been sequenced are listed by their species name (ie. Arabidopsis, Methanobacterium). *Methanococcus jannaschii* and *Escherichia coli* have been shortened to MJ and E.coli respectively. *Schziosaccharomyces pombe* is shortened to pombe.

| Gene/protein/function at replication fork | Orthologous sequences in database | Other paralogs |
|---|---|---|
| **ORC complex**<br>• origin recognition; ATP-dependent DNA binding<br>• 6 non-homologous subunits (ORC1-6) | ORC1 - A/F/Methanobacterium<br>ORC2 - A/F<br>ORC3 - yeast<br>ORC4 - yeast<br>ORC5 - yeast/Drosophila<br>ORC6 - yeast | ORC1-related: CDC6 yeast (P09119), CDC18 pombe (P41411) |
| **Primase**<br>• synthesis of RNA/DNA primer. In eukaryotes, DNA polymerase α is the primase. | A/F/Trypanosoma/Plasmodium/Oxytricha/many archaeal sequences/E. coli | |
| **Initiation**<br>• MCM (mini-chromosome maintenance) proteins are required for iniitation of DNA replication (often called licensing factors).<br>• 6 proteins identified in yeast MCM 2-7 | MCM2 - A/yeast/ MJ<br>MCM3 - yeast/A/MJ/ Entamoeba<br>MCM4 - A/F/MJ<br>MCM5 - A/F/MJ<br>MCM6 - A/F/MJ<br>MCM7 - A/F/Arabidopsis | |
| **Replication Protein-A**<br>• single-stranded DNA binding, stimulates DNA polymerase and loading of helicase<br>• 3 non-homologous subunits (RPA 1-3) | RPA1 14kDa - human/yeast<br>RPA2 32kDa - A/F/Crithidia<br>RPA3 70kDa - A/F/Crithidia | |
| **Proliferating Cell Nuclear Antigen (PCNA)**<br>• processivity factor, stimulates DNA polymerases and DNA-dependent ATPase (RF-C complex). | A/F/plants/Plasmodium/MJ | DNA-repair protein XPG (U40796); Excision repair protein ERCC5 (A54439) |

| Gene/protein/function at replication fork | Orthologous sequences in database | Other paralogs |
|---|---|---|
| RNase H1<br>• nuclease for removal of RNA primer | A/F/Crithidia/MJ/Eub | |
| FEN-1 endonuclease<br>• nuclease for removal of RNA primers | A/F/MJ | |
| Replication Factor-C<br>• DNA-dependent ATPase; binds primer template, stimulates loading of DNA polymerase.<br>• 5 homologous subunits (RFC 1-5) | RFC1 - A/F/MJ/Eub<br>RFC2 - A/F/MJ/Eub<br>RFC3 - A/F/MJ/Eub<br>RFC4 - A/F/MJ/Eub<br>RFC5 - A/F/MJ/Eub | RFC1-related: DSEB mouse (A56284); ISRE-mouse (U07157); human PO-GA (JN0599); CHL12-yeast (S50340); GNF1-Drosophila (P35600) |
| DNA ligase (ATP-dependent)<br>• ligation of Okazaki fragments on lagging strand | A/F/Arabidopsis/MJ/ Crithidia/Methanobacterium | DNA ligase II, III (X84740), IV (X83441) |
| DNA-dependent DNA polymerases<br>**alpha** - primer synthesis<br>**delta** - leading/lagging strand synthesis<br>**epsilon** - leading/lagging strand synthesis | **alpha** - A/F/Plasmodium/ Oxytricha/Trypanosoma/ many archaeal sequences/ E. coli<br>**delta** - A/F/Plasmodium/ many archaeal sequences/ E. coli<br>**epsilon** - A/F/many archaeal sequences/E. coli | REV3 - yeast DNA repair polymerase (P14284) |

databases with proteins that had been identified as essential for DNA replication in *S. cerevisiae* (table 3.1; based on Kornberg and Baker, 1992; Stillman, 1994). Initially, I used P-values at or below $10^{-6}$ in BLASTP and TBLASTX searches as a cutoff for identifying similar sequences in the database. I found little difference in the sensitivity of the BLAST (Altschul et al., 1990) or FASTA (Pearson and Lipman, 1988) algorithms to detect similarities between eukaryotic replication proteins and similar sequences in databases. Database sequences that were similar to query sequences, but at significance less than $10^{-6}$, were individually examined to determine whether matches were significant. Except for one example, the HolB proteins (see below), all of these potential similarities were rejected as not significant.

Of all eukaryotic protein components used to search databases, only three gene families could be identified whose members performed replication functions: DNA polymerases, MCM and RF-C proteins (table 3.1). Eubacterial proteins with significant database scores to eukaryotic DNA polymerases were found, but these do not have a replicative function; this includes the *E. coli* family B DNA polymerase which functions in DNA repair, not replication (see chapter II). Other eukaryotic paralogs were identified for a number of replication-associated proteins. For instance, there were many significant database hits to RFC-1, but these proteins have been identified as transcription factors or DNA-binding proteins. One significant database match, the Differentiation-specific element binding protein (DSEB), was isolated from human cell lines by the ability to bind to a *cis*-acting transcriptional sequence located in the 5' non-coding region of the angiotensinogen gene (McGehee and Habener, 1995). Another eukaryotic replication-associated protein, FEN-1, a 5'-3' exonuclease required for removal of primers, also had many hits to proteins with functions other than replication. FEN-1 is a member of a large paralogous family of eukaryotic nucleases which includes excision/repair proteins (reviewed in Lieber, 1997).

Database searches with MCM proteins did not reveal any paralogous proteins with non-replication associated functions, as was the case with RF-C or FEN-1 proteins (table 3.1). Eukaryotic MCM, FEN-1, and RF-C proteins all have homologs in the complete genome sequence of *M. jannaschii*, but only the RF-C proteins have homologs in eubacteria; no significant scores were obtained in database searches with MCM or FEN-1 proteins against eubacterial genomes.

In addition to significant search scores to eubacterial clamp loading proteins, database searches with the largest subunit of the eukaryotic clamp loading complex (RFC-1) revealed significant similarity to eubacterial NAD-dependent DNA ligases (figure 3.1). RFC-1 possesses a long amino-terminal extension relative to other eukaryotic RF-C paralogs; only this amino-terminal extension of the RFC-1 protein shows significant similarity to eubacterial DNA ligases. No other eukaryotic, archaebacterial or eubacterial RF-C paralogs possess this extension and none exhibits similarity to eubacterial DNA ligases.

Although the HolB protein is a component of the eubacterial clamp loading complex, encoding the δ′ subunit (Carter et al., 1993; Dong et al., 1993), it shows limited sequence similarity to the eubacterial *dna*X gene (the γ subunit), archaebacterial and eukaryotic RF-C proteins (O'Donnell et al., 1993). Since sequence similarity is confined to the amino-terminal portion of the protein (domains II-IV of figure 3.4), and since HolB proteins are difficult to align with eubacterial DnaX and archaebacterial and eukaryotic RF-C proteins, no HolB proteins were included in phylogenetic analysis. The other components of the eubacterial clamp loading complex (the HolA, C, and D

```
                          [411]                                                  [466]
       H.sapeins RFC-1    IFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDSGQSKSDKAAAL
     M.musculus RFC-1    TFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDSGQSKSDKAAAL
 D.melanogaster RFC-1    TFVVTGVLESMEREEAESVIKEYGGKVMTVVGKKLKYLVVGEEAGPKKLAVAEEL
 D.melanogaster GNF-1    TFVVTGVLESMEREEAESVIKEYGGKVMTVVGKKLKYLVVGEEAGPKKLAVAEEL
       H.sapiens PO-GA    IFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDSGQSKSDKAAAL
      M.musculus DSEB    TFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDSGQSKSDKAAAL
      M.musculus ISRE    TFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDSGQSKSDKAAAL
 A.platyrhynchos RFC-1    TFVITGVLESIERDEAKSLIERYGGKVTGNVSKKTNYLVMGRDCGQSKCEKASAL
   S.cerevisiae RFC-1    TIVFTGVLPTLERGASEALAKRYGARVTKSISSKTSVVVLGDEAGPKKLEKIKQL
     E.nidulans RFC-1    TFVFTGVLDTLGREEGQALVKRYGGKVTTAPSGKTSFVVLGSDAGPSKLATISKH
       E.coli DNA lig    TVVLTGSLSQMSRDDAKARLVELGAKVAGSVSKKTDLVIAGEAAG.SKLAKAQEL
  H.influenzae DNA lig    TVVLTGTLTQMGRNEAKALLQQLGAKVSGSVSSKTDFVIAGDAAG.SKLAKAQEL
Synechocystis sp.DNA lig    TFVLTGTLPNLSRLEAQELIEQSGGKVTSSVSTKTDYVLLGDKPG.SKAAKAESL
  M.genitalium DNA lig    RFLITGSFNIS.RDQIKDLLSAKFDCQFASEVKPTVDFVIAGNKP.TLRKINHAK
    R.marianus DNA lig    TFVLTGALPHLTRKEAEELIKRAGGRVASSVSRNTDYVVVGENPG.SKYDRARQL
   T.aquaticus DNA lig    TFVITGELSRP.REEVKALLRRLGAKVTDSVSRKTSYLVVGENPG.SKLEKARAL
    Z.mobilis DNA lig    IIVFTGSLQKITRDEAKRQAENLGAKVASSVSKKTNLVVAGEAAG.SKLSKAKEL
                          [654]                                                  [707]
```

**Figure 3.1** Amino acid alignment of eubacterial NAD-dependent DNA ligases and a portion of the amino-terminal region of RFC-1 from a variety of eukaryotes. Paralogs of RFC-1 with functions other than replication are also included in the alignment. Numbering of amino acids corresponds to *Homo sapiens* RFC-1 (top) and to *Zymomonas mobilis* NAD-dependent DNA ligase (bottom). Amino acids highlighted in bold indicate identical or conserved amino acids shared between 95% of taxa included in the alignment.

proteins; Xiao et al., 1993; Dong et al., 1993) are not similar to the HolB or DnaX subunits, nor to archaebacterial or eukaryotic RF-C proteins.

### Phylogenetic analyses suggests that multiple independent gene duplications occurred in the evolution of MCM and RF-C proteins

Two protein families, RF-C and MCM, were chosen for phylogenetic analyses. All available eukaryotic, archaebacterial and, when available, eubacterial homologs, were retrieved from public databases. Two unpublished MCM sequences, one from *S. solfataricus* P2 and the other from *Nosema locustae* (Logsdon and Doolittle, unpublished), were included in analyses. All available eubacterial *dna*X sequences, except those that are almost identical (ie. *Mycoplasm genitalium* and *Mycoplasma pneumoniae,* in which case one was used) were included in analyses. The final RF-C dataset consisted of 26 taxa and 110 amino acid positions (figure 3.4), while the MCM dataset consisted of 36 taxa and 309 amino acid positions (figure 3.2).

### MCM phylogeny

Both parsimony and distance analyses of the MCM proteins resolved the eukaryotic paralogs into six distinct monophyletic groups, each supported by high bootstrap values (figure 3.3). This is the expected tree topology of a gene family that has evolved by a series of duplications. However, the relationship between eukaryotic paralogs was not very well resolved, as low bootstrap values were recovered for nodes uniting various paralogs with one another (figure 3.3). One significant finding, supported by parsimony and distance methods, was the branching of the *S. solfataricus* P2 MCM homolog with eukaryotic MCM2 paralogs to the exclusion of other archaebacterial MCM sequences from *M. jannaschii.* This phylogenetic result is also

**Figure 3.2** Amino acid alignment of archaebacterial and eukaryotic MCM proteins. Numbering is according to *Saccharomyces cerevisiae* MCM2. Horizontal lines divide the alignment into orthologous groups (ie. MCM2-7) and separate the archaebacterial from eukaryotic paralogs. The *Nosema locustae* sequence (Nosema) is separated from the other eukaryotic paralogs because it does not show sequence or phylogenetic affinity for any one particular ortholog. The *M. jannaschii* MCM sequences are listed along with their TIGR database accession number. Gaps introduced into the alignment are indicated by a period (.). Bullets (•) below the *Sulfolobus solfataricus* P2 MCM sequence indicate similar or identical amino acids shared with eukaryotic MCM2 paralogs. Amino acids highlighted in bold indicate residues that are similar or identical in at least half of each of the eukaryotic and archaebacterial paralogs.

```
                      307                          336  369
  Saccharomyces MCM2  RELRESNLSSLVR.VGVVTRRTGVFPQLKY  SKGPFRVNGEKTYRNYQRVTLQEAPGTVPP
     Drosophila MCM2  RTFRKLHLNQLVRTLGVVTATTGVLPQLSV  TGPFSINMEQTLYRNYQKITLQESPGRIPA
        Xenopus MCM2  RSLRQLHLNQLIRTSGVVTCCTGVLPQLSM  FGPFEINMEETVYQNYQRITIQESPGKVAA
  Saccharomyces MCM3  RTLTAQHLNKLVSVEGIVTKTSLVRPKLIR  GNKLTTEYGYSTFIDHQRITVQEMPEMAPA
       Triturus MCM3  RTLGAQFLGNLLCVEGIVTKCSLVRPKVMR  NNPLETEYGLCTYKDHQTLTIQEMPEKAPA
        Xenopus MCM3  RTLTASLLGSLVCVEGIVTKCSLVRPKVMR  NNPLETEYGLSTYKDHQTLSIQEMPEKAPA
      Entamoeba MCM3  RNITASILQQKVAVQGIITKSSQIRPLLQT  GKPLELEPGLSTYKDFQTLVVQEMPESAPT
        Xenopus MCM4  RSLNPEDIDQLITISGMVIRTSQIIPEMQE  THSMALIHNRSMFSDKQMIKLQESPEDMPA
           Homo MCM4  RNLNPEDIDQLITISGMVIRTSQLIPEMQE  THSMALIHNRSLFSDKQMIKLQESPEDMPA
  Saccharomyces MCM4  RELNPNDIDKLINLKGLVLRSTPVIPDMKV  THSMALIHNRSFFSDKQMIKLQESPEDMPA
            Mus MCM4  RNLNPEDIDQLITISGMVIRTSQLIPEMQE  PNSMSLIHNRCSFADKQVIKLQETPDFVPD
        S.pombe MCM4  RDLNPGDIDKLISIKGLVLRCTPVIPDMKQ  TNAMQLIHNRSEFADKQVIKLQETPDVVPD
      Drosophila dpa  RSLNPEDMDQLISISGMVIRSSNVIPEMRE  A.AGQTPHNVLLYAHDLVDKVQPGDRVTVT
  Saccharomyces MCM5  RDLDSEHVSKIVRLSGIIISTSVLSSRATY  PDPYIIIHESSKFIDQQFLKLQEIPELVPV
        S.pombe MCM5  RNLTASHISKLVRVPGIIIGASTLSCRATA  MDPFIIDHSKSTFIDQQVLKLQEAPDMVPV
           Homo MCM5  RSLKSDMMSHLVKIPGIIIAASAVRAKATR  LDPYFIMPDKCKCVDFQTLKLQELPDAVPH
     Drosophila MCM5  RQLKSDCVSKLVKIAGIIVAASGISAKATR  LDPFFIMPDKCKCVDFQTLKLQELPDFVPQ
        Xenopus MCM5  RSLKSEQMSHLVKIPGIIIAATAVRAKATK  LDPYFIIPDKCKCVDFQTLKLQESPDAVPH
      C.elegans MCM5  RQVKSAQVSQVVKISGIIVAAAQVRSKATK  IDPYIMLPDKCECVDYQTLKLQENPEDVPH
            Mus MCM6  RELTSSRIGLLTRISGQVVRTHPVHPELVS  RKRFLLDTNKSRFVDFQKVRIQETQAELPR
        S.pombe MCM6  RDLRTDRIGRLTTITGTVTRTSEVRPELAQ  KRSWRLNISQSSFQDCQKVRIQENSNEIPT
           Homo MCM6  RELTSSRIGLLTRISGQVVRTHPVHPELVS  RRRFLLDTNKSRFVDFQKVRIQETQAELPR
    Arabidopsis MCM6  REVKASHIGQLVRISGIVTRCSDVKPLMAV  RLNSKAGNPILQLRASKFLKFQEAKMQELA
  Saccharomyces MCM6  RDIRSEKIGSLLSISGTVTRTSEVRPELYK  RAFWTLNVTRSRFLDWQKVRIQENANEIPT
      C.elegans MCM6  RELSADKVGGLVRIAGQIVRTHPVHPELSR  RTRFSLDVNSSTFVDFQKIRIQETQAELPR
 C.elegans MCM6.like  RELSADKVGGLVRIAGQIVRTHPVHPELSR  RTRFSLDVNSSTFVDFQKIRIQETQAELPR
  Saccharomyces MCM7  RQIKGDFLGQLITVRGIITRVSDVKPAVEV  KGQLFMSTRASKFSAFQECKIQELSQQVPV
        Xenopus MCM7  RDVKADSIGKLVNVRGIVTRVTEVKPMMVV  GGRLYLQTRGSKFIKFQELKIQEHSDQVPV
            Mus MCM7  REVRADSVGKLLTVRGIVTRVSEVKPRMVV  GGRLYLQTRGSKFVKFQEMKIQEHSDQVPV
           Homo MCM7  REVRADSVGKLVTVRGIVTRVSEVKPKMVV  GGRLYLQTRGSRFIKFQEMKMQEHSDQVPV
              Nosema  ..............................  LFNVLTDIPDIRCRDFQEIKIQEMFYESKM
Methanococcus MJ0363  SELSSAHKGKLVEFRAMILQATKLKLRYAK  CKGLIFDEDLSGKVDFQEIKVQTPLQESIY
Methanococcus MJ0961  YTYTPCDGRVEIEIDDYFSEGEFIKDMLSP  EIKFILDEYDSIYVNIQEMEIQQPIDLMKN
Methanococcus MJ1489  PKCGREVVR.EIDILN..TD.SE.KAVCEC  GAELNLIEHKSIYTDFQEIKVQQPLDLMEN
          Sulfolobus  ..............................  PGQFRLIPEKTKLIDWQKAVIQERPEEVPS
                                                                  •   ••• •

                                                429  489
  Saccharomyces MCM2  GRLPRHREVILLADLVDVSKPG.EEVEVTGI  RDRGIIDKIISS.APSIYGHRDIKTAVAC
     Drosophila MCM2  GRIPRSKDVILLADLCDQCKPG.DELEVTGI  KDPRIVERVVASMAPSIYGHDYIKRALAL
        Xenopus MCM2  GRLPRSKDAILLADLVDSCKPG.DEIELTGI  KDERIGERIFASIAPSIYGHEDIKRGLAL
  Saccharomyces MCM3  GQLPRSIDVILDDDLVDKTKPG.DRVNVVGV  KKKDIFDILSQSLAPSIYGHDHIKKAILL
       Triturus MCM3  GQLPRSIDIIADDDLVDSCKPGSDRVQIVGI  HSKDIFEHLSKSLAPSIHGHEYIKKAILC
        Xenopus MCM3  GQLPRSVDIIADDDLVDKCKPG.DRVQIVGI  HSKDIFEHLSKSLAPSIHGHEYIKKAILC
      Entamoeba MCM3  GQMPRSVIVILLDQLVDKGKPG.DRVIINGT  KEENPINLFSKSIAPSIYGHSDVKKAILL
        Xenopus MCM4  GQTPHTTILYGHNDLVDKVQPG.DRVNVTGI  AKPDIYERLAAALAPSIYEHEDIKKGILL
           Homo MCM4  GQTPHTVILFAHNDLVDKVQPG.DRVNVTGI  RKPDIYERLASALAPSIYEHEDIKKGILL
  Saccharomyces MCM4  GQTPHTIVLFAHNDLVDKVQPG.DRVNVTGI  RKPDIYERLASALAPSIYEHEDIKKGILL
            Mus MCM4  GQTPHSISLCVYDELVDSCRAG.DRIEVTGT  AREDLYSLLARSIAPSIYELEDVKKGILL
        S.pombe MCM4  GQTPHSVSLCVYDEL.DSARAG.DRIEVTGI  KRDDIYDILSRSLAPSIYEMDDVKKGLLL
      Drosophila dpa  GIYRATPLKTGGLSSSVKSVYK.THVDVVHF  LAKKPDIYDRLARAIAPSIYENDDIKKGI
  Saccharomyces MCM5  GEMPRNLTMTCDRYLTNKVIPG.TRVTIVGI  RNPKLYEILTNSIAPSIFGNEDIKKAIVC
        S.pombe MCM5  GELPRHILLNADRYLTNQITPG.TRCVITGI  RTPNLYDIISNSISPAIYGNVDIKKAIAC
           Homo MCM5  GEVPRHMQLYCDRYLCDKVVPG.NRVTIMGI  ALPNVYEVISKSIAPSIFGGTDMKKAIAC
     Drosophila MCM5  GEIPRHLQLFCDRSLCERVVPG.NRVLIQGI  ASGDIYERLSQSLAPSIFGSRDIKKAITC
        Xenopus MCM5  GELPRHMQLYCDRYLCDKVVPG.NRVTIMGI  AKPDIYETVAKSIAPSIYGSSDIKKAIAC
      C.elegans MCM5  GEMPRHLQLFTERYLTDKVVPG.NRVTIVG.  QRKDAYELIAKSIAPSIYGSADIKKSIAC
            Mus MCM6  GSIPRSLEVILRAEAVESAQAG.DRCDFTGA  QDKNLYHNLCTSLFPTIHGNDEVKRGVLL
        S.pombe MCM6  GSMPRTLDVILRGDIVERAKAG.DKCAFTGI  HSDHIYSRLSNSLAPSVYGHEIIKKGILL
           Homo MCM6  GSIPRSLEVILRAEAVESAQAG.DKCDFTGT  QDKNLYHNLCTSLFPTIHGNDEVKRGVLL
    Arabidopsis MCM6  EHVPKGHIPRSMTVHLRGELTR.KVSPGDVV  EDGDIYNKLSRSLAPEIYGHEDIKKALLL
  Saccharomyces MCM6  GSMPRTLDVILRGDSVERAKPG.DRCKFTGV  KDEHIYDKLVRSIAPAVFGHEAVKKGILL
      C.elegans MCM6  GSIPRTVDVIVRGEMVETVQPG.DKCDIVGT  DDKKIEKNIVDSLFPNIYGNHEVKLGVLL
 C.elegans MCM6.like  GSIPRTVDVIVRGEMVETVQPG.DKCDIVGT  DDKKIEKNIVDSLFPNIYGNHEVKLGVLL
  Saccharomyces MCM7  GHIPRSLNIHVNGTLVRSLSPG.DIVDVTGI  TSGDVYNRLAKSIAPEIYGNLDVKKALLL
        Xenopus MCM7  GNIPRCMSVYVRGENTRLAQPG.DHVGITGV  TEEDFYEKLAASIAPEIYGHEDVKKALLL
            Mus MCM7  GNIPRSITVVLEGENTRIAQPG.DHVSVTGI  AEEDFYEKLAASIAPEIYGHEDVKKALLL
           Homo MCM7  GNIPRSITVLVEGENTRIAQPG.DHVSVTGI  AEEDFYEKLAASIAPEIYGHEDVKKALLL
              Nosema  PRVIEAVLYDDLVG...CLAPG.EVVHVLGI  KTPNLLSSLTHSVFPEIFGNEIIKAGLVL
Methanococcus MJ0363  ANKHSTTVFYEFNKPKKAVYSG..YVKIVGV  KDKNVIQKLSDYAFREVTGYDMIKRAVLL
Methanococcus MJ0961  PEEPARSIRVFLENTP.GIYAG.R.VNVIGR  RKKNIIDILSNYLISQIKGYELVKKAIFL
Methanococcus MJ1489  PEEPPKYITVFLENSP.GIYAG.R.VKITGI  KRKDVVNILADRLIPEIKGHSAIKKAVLL
          Sulfolobus  GQLPRQLEIILEDDLVDSARPG.DRVKVTGI  KDPWIRDRIISSIAPSIYGHWELKEALAL
                      • •••  • ••  •• • •••• • • ••••  ••  • ••• •• •••••••  ••  •••
  Saccharomyces MCM2  SLFG.GVPK..NVNG.KHSIRGDINVLLLGDPGTAKSQILK.YVEKTARAVFATGQGASAVG
     Drosophila MCM2  ALFG.GESK...NPGEKHKVRGDINLLICGDPGTAKSQFLKYTEKVAPRAVFTTGQGASAVG
        Xenopus MCM2  ALFG.GEAKNP..GGK.HKVRGDINVLLCGDPGTAKSQFLKYVEKVASRAVFTTGQGASAVG
  Saccharomyces MCM3  MLM..GGVEKNLENG..SHLRGDINILMVGDPSTAKSQLLRFVLNTASLAIATTGRGSSGVG
       Triturus MCM3  MLL..GGNEKILENG..TRIRGDINVLLIGDPSVAKSQLLRYVLHTGPRAIPTTGRGSSGVG
        Xenopus MCM3  MLL..GGNEKVLENG..TRIRGDINVLLIGDPSVAKSQLLRYVLHTAPRAIPTTGRGSSGVG
      Entamoeba MCM3  MLV..GATPKIRLRS...RVRGDIHVMLCGDPSTAKSQLLRYVMSIAPLAVSTNGRGATGVG
        Xenopus MCM4  QLFG.GTRKDFSHTGRG.KFRAEVNILLCGDPGTSKSQLLQYVFNLVPRGQYTSGKGSSAVG
           Homo MCM4  QLFG.GTRKDFSHTGRG.KFRAEINILLCGDPGTSKSQLLQYVYNLVPRGQYTSGKGSSAVG
  Saccharomyces MCM4  QLFG.GTRKDFSHTGRG.KFRAEINILLCGDPGTSKSQLLQYVYNLVPRGQYTSGKGSSAVG
            Mus MCM4  QLFG.GTNKTFTKGGR...YRGDINILLCGDPSTSKSQILQYVHKITPRGVYTSGKGSSAVG
        S.pombe MCM4  QLF..GTNKSFHKGASP.RYRGDINILMCGDPSTSKSQILKYVHKIAPRGVYTSGKGSSAVG
      Drosophila dpa  LLQLFGGTKKKHATLGRQNFRSEIHLLLCGDPGTSKSQMLQYVFNLVPRSQYTSGRGSSAVG
  Saccharomyces MCM5  LLM..GGSKKILPDG..MRLRGDINVLLLGDPGTAKSQLLKFVEKVSPIAVYTSGKGSSAAG
        S.pombe MCM5  LLFS.GSKK.ILPDG..MRLRGDINVLLLGDPGTAKSQFLKFVERLAPIAVYTSGKGSSAAG
           Homo MCM5  LLF..GGSRKRLPDG..LTRRGDINLLMLGDPGTAKSQLLKFVEKCSPIGVYTSGKGSSAAG
     Drosophila MCM5  MLF..GVSRKRLPDGLCR..RGDINVLLLGDPGTAKSQLLKFVEKVAPIAVYTSGKGSSAAG
        Xenopus MCM5  LLF..GGSRKRLPDG..LTRRGDVNLLMLGDPGTAKSQLLKFVERCSPIGVYTSGKGSSAAG
      C.elegans MCM5  LLF..GGARKKLPDG..ITRRGDINVLLLGDPGTAKSQLLKFVEQVSPIGVYTSGKGSSAAG
            Mus MCM6  .MLF.GGVP.KTTGE.GTSLRGDINVCIVGDPSTAKSQFLKHVDEFSPRAVYTSGKASSAAG
        S.pombe MCM6  Q.LM.GGVH.KLTPE.GINLRGDLNICIVGDPSTSKSQFLKYVCNFLPRAIYTSGKASSAAG
           Homo MCM6  .MLF.GGVP.KTTGE.GTSLRGDINVCIVGDPSTAKSQFLKHVEEFSPRAVYTSGKASSAAG
    Arabidopsis MCM6  .LV..GAPH.RQLKD.GMKIRGDVHICLMGDPGVAKSQLLKHIINVAPRGVYTTGKGSSGVG
  Saccharomyces MCM6  QMLG.GVHK.STVEG..IKLRGDINICVVGDPSTSKSQFLKYVVGFAPRSVYTSGKASSAAG
      C.elegans MCM6  MLLG.GVAKKSRDEG..TSLRGDINVCLVGDPSTAKSQVLKAVEEFSPRAIYTSGKASSAAG
 C.elegans MCM6.like  .MLL.GGVAKKSRDE.GTSLRGDINVCLVGDPSTAKSQVLKAVEEFSPRAIYTSGKASSAAG
  Saccharomyces MCM7  LLVG.GVDK...RVGDGMKIRGDINVCLMGDPGVAKSQLLKAICKISPRGVYTTGKGSSGVG
        Xenopus MCM7  LLV..GGVD.NSPRG..MKIRGNINICLMGDPGVAKSQLLSYIDRLAPRSQYTTGRGSSGVG
            Mus MCM7  LLV..GGVD.QSPQG..MKIRGNIHICLMGDPGVAKSQLLSYIDRLAPRSQYTTGRGSSGVG
           Homo MCM7  LLV..GGVD.QSPRG..MKIRGNINICLMGDPGVAKSQLLSYIDRLAPRSQYTTGRGSSGVG
              Nosema  AMF..G...GSEKFAGESKVRGEIHVLIIGDPGLGKSXMLLSVCNILPKSTYVVWKFYDNRR
Methanococcus MJ0363  QLVSSG........TN.IDMRTSIHILMISDPGVGKSTLMESLIQKFPFVKKVYAVTSSGPG
Methanococcus MJ0961  QQIK.G...AFKFLPDGTPLRRDSHILLITDPGIGKSTMLRRIARLFPQNAYASVTTATGGG
Methanococcus MJ1489  QQIK.G...AFKFLPDGTPLRRDSHILLITDPGIGKSTMLRRIARLFPQNAYASVTTATGGG
          Sulfolobus  ALF.GGVPK....VLEDTRIRGDIHILIIGDPGTAKSQMLQFISRVAPRAVYTTGKGSTAAG
                      •••  •  •          ••••• ••• ••••••••• •      • ••••••• •••• •

  Saccharomyces MCM2  LTASVRKD..PITKEWTLEGGALVLADKGVCLIDEFDKMNDQDRTSI.EAMEQQSISISKAG
     Drosophila MCM2  LTAYVRRN..PVSREWTLEAGALVLADQGVCLIDEFDKMNDQDRTSIHEAMEQQSISISKAG
        Xenopus MCM2  LTAYVQRH..PVTKEWTLEAGALVFADRGVCLIDEFDKMNDQDRTSIHEAMEQQSISISKAG
  Saccharomyces MCM3  LTAAVTTD..RETGERRLEAGAMVLADRGVVCIDEFDKMTDVDRVAIHEVMEQQTVTIAKAG
       Triturus MCM3  LTAAVTTD..QETGERRLDVGAMVLADRGVVCIDEFDKMSDMDRTAIHEVMEQGRVTIAKAG
        Xenopus MCM3  LTAAVTTD..QETGERRLEAGAMVLADRGVVCIDEFDKMSDMDRTAIHEVMEQGRVTIAKAG
      Entamoeba MCM3  LTAAVVND..PDTNQRTLEAGAMVLADRGICCVDEFDKMSIEDRAAMHEVMEQQTVTVQKAG
        Xenopus MCM4  LTAYVMKD..PETRQLVLQTGALVLSDNGICCIDEFDKMNESTRSVLHEVMEQQTLSIAKAG
           Homo MCM4  LTAYVMKD..PETRQLVLQTGALVLSDNGICCIDEFDKMNESTRSVLHEVMEQQTLSIAKAG
  Saccharomyces MCM4  LTAYVMKD..PETRQLVLQTGALVLSDNGICCIDEFDKMNESTRSVLHEVMEQQTLSIAKAG
            Mus MCM4  LTAYITRD..VDTKQLVLESGALVLSDGGVCCIDEFDKMSDSTRSVLHEVMEQQTISIAKAG
        S.pombe MCM4  LTA.ITRD..QDTKQLVLESGALVLSDGGICCIDEFDKMSDATRSILHEVMEQQTVTVAKAG
      Drosophila dpa  LTAYVTKD..PETRQLVLQTGALVLADNGVCCIDEFDKMNDSTRSVLHEVMEQQTLSIAKAG
  Saccharomyces MCM5  LTASVQRD..PMTREFYLEGGAMVLADGGVVCIDEFDKMRDEDRVAIHEAMEQQTISIAKAG
        S.pombe MCM5  LTASIQRD..SVTREFYLEGGAMVLADGGIVCIDEFDKMRDEDRVAIHEAMEQQTISIRKAG
           Homo MCM5  LTASVMRD..PSSRNFIMEGGAWVLADGGVVCIDEFDKMREDDRVAIHEAMEQQTISIAKAG
     Drosophila MCM5  LTASVMKD..PQTRNFVVEGGAMVLADGGVVCIDEFDKMREDDRVAIHEAMEQQTISIAKAG
        Xenopus MCM5  LTASVMRD..PVSRNFIMEGGAMVLADGGVVCIDEFDKMREDDRVAIHEAMEQQTISIAKAG
      C.elegans MCM5  LTASVIRD..PQSRSFIMEGGAMVLADGGVVCIDEFDKMREDDRVAIHEAMEQQTISIAKAG
            Mus MCM6  LTAAVVRD..EESHEFVIEAGALMLADNGVCCIDEFDKMDMRDQVAIHEAMEQQTISITKAG
        S.pombe MCM6  LTAAVVKD..EETGDFTIEAGALMSADNGICAIDEFDKMDLSDQVAIHEAMEQQTISIAKAG
           Homo MCM6  LTAAVVRD...ESHEFVIEAGALMLADNGVCCIDEFDKMDVRDQVAIHEAMEQQTISITKAG
    Arabidopsis MCM6  LTAAVMRD..QVTNEMVLEGGALVLADMGICAIDEFDKMDESDRTAIHEVMEQQTVSIAKAG
  Saccharomyces MCM6  LTAAVVRD..EEGGDYTIEAGALMLADNGICCIDEFDKMDISDQVAIHEAMEQQTISIAKAG
      C.elegans MCM6  LTAAVVKD..EESFEFVIEAGALMLADNGVCCIDEFDKMDLKDQVAIHEAMEQQTISITKAG
 C.elegans MCM6.like  LTAAVVKD..EESFEFVIEAGALMLADNGVCCIDEFDKMDLKDQVAIHEAMEQQTISITKAG
  Saccharomyces MCM7  LTAAVMKD..PVTDEMILEGGALVLADNGICCIDEFDKMDESDRTAIHEVMEQQTISISKAG
        Xenopus MCM7  LTAAVMKD..PVTGEMTLEGGALVLADQGVCCIDEFDKMMDTDRTAIHEVMEQQTISIAKAG
            Mus MCM7  LTAAVLRD..SVSGELTLEGGALVLADQGVCCIDEFDKMAEADRTAIHEVMEQQTISIAKAG
           Homo MCM7  LTAAVLRD..SVSGELTLEGGALVLADQGVCCIDEFDKMAEADRTAIHEVMEQQTISIAKAG
              Nosema  LTIALTHD..SASGDFIAEAGALVLSDNGICCLDEFDKIENHRSLY..EVMEQQRVTVAKCG
Methanococcus MJ0363  LVGSVVREKAEFGDSWVLKAGVLTEADGGVVCIDEFSRNKEVYDYLLG.VMEQQKIEINKAG
Methanococcus MJ0961  LTAIVTREATEIGDGWVVKPGVFVRANEGTACIDELTVDKNVMKYIL.EAMESQTIHVNKGG
Methanococcus MJ1489  LTAIVTREATEIGDGWVVKPGVFVRANEGTACIDELTVDKNVMKYIL.EAMESQTIHVNKGG
          Sulfolobus  LTAAVVRE..KGTGEYYLEAGALVLADGGIAVIDEIDKMRDEDRVAIHEAME..........
                      ••• • •     • •  •• •••• •• ••••••• ••• •••• •• ••••
```

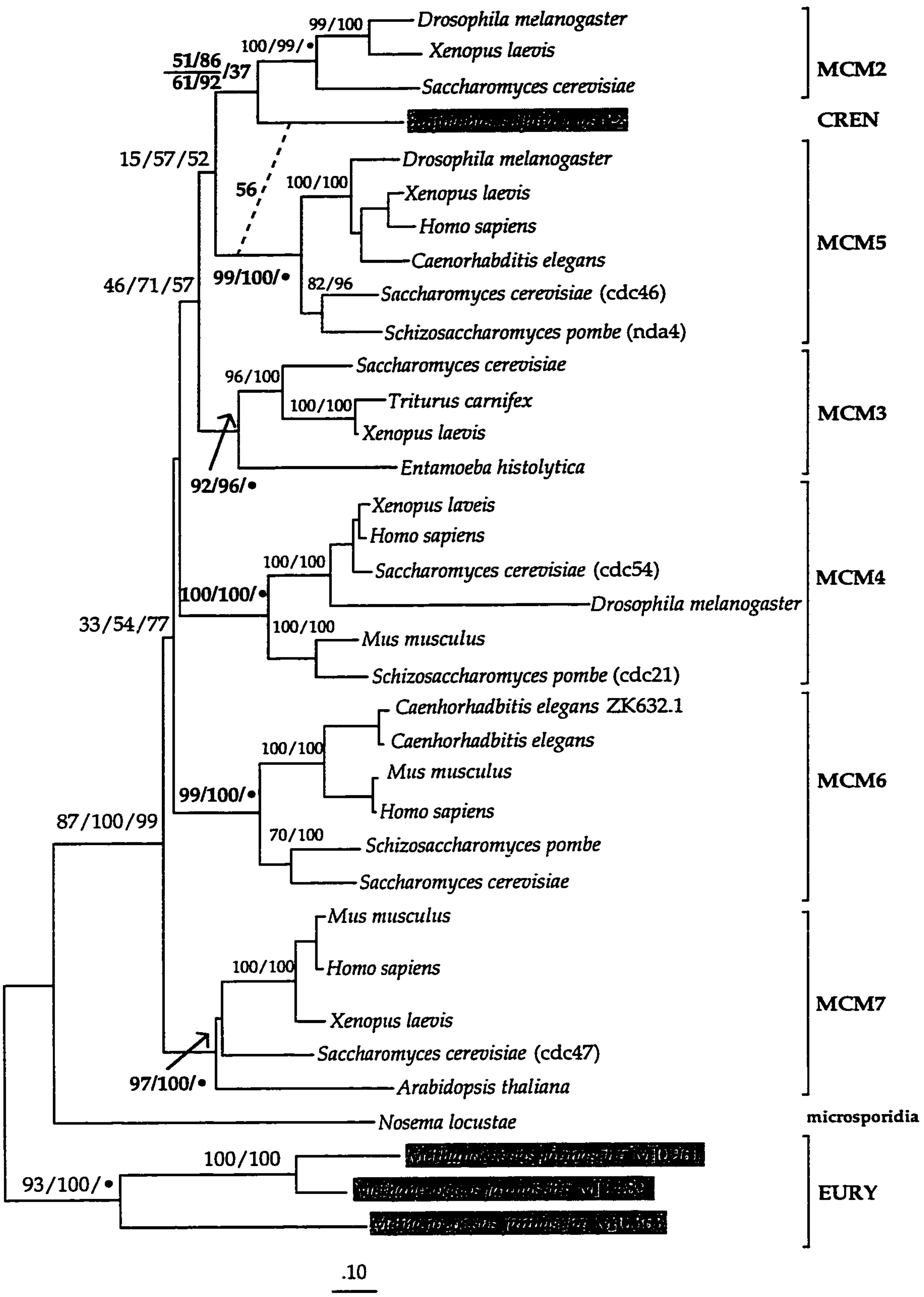
```
                                                                                 698
   Saccharomyces MCM2   IVT.TLQARCSIIAAANPNGGRYNSTLPLAQNVSLTEPILSRFDI.LVVRDLVDEEADERLATF
     Drosophila MCM2   IVT.SLQARCTVIAAANPIGGRYDPSMTFSENVNLSEPILSRFDVLCVVKDEFDPMQDQQLAKF
        Xenopus MCM2   IVT.SLQARCTVIAASNPIGGRYDPSLTFSENVDLTEPIVSRFDILCVVRDTVDPVQDEMLARF
   Saccharomyces MCM3   IHT.TLNARCSVIAAANPVFGQYDVNRDPHQNIALPDSLLSRFDLLFVVTDDINEIRDRSISEH
       Triturus MCM3   IQA.RLNARCSVLAAANPVYGRYDQYKTPMENIGLQDSLLSRFDLLFIVLDQMDADNREISDHV
        Xenopus MCM3   IQA.RLNARCSVLAAANPVYGRYDQYRTPMENIGLQDSLLSRFDLLFIVLDKMDADNDQEIADH
      Entamoeba MCM3   IHT.GIKCKMSILAAANPSNGNYDFKKSPMENLYFPESLLSRFDLIFIILDSSTEELDRKLSQH
        Xenopus MCM4   IIC.QLNARTSVLAAANPVESQWNPKKTTIENIQLPHTLLSRFDLIFLMLDPQDEAYDRRLAHH
           Homo MCM4   IIC.QLNARTSVLAAANPIESQWNPKKTTIENIQLPHTLLSRFDLIFLMLDPQDEAYDRRLAHH
   Saccharomyces MCM4   IIC.QLNARTSVLAAANPIESQWNPKKTTIENIQLPHTLLSRFDLIFLMLDPQDEAYDRRLAHH
            Mus MCM4   IIT.TLNARSSILASANPIGSRYNPNLPVTENIDLPPPLLSRFDLVYLVLDKVDEKNDRELAKH
        S.pombe MCM4   II..TLNARTSILASANPIGSKYNPDLPVTKNIDLPPTLLSRFDLVYLILDRVDETLDRKLANH
      Drosophila dpa   IIC.QLNARTSILAAANPAESQWNKRKNIIDNVQLPHTLLSRFDLIFLVLDPQDEIFDKRLASH
   Saccharomyces MCM5   ITT.VLNSRTSVLAAANPIYGRYDDLKSPGDNIDFQTTILSRFDMIFIVKDDHNEERDISIANH
        S.pombe MCM5   ITT.ILNSRTSVLAAANPIFGRYDDMKTPGENIDFQSTILSRFDMIFIVKDEHDETKDRNIARH
           Homo MCM5   ITT.TLNSRCSVLAAANSVFGRWDETKG.EDNIDFMPTILSRFDMIFIVKDEHNEERDVMLAKH
     Drosophila MCM5   ITT.TLNSRCSVLAAANSIFGRWDDTKG.EENIDFMPTILSRFDMIFIVKDIHDESRDITLAKH
        Xenopus MCM5   ITT.TLNSRCSVLAAANSVYGRWDDTKG.EENIDFMPTILSRFDMIFIVKDEHNEQRDMTLAKH
      C.elegans MCM5   ITT.TLNSRCSVLAAANSVYGRWDESRG.DDNIDFMPTILSRFDMIYIVKDTHDVLKDATLAKH
            Mus MCM6   VKA.TLNARTSILAAANPVSGHYDRSKSLKQNINLSAPIMSRFDLFFILVDECNEVTDYAIARR
        S.pombe MCM6   IQA.TLNARTSILAAANPIGGRYNRKTTLRNNINMSAPIMSRFDLFFVVLDECNESVDRHLAKH
           Homo MCM6   VKA.TLNARTSILAAANPISGHYDRSKSLKQNINLSAPIMSRFDLFFILVDECNEVTDYAIARR
    Arabidopsis MCM6   ITT.SLNARTAVLAAANPAWGRYDLRRTPAENINLPPALLSRFDLLWLILDRADMDSDLELAKH
   Saccharomyces MCM6   IHA.TLNARTSILAAANPVGGRYNRKLSLRGNLNMTAPIMSRFDLFFVILDDCNEKIDTELASH
      C.elegans MCM6   VKA.TLNARASILAAANPVNGRYDRSRPLKYNVQMSAPIMSRFDLFFVLVDECNEVTDYAIARR
 C.elegans MCM6-like   VKA.TLNARASILAAANPVNGRYDRSRPLKYNVQMSAPIMSRFDLFFVLVDECNEVTDYAIARR
   Saccharomyces MCM7   INT.TLNARTSILAAANPLYGRYNPRLSPLDNINLPAALLSRFDILFLMLDIPSRDDDEKLAEH
        Xenopus MCM7   IMT.TLNARCSILAAANPAYGRYNPKKTVEQNIQLPAALLSRFDVLWLIQDKPDRDNDLRLAQH
            Mus MCM7   ILT.TLNARCSILAAANPAYGRYNPRRSLEQNVQLPAALLSRFDLLWLIQDRPDRDNDLRLAQH
           Homo MCM7   ILT.TLNARCSILAAANPAYGRYNPRRSLEQNIQLPAALLSRFDLLWLIQDRPDRDNDLRLAQH
              Nosema   VVC.SVPTRATVVAATNPKFGHFKKDKSLRQNIGFDSALLSRFDLVFVLQDNLDESHNLDVSNY
Methanococcus MJ0363   VIDAVLPARVAILAACNPRFGRFNPDLTVWEQINLPKELLDRFDLIFVIKDKIDKKKDEDIADF
Methanococcus MJ0961   IN.VKLPARCAVLAACNPKRGRFDRNLTVIEQIDIPAPLLSRFDLIFPLMDKPNRKSDEEIAEH
Methanococcus MJ1489   IN.VKLPARCAVLAACNPRWGRFNPEVSVAEQINIPAPLLSRFDLIFPIRDVSDKDKDKDIAEY
          Sulfolobus   ................................................................
```

supported by a visual inspection of the MCM alignment as the *S. solfataricus* P2 sequence is more similar to eukaryotic paralogs than to the *M. jannaschii* paralogs. ML analysis agrees with distance and parsimony methods in placing the *S. solfataricus* MCM within eukaryotic paralogs to the exclusion of the *M. jannaschii* paralogs. However, ML does not find a grouping of the *S. solfataricus* P2 MCM with eukaryotic MCM2 paralogs (only 37% bootstrap support) but instead places the *S. solfataricus* P2 MCM sequence together with eukaryotic MCM5 paralogs (56% bootstrap). All three methods group the *S. solfataricus* P2, MCM2 and MCM5 paralogs together to the exclusion of all other sequences.

This result is not expected if one accepts archaebacterial monophyly; all archaebacterial MCM proteins should show a greater sequence similarity to one another than to any eukaryotic sequence and should branch together in phylogenetic analysis. If archaebacteria are not monophyletic, but paraphyletic with crenarchaeotes (ie. *S. solfataricus* P2) sharing a more recent common ancestor with eukaryotes than euryarchaeotes (ie. *M. jannaschii*), the recovered topology is not altogether unexpected. For instance, the gene duplications that gave rise to the six MCM paralogs found in eukaryotes would have occurred in crenarchaeotes, after their divergence from a common ancestor with euryarchaeotes. Crenarchaeotes would thus possess an ortholog of each of the six eukaryotic MCM proteins and, if sequences were available, would branch specifically with each MCM family, as does the *S. solfataricus* P2 MCM protein.

Alternatively, the common ancestor of euryarchaeotes and crenarchaeotes might have possessed two MCM paralogs, one similar to the *S. solfataricus* P2 sequence and the other to the *M. jannaschii* sequences. One of these paralogs (the *S. solfataricus* type) was lost along the euryarchaeote

**Figure 3.3** Phylogenetic analysis of archaebacterial (shaded) and eukaryotic MCM proteins. The tree shown is from PROTDIST analysis with all taxa and rooted with the three *M. jannaschii* paralogs. Nodes constrained for ML analysis are indicated by a •. Bootstrap values supporting nodes are indicated in the order parsimony/distance/ML. In some instances, alternative names for MCM proteins are included in parentheses (ie. CDC54). The dashed line joining the *S. solfataricus* P2 MCM sequence with eukaryotic MCM5 paralogs indicates that this topology was found in ML over the shown topology found by PROTDIST and parsimony analysis. For the node connecting the S. solfataricus P2 MCM sequence with eukaryotic MCM2 paralogs, parsimony and distance bootstrap values are present with the *Nosema locustae* sequence included in the analysis (above line), and with it removed (below line). The ML bootstrap value (36%) is given even though this topology was not the most likely.


99/100
Drosophila melanogaster
100/99/•
Xenopus laevis
51/86
61/92
/37
Saccharomyces cerevisiae
MCM2
Sulfolobus solfataricus P2
CREN
15/57/52
Drosophila melanogaster
56
100/100
Xenopus laevis
Homo sapiens
Caenorhabditis elegans
MCM5
46/71/57
99/100/•
82/96
Saccharomyces cerevisiae (cdc46)
Schizosaccharomyces pombe (nda4)
96/100
Saccharomyces cerevisiae
100/100
Triturus carnifex
Xenopus laevis
MCM3
Entamoeba histolytica
92/96/•
Xenopus laveis
Homo sapiens
100/100
Saccharomyces cerevisiae (cdc54)
MCM4
100/100/•
Drosophila melanogaster
33/54/77
100/100
Mus musculus
Schizosaccharomyces pombe (cdc21)
Caenhorhadbitis elegans ZK632.1
Caenhorhadbitis elegans
100/100
Mus musculus
MCM6
99/100/•
Homo sapiens
87/100/99
70/100
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Mus musculus
100/100
Homo sapiens
MCM7
Xenopus laevis
Saccharomyces cerevisiae (cdc47)
97/100/•
Arabidopsis thaliana
Nosema locustae
microsporidia
100/100
Methanococcus jannaschii MJ0961
93/100/•
Methanococcus jannaschii MJ1489
EURY
Methanococcus jannaschii MJ0363
.10

lineage after euryarchaeotes and crenarchaeotes diverged from a common ancestor (evidenced by the lack of this paralog in the complete genome sequence of *M. jannaschii*), while the *M. jannaschii*-like paralog has not yet been sequenced from *S. solfataricus* P2. The multiple paralogs in *M. jannaschii* can be best explained by a series of euryarchaeote-specific gene duplications. Both of these evolutionary scenarios can satisfactorily explain the MCM phylogeny if it were not for the branching position of the *N. locustae* MCM sequence as an outgroup to all eukaryotic sequences, including the *S. solfataricus* P2 sequence.

Although the *N. locustae* sequence is partial, it does cover the most conserved region of MCM proteins (amino acids 489-698 of *S. cerevisiae* MCM2). It is divergent in sequence from eukaryotic MCM proteins and is not specifically similar to any one eukaryotic paralog as judged by amino acid identity (not shown). Removal of the *N. locustae* sequence from phylogenetic analyses did not result in tree topologies that were different from when it was included (not shown). Parsimony and distance bootstrap values for the inclusion of the *S. solfataricus* P2 sequence as an ortholog of eukaryotic MCM2 proteins increased when the *N. locustae* sequence was removed (figure 3.3), as did values for nodes supporting other relationships. The position of *N. locustae* as an outgroup to the eukaryotic and *S. solfataricus* P2 paralogs was also found by ML analysis (99% support); it seems unlikely that the long branch length of the *N. locustae* sequence is adversely affecting phylogenetic analyses.

## RF-C phylogeny

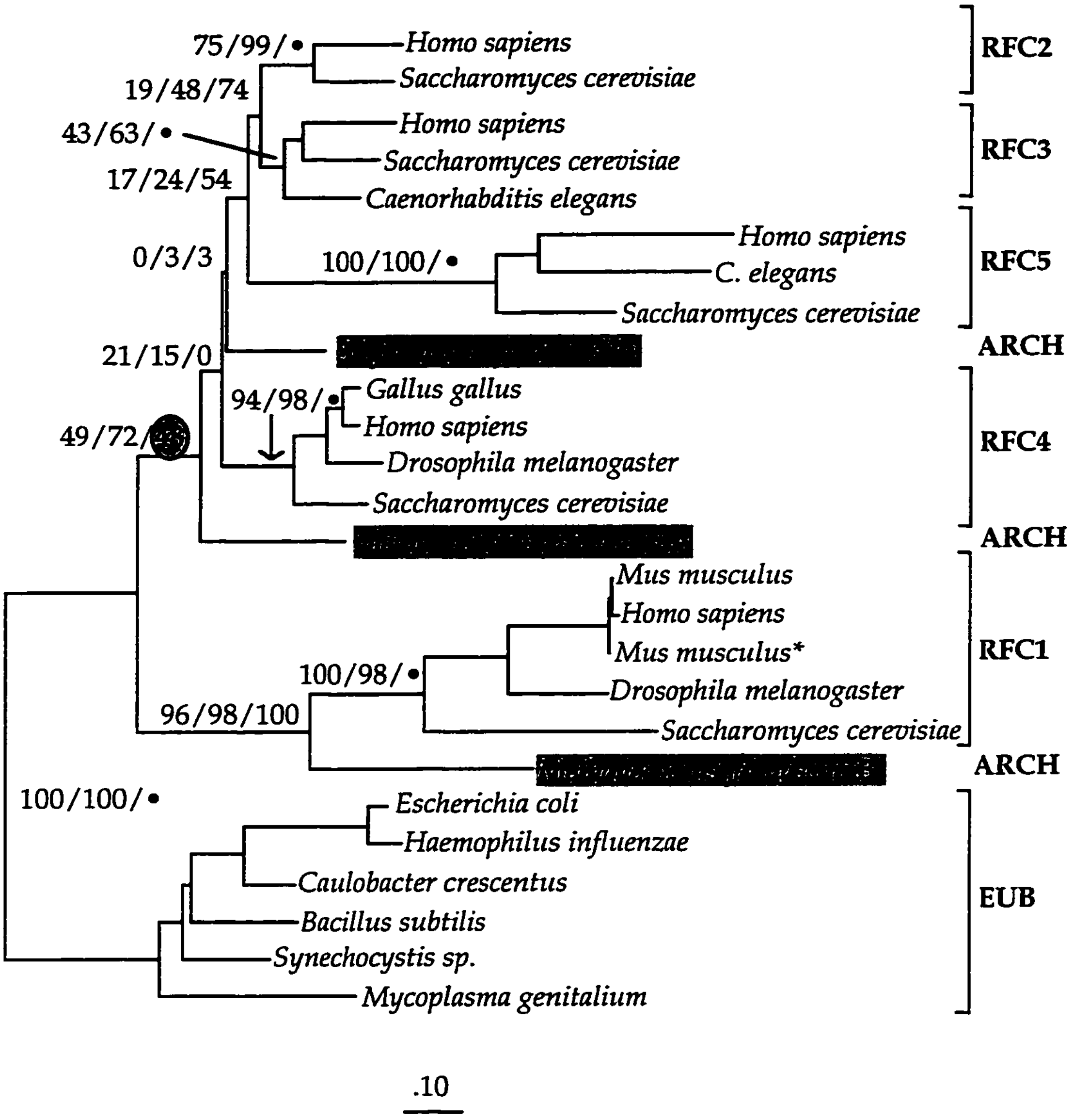
As for MCM proteins, phylogenetic analyses (figure 3.5) of the RF-C dataset (figure 3.4) resolved the five eukaryotic paralogs into monophyletic groups, consistent with these proteins having evolved by a series of gene duplications. However, the optimal trees found by parsimony, distance and ML were not similar in topology as all differed in the placement of the *S. solfataricus* P2 and *M. jannaschii* A paralogs, and in the relationship of the eukaryotic RF-C paralogs to each other. In all methods, the *M. jannaschii* B sequence consistently branched with eukaryotic RFC-1 sequences suggesting that this archaebacterial protein is an ortholog of eukaryotic RFC-1 proteins. It is unclear to which eukaryotic paralog the *S. solfataricus* P2 and *M. jannaschii* A sequences are most related. Parsimony and distance analyses are unresolved as bootstrap support for a specific branching of these paralogs with various eukaryotic paralogs is spread between a number of alternatives. ML analysis places the *S. solfataricus* and *M. jannaschii* A sequences as members of a clade consisting of the eukaryotic RFC-2,-3,-4 and -5 paralogs with 48% bootstrap support. If nodes with under 50% bootstrap support in parsimony and distance analyses are collapsed, a similar topology is obtained. These results are all consistent with the common ancestor of archaebacteria and eukaryotes possessing two RF-C paralogs. The RFC-2,-3,-4, and -5 paralogs of eukaryotes evolved by gene duplications from an ancestral sequence most closely resembling the *S. solfataricus* P2 and *M. jannaschii* A proteins. *M. jannaschii* B and eukaryotic RFC-1 sequences most closely resemble the second ancestral RF-C paralog of archaebacteria and eukaryotes.

**Figure 3.4** Amino acid alignment of conserved domains of eukaryotic, archaebacterial and eubacterial clamp loading proteins. Domains are numbered according to their linear order from N- to C-terminal (Cullman et al., 1995). Domain I is found only in eukaryotic RFC-1 paralogs and shares sequence similarity with eubacterial NAD-dependent DNA ligases (see figure 3.1). Methanococcus A and B refers to two RF-C homologs from *M. jannaschii* (MJ1422 and MJ0884 respectively). Amino acids that are identical or conserved between 95% of taxa included in the alignment are in bold type. Gaps introduced into the alignment are indicated by a period (.).

| | Domain II | Domain III | Domain IV |
|---|---|---|---|
| Bacillus | ALYRVFRPQRFEDVVGQEHIT | .KKFSHAYLFSGPRGTGKTSAAKIFA | D.VIEIDAASNNG.VDEIRDI |
| Escherichia | VLARKWRPQTFADVVGQEHVL | .GRIHHAYLFSGTRGVGKTSIARLLA | D.LIEIDAASRTK.VEDTRDL |
| Haemophilus | VLARKWRPKTFADVVGQEHII | .NRLHHAYLFSGTRGVGKTSIARLFA | D.LIEIDAASRTK.VEDTREL |
| Caulobacter | VLARKYRPRTFEDLIGQEAMV | .GRIAHAFMLTGVRGVGKTTTARLLA | D.VLELDAASRTK.VDEMREL |
| Mycoplasm | VFYQKYRPINFKQTLGQESIR | .DKLPNGYIFSGERGTGKTTFAKIAI | D.IVEIDAASKNG.INDIREL |
| Synechocystis | PLHHKYRPQTFADLVGQNAIE | .ERIVPAYLFTGPRGTGKTSSARIAL | D.VIEIDAASNTG.VDNIREI |
| Sulfolobus | LWAEKYRPKTLDDIVNQREIV | KEKNMPHLLFAGPPGTGKTTAALALV | EYFLELNA.SDERGIDVIRNK |
| Methanococcus A | PWVEKYRPKTLDDIVGQDEIV | VKKSMPHLLFSGPPGVGKTTAALCLA | DNFLELNA.SVSKDTPILVKI |
| Methanococcus B | SWVEKYRPKSLKDVAGHEKVK | EKLTPKPILLVGPPGCGKTTLAYALA | FEVIELNA.SDKRNSSAIKKV |
| Gallus RFC4 | PWVEKYRPLKLCEVVGNEDTV | KEGNVPNIIIAGPPGTGKTTSILCLA | DAVLELNA.SNDRGIDVVRNK |
| Homo RFC4 | PWVEKYRPVKLNEIVGNEDTM | REGNVPNIIIAGPPGTGKTTSILCLA | DAMLELNA.SNDRGIDVVRNK |
| Drosophila RFC4 | PWIEKYRPVKFKEIVGNEDTV | TQGNAPNIIIEGPPGVGKTTTIQCLA | EAVLELNA.SNERGIDVVRNK |
| Saccharomyces RFC4 | PWVEKYRPQVLSDIVGNKETI | KDGNMPHMIISGMPGIGKTTSVHCLA | DGVLELNA.SDDRGIDVVRNQ |
| Homo RFC2 | PWVEKYRPKCVDEVAFQEEVV | EGADLPNLLFYGPPGTGKTSTILAAA | LRVLELNA.SDERGIQVVREK |
| Saccharomyces RFC2 | PWVEKYRPKNLDEVTAQDHAV | KSANLPHMLFYGPPGTGKTSTILALT | SRILELNA.SDERGISIVREK |
| Homo RFC3 | PWVEKYRPQTLNDLISHQDIL | NEDRLPHLLLYGPPGTGKTSTILACA | SMVLELNA.SDDRGIDIIRGP |
| Saccharomyces RFC3 | PWVEKYRPETLDEVYGQNEVI | DEGKLPHLLFYGPPGTGKTSTIVALA | NMVLELNA.SDDRGIDVVRNQ |
| Caenorhabditis RFC3 | PWVEKYRPSKLDELVAHEQIV | ENRTLPHLLFYGPPGTGKTTTVLAAA | SMVLELNA.SDERGIDVVRNT |
| Mus RFC1 | LWVDKYKPASLKNIIGQQGDQ | DGSSFKAALLSGPPGVGKTTTASLVC | YSYVELNA.SDTRSKNSLKAV |
| Mus RFC1* | LWVDKYKPASLKNIIGQQGDQ | DGSSFKAALLSGPPGVGKTTTASLVC | YSYVELNA.SDTRSKNSLKAV |
| Homo RFC1 | LWVDKYKPTSLKTIIGQQGDQ | DGSSFKAALLSGPPGVGKTTTASLVC | YSYVELNA.SDTRSKSSLKAI |
| Drosophilia RFC1 | AWVDKHKPTSIKEIVGQAGAA | DGSFYKAALLSGPPGIGKTTTATLVV | FDAVEFNA.SDTRSKRLLKDE |
| Saccharomyces RFC1 | LWTVKYAPTNLQQVCGNKG.. | GSGVFRAAMLYGPPGIGKTTAAHLVA | YDILEQNA.SDVRSKTLLNAG |

| | Domain V | Domain VI | Domain VII | Domain VIII |
|---|---|---|---|---|
| Bacillus | VYIIDEVH | .MLSIGAFNALL | LTIISR...C | HGGMRDALSLL |
| Escherichia | VYLIDEVH | .MLSRHSFNALL | VTILSR...C | EGSLRDALSLT |
| Haemophilus | VYLIDEVH | .MLSRHSFNALL | VTILSR...C | QGSIRDSLSLT |
| Caulobacter | VYIIDEVH | .MLSTAAFNALL | KTILSR...C | EGSVRDGLSLL |
| Mycoplasm | VYILDEAH | .MLTTQSWGGLL | KTILSR...C | QGSLRDGLSLL |
| Synechocystis | VYVIDEVH | .MLSTAAFNALL | KTIISR...C | NGGLRDAESLL |
| Sulfolobus | VVLLDEAD | NM.TADAQQALR | EPIQSR...C | MGDMRKSINIL |
| Methanococcus A | IIFLDESD | AL.TADAQNALR | PPIQSR...C | EGDMRKAINVL |
| Methanococcus B | LIVLDEVD | GI.SGKEDAG.. | PSIRSLLPYV | AGDLRSAINDL |
| Gallus RFC4 | IIILDEAD | SM.TDGAQQALR | EPIQSR...C | QGDMRQALNNL |
| Homo RFC4 | IIILDEAD | SM.TDGAQQALR | EPIQSR...C | QGDMRQALNNL |
| Drosophila RFC4 | IVILDEAD | SM.TEGAQQALR | EPIQSR...C | QGDMRQGLNNL |
| Saccharomyces RFC4 | IVILDEAD | SM.TAGAQQALR | EPLQSR...C | EGDMRQAINNL |
| Homo RFC2 | IVILDEAD | SM.TSAAQAALR | EPLTSR...C | EGDLRKAITFL |
| Saccharomyces RFC2 | IIILDEAD | SM.TADAQSALR | DPLASR...C | AGDLRRGITLL |
| Homo RFC3 | LVILDEAD | AM.TQDAQNALR | PALQSR...C | SGDMRRALNIL |
| Saccharomyces RFC3 | LIILDEAD | AM.TNAAQNALR | PALLSR...C | NGDMRRVLNVL |
| Caenorhabditis RFC3 | LVILDEAD | AM.TKDAQNALR | PAIQSR...C | KGDMRTVINTL |
| Mus RFC1 | ALIMDEVD | GM.AGNEDRG.. | PKIRSLVHYC | NQDVRQVLHNL |
| Mus RFC1* | ALIMDEVD | GM.AGNEDRG.. | PKIRSLVHYC | NQDVRQVLHNL |
| Homo RFC1 | ALIMDEVD | GM.AGNEDRG.. | PKIRSLVHYC | NQDIRQVLHNL |
| Drosophilia RFC1 | VLIMDEVD | AM.AGNEDRG.. | PKIRSLVNYC | NNDIRQSINHI |
| Saccharomyces RFC1 | VIIMDEVD | GM.SG.GDRG.. | PKMRPFDRVC | RGDIRQVINLL |

**Figure 3.5** Phylogenetic analysis of eubacterial, archaebacterial (shaded) and eukaryotic clamp loading proteins. The tree shown was the optimal tree found by PROTDIST analysis. A similar topology was found by ML and parsimony analyses (five shortest trees, 620 steps, CI=0.681, HI=0.319) but differing in the arrangement of the eukaryotic RF-C paralogs to one another. Nodes constrained in ML analysis are indicated by a •. Bootstrap values supporting nodes are indicated in the order parsimony/distance/ML. The ML bootstrap value that is circled and shaded represents the value for a clade comprising the *M. jannaschii* A, *S. solfataricus* P2, and eukaryotic RFC-2,-3,-4,-5 paralogs, but the branching order of these paralogs was identical to that found by parsimony and distance methods. *Mus musculus* possesses two closely related RFC-1 sequences and these are differentiated from each other by a *.


75/99/•
Homo sapiens
Saccharomyces cerevisiae
RFC2
19/48/74
43/63/•
Homo sapiens
Saccharomyces cerevisiae
RFC3
17/24/54
Caenorhabditis elegans
0/3/3
100/100/•
Homo sapiens
C. elegans
RFC5
Saccharomyces cerevisiae
[redacted]
ARCH
21/15/0
94/98/•
Gallus gallus
Homo sapiens
49/72/
Drosophila melanogaster
RFC4
Saccharomyces cerevisiae
[redacted]
ARCH
Mus musculus
Homo sapiens
Mus musculus*
RFC1
100/98/•
Drosophila melanogaster
96/98/100
Saccharomyces cerevisiae
[redacted]
ARCH
100/100/•
Escherichia coli
Haemophilus influenzae
Caulobacter crescentus
EUB
Bacillus subtilis
Synechocystis sp.
Mycoplasma genitalium
.10

## Discussion

### Recurrent gene duplications in the evolution of archaebacterial and eukaryotic DNA replication proteins

The phylogenies of family B DNA polymerases (Edgell et al., 1997), MCM, and RF-C proteins of archaebacteria and eukaryotes are not congruent with respect to species relationships, but are congruent in showing that these gene families have each undergone a number of independent gene duplications during their evolution. The lack of functional data for archaebacterial homologs of eukaryotic replication proteins greatly hampers any attempts to understand the significance of these gene duplication events. For instance, two of the three family B DNA polymerases found in crenarchaeotes are extremely divergent in sequence, each exhibiting a number of non-conserved amino acid substitutions in functional domains relative to other archaebacterial and eukaryotic paralogs. Likewise, the MCM paralogs of *M. jannaschii* are divergent from each other, the *S. solfataricus* P2 paralog, and all six eukaryotic paralogs. In each of these instances, the degree of sequence divergence is sufficient to question whether or not the archaebacterial paralogs have functions analogous to those of eukaryotic paralogs.

Searching for protein families by computer-based methods is limited by the sensitivity of search algorithms to detect similarity between query and database sequences, and it is entirely possible that all paralogous proteins will not be found. Even if functional studies do identify ORFs missed by computer-based methods as members of a protein family, it is likely that these proteins have accumulated non-synonymous substitutions such that amino acid alignments with other paralogs are not significant.

What is the function of archaebacterial MCM proteins?

Genome sequencing projects are continually uncovering examples of proteins in archaebacterial genomes that were thought to be "eukaryote-specific"; such is the case with archaebacterial homologs of MCM proteins. The presence of these proteins is particularly confusing given their function in eukaryotes as limiting initiation of replication to once per cell cycle (Kearsy et al., 1996). It is unlikely that the *M. jannaschii* and *S. solfataricus* P2 paralogs function in a similar manner as archaebacteria are not thought to possess a eukaryote-style cell cycle, nor the elaborate check point controls associated with one.

Recent experimental evidence has shown that *S. cerevisiae* MCM proteins are deposited onto chromatin at the start of G1 by the CDC6 protein at levels that exceed the number of active replication origins (Donovan et al., 1997). This binding is thought to be involved in limiting replication initiation to once per cell cycle and in preventing re-initiation from origins that have already fired. It is possible that archaebacterial MCM homologs also bind DNA and, in doing so, function to control replication initiation. However, no homolog of the CDC6 protein is present in the genome of *M. jannaschii* (Bult et al., 1996), nor in the partially completed genome of *S. solfataricus* P2 (not shown), so the "loading" of archaebacterial MCM proteins onto DNA must proceed by a different pathway than that of *S. cerevisiae.* In addition, there is substantial biochemical and genetic evidence from studies in *S. cerevisiae* pointing to a physical interaction of CDC6 and the Origin Recognition Complex (ORC) proteins (see chapter IV for a discussion of ORC proteins); this interaction is thought to be crucial for the control of initiation of replication (Liang et al., 1995; Cocker et al., 1996; Kearsey et al., 1996). As no homologs of eukaryotic ORC proteins, nor of the eubacterial origin-binding protein DnaA, are present in the *M. jannaschii* genome sequence, archaebacterial MCM proteins must control replication initiation though interactions with an as-yet-unidentified set of origin-binding protein(s) (if indeed that is the function of archaebacterial MCM proteins).

Phylogenetic analysis of MCM proteins indicated that the *S. solfataricus* P2 sequence is likely an ortholog of MCM2/5 proteins, while the three *M. jannaschii* MCM homologs do not appear specifically related to any one eukaryotic MCM protein (figure 3.3). This phylogeny is consistent with independent gene duplications in the history of this gene family: one event giving rise to the multiple paralogs of *M. jannaschii,* and the other event(s) giving rise to the multiple paralogs of eukaryotes. However, the long branch lengths of the *M. jannaschii* paralogs suggest that the duplication events did not occur recently, as each paralog has undergone a number of non-conserved amino acid substitutions relative to each other and to eukaryotic paralogs. These non-conserved amino acid substitutions suggest that these divergent paralogs of *M. jannaschii* may have been recruited for some function other than replication initiation. In addition, the large extra-chromosomal element of *M. jannaschii* possesses an ORF with sequence similarity to the chromosomally-encoded MCM homologs (Bult et al., 1996; this sequence was not included in phylogenetic analyses because it is very divergent); whether or not this paralog has some function(s) in controlling plasmid replication remains to be determined.

### A DNA-ligase domain was added to RFC-1 early in the evolution of eukaryotes

Although the similarity of the amino-terminal extension of eukaryotic RFC-1 and eubacterial ATP-dependent DNA ligases has been well documented by other researchers, none has presented an explanation as to how eukaryotic RFC-1 acquired this DNA ligase-like domain (Burbelo et al., 1993; Cullman et al., 1995). The fact that no other eukaryotic, archaebacterial, or eubacterial RF-C homolog possesses the DNA ligase-like domain suggests that this domain was not an ancestral feature of clamp loading proteins. Rather, it is more likely that this domain was added to the eukaryotic RFC-1 protein after the gene duplication event that gave rise to this paralog. If phylogenetic analysis suggesting that the *M. jannaschii* B sequence is an ortholog of eukaryotic RFC-1 proteins is accurate (figure 3.5), the ligase domain must have been added after eukaryotes and archaebacteria diverged from a common ancestor.

Since the origin of the DNA ligase domain is obviously eubacterial (Cullman et al., 1996; see above), it is tempting to speculate that the source ligase domain was the endosymbiont that in extant eukaryotes is recognizable as the mitochondrion (or chloroplast). We know that much of the genome of the mitochondrial and chloroplast endosymbiont was, and still is being, transferred to nuclear genomes (reviewed in Gray, 1992). No known organellar genomes encode DNA ligases, either ATP- or NAD-dependent (Palmer, 1997), suggesting that this gene was indeed transferred to the nuclear genome early in the evolution of eukaryotes. A fortuitous recombination event could have fused the DNA-ligase to the amino-terminal region of RFC-1, but not to other RF-C paralogs. A test of this theory would be to clone and sequence RFC-1 orthologs from a variety of amitochondrial eukaryotes. For instance, the RFC-1 ortholog of *G. lamblia* might not possess the DNA ligase-like domain if the source of the ligase gene was the mitochondrial symbiont, as diplomonads such as *G. lamblia* are perhaps the only remaining archezoan as initially proposed by Cavalier-Smith (Cavalier-Smith, 1987 a, b). Although it has been shown that *G. lamblia* possess a eubacterial-like triosphosphate isomerse (Keeling and Doolittle, 1997), it is not clear if the source of this gene is in fact the same lineage of α-proteobacteria that became the mitochondrial symbiont. However, the RFC-1 orthologs of secondarily amitochondrial protists, such as *T. vaginalis* (Bui et al., 1996; Germont et al., 1996; Horner et al., 1996; Roger et al., 1996) and *Entamoeba histolytica* (Clark and Roger, 1995), might possibly possess the DNA ligase domain since they once harboured mitochondria. It is also possible that eukaryotes acquired the DNA ligase domain from a eubacterial lineage unrelated to the α-proteobacterial lineage.

Biochemical characterization of RF-C proteins has not resulted in the isolation of a ligase activity (Burbello et al., 1993; Cullman et al., 1995), so it appears that the DNA ligase domain has been co-opted for some biochemical function other than ligation. Indeed, mutational studies have identified the DNA ligase region of RFC-1 as functioning in DNA binding (Burbelo et al., 1993). The presence of multiple RFC-1 paralogs in animal genomes (table 3.1) would seem to suggest that the DNA-binding activity of the ligase domain has been recruited independently for sequence-specific binding as all function as transcriptional activators (see for example McGehee and Habener, 1995). Unfortunately, there are no mutational studies on eubacterial NAD-dependent DNA ligases to confirm whether or not this region of the protein also functions in DNA binding.

# IV. Comparisons of eubacterial, archaebacterial, and eukaryotic DNA replication proteins

## Introduction

Studies on the mechanism(s) of DNA replication in eubacteria and eukaryotes have led to the identification of many similar (analogous) biochemical activities (Kornberg and Baker, 1992). Cloning, sequencing, and in-depth biochemical characterization of proteins performing analogous activities in eubacteria and eukaryotes has led to an appreciation of the sophisticated molecular mechanisms responsible for ensuring accurate and processive DNA replication. It would not be unreasonable to expect proteins performing analogous replication functions in eubacteria and eukaryotes to be homologous, having diverged from a common set of replication proteins present in the last common ancestor of life. This is certainly true for proteins involved in other genetic processes such as transcription and translation (see reviews by Amils et al., 1993; Zillig et al., 1993). This assumption can be tested by comparisons of the amino acid sequences of replication proteins that perform analogous functions. However, genetic screens in *E. coli* and *S. cerevisiae* for abnormal DNA replication phenotypes have resulted in the identification of a vast number of proteins (Kornberg and Baker, 1992); often the involvement of these proteins in replication is indirect. As such, I have limited comparisons to proteins that are localized at the replication fork and that are involved in origin recognition and initiation of replication (table 4.1).

**Table 4.1** Summary of replication proteins of eubacteria and eukaryotes that perform analogous functions used in sequence comparisons. Based loosely on Stillman, 1994. All eubacterial replication proteins are from *E. coli,* except the catalytic subunit of DNA polymerase III (polC), which is from *Bacillus subtilus*. The *B. subtilus* polC polypeptide is encoded by a single gene, whereas the *E. coli* polypeptide is encoded by two separate genes.

| Protein (abbreviaton) | Function of protein |
|---|---|
| Eubacterial | |
| *E. coli* DnaA (EcDnaA) | origin-binding protein |
| *E. coli* SSB (EcSSB) | single-stranded DNA binding protein |
| *E. coli* DnaB (EcDnaB) | replication fork helicase (5′-3′) |
| *E. coli* PriA (EcPriA) | unwinding of origin (3′-5′) |
| *E. coli* DnaG (EcDnaG) | synthesis of primer |
| *E. coli* DnaX (EcDnaX) | clamp loading protein (γ subunit) |
| *E. coli* HolB (EcHolB) | clamp loading protein (δ′ subunit) |
| *E. coli* DnaN (EcDnaN) | processivity factor ('sliding clamp') |
| *E. coli* PolA (EcPolA) | removal of primers |
| *Bacillus subtilus* PolC (BsPolC) | α-subunit of holoenzyme complex; polymerization |
| *E. coli* DNA ligase (EcLigase) | ligation of Okazaki fragments (NAD-dependent) |
| Eukaryotic | |
| *S. cerevisiae* Origin Recognition Complex (ORC) proteins 1-6 (ScORC1-6) | origin binding proteins |
| *S. cerevisiae* Replication Protein A, subunits 1-3 (ScRPA1-3) | single-stranded binding proteins |
| *S. cerevisiae* DNA polymerase α (ScDnapα) | synthesis of primer |
| *S. cerevisiae* Dna2 | helicase (3′-5′) |
| *S. cerevisiae* Replication Factor-C, subunits 1-5 (ScRFC1-5) | clamp loading proteins |
| *S. cerevisiae* Proliferating Cell Nuclear Antigen (ScPCNA) | processivity factor ('sliding clamp') |
| *S. cerevisiae* DNA polymerase δ, ε (ScDnapδ, ScDnapε) | synthesis of leading and lagging strands |
| *S. cerevisiae* DNA ligase (ScLigase) | ligation of Okazaki fragments (ATP-dependent) |
| *S. cerevisiae* FEN-1/Rad2 (ScFen-1) | removal of primers |

## Results and Discussion

### Many eubacterial and eukaryotic replication proteins are not similar at the amino acid level, yet perform analogous function(s)

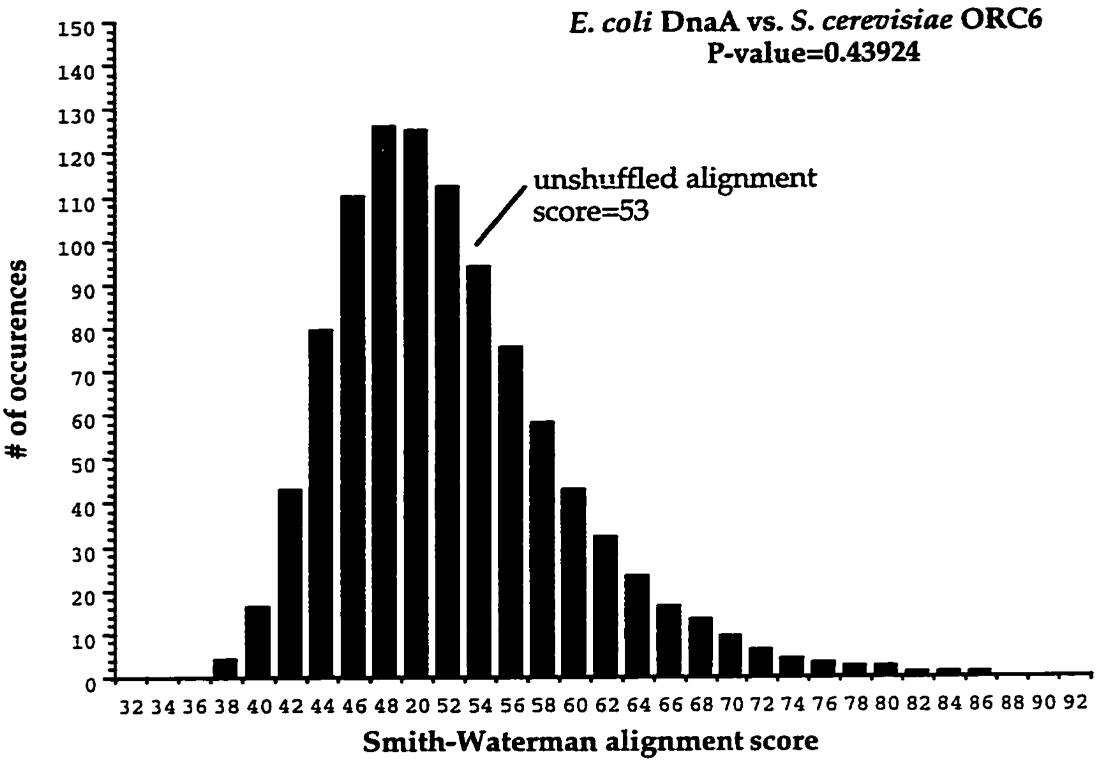
To determine whether eubacterial and eukaryotic replication proteins performing analogous replication functions are significantly similar at the amino acid level, I compared pairwise alignment scores of unshuffled alignments against scores of shuffled pairwise alignments of the same two proteins (table 4.2). This method can help resolve two problems commonly encountered when aligning amino acid sequences: length and biased composition (ie. strings of similar or identical amino acids; Doolittle, 1986). Proteins with simple repetitive sequences will artificially increase alignment scores as the chances of aligning similar or identical amino acids is greater than for proteins without a biased amino acid composition. For example, the eubacterial single-stranded DNA binding proteins, SSB, has a high number of glycine, proline and glutamine residues relative to other *E. coli* replication proteins and other eubacterial SSB-proteins.

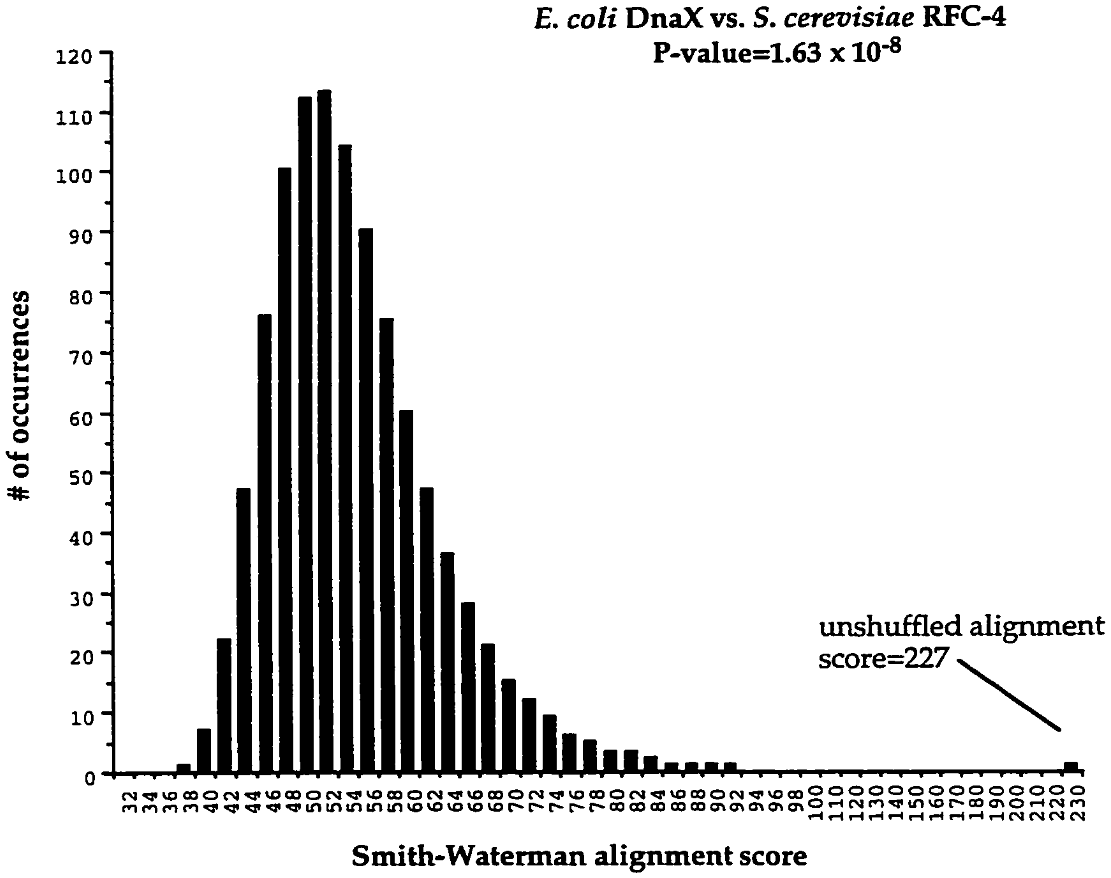
The majority of alignments of replication proteins of eukaryotes and eubacteria are not significantly better at the 95% confidence level than would be expected from 1000 shuffled alignments (table 4.2). Alignments of proteins involved in clamp loading (the DnaX and HolB proteins in *E. coli* and the Replication Factor-C proteins in eukaryotes) were significantly better than shuffled alignments (figure 4.1), but this result is expected based on work presented in chapter III and by other researchers (O'Donnell et al., 1993; Cullmann et al., 1995).

**Table 4.2** Comparisons of amino acid sequences of eubacterial and eukaryotic replication proteins performing analogous replication functions. Each pairwise comparison is listed along with the length (in amino acids) of the two proteins (see table 4.1 for list of abbreviations). Optimal score refers to the best alignment score of the unshuffled protein sequences obtained by the Smith-Waterman algorithm. Range refers to the range of alignment scores of 1000 shuffled pairwise alignments. P-value refers to the probability that the unshuffled alignment will fall outside of the range of shuffled alignment scores.

| Comparison | Optimal score | Range | P-value |
|---|---|---|---|
| EcDnaA (467aa) vs. ScOrc1 (913aa) | 72 | 43-125 | 0.160 |
| ScOrc2 (621aa) | 53 | 44-84 | 0.633 |
| ScOrc3 (614aa) | 56 | 43-91 | 0.596 |
| ScOrc4 (529aa) | 63 | 41-81 | 0.186 |
| ScOrc5 (279aa) | 89 | 39-93 | 0.00162 |
| ScOrc6 (435aa) | 53 | 37-101 | 0.449 |
| ScOrc1 vs. ScOrc2 | 95 | 51-126 | 0.0401 |
| ScOrc3 | 76 | 48-123 | 0.186 |
| ScOrc4 | 87 | 47-108 | 0.186 |
| ScOrc5 | 59 | 46-93 | 0.549 |
| ScOrc6 | 108 | 46-111 | 0.00203 |
| EcDnaX (643aa) vs. ScRFC-1 (862aa) | 105 | 46-160 | 0.00203 |
| ScRFC1-lig.(598aa) | 43 | 28-92 | 0.465 |
| ScRFC-2 (354aa) | 263 | 27-88 | 2.13 x $10^{-5}$ |
| ScRFC-3 (340aa) | 190 | 26-84 | 4.85 x $10^{-7}$ |
| ScRFC-4 (323aa) | 166 | 25-78 | 1.57 x $10^{-8}$ |
| ScRFC-5 (354aa) | 45 | 26-79 | 0.315 |
| EcDnaX vs. EcHolB (334aa) | 271 | 28-101 | 8.89 x $10^{-14}$ |
| EcHolB vs. ScRFC-1 | 39 | 27-83 | 0.573 |
| ScRFC1-lig. | 40 | 24-80 | 0.466 |
| ScRFC-2 | 48 | 22-91 | 0.108 |
| ScRFC-3 | 127 | 24-71 | 1.74 x $10^{-7}$ |
| ScRFC-4 | 57 | 21-85 | 0.038 |
| ScRFC-5 | 91 | 24-75 | 6.58 x $10^{-5}$ |
| ScRFC-1 vs. ScRFC-2 | 265 | 29-104 | 1.11 x $10^{-14}$ |
| ScRFC-3 | 111 | 28-94 | 5.40 x $10^{-5}$ |
| ScRFC-4 | 259 | 27-96 | 4.38 x $10^{-14}$ |
| ScRFC-5 | 59 | 28-89 | 0.0715 |
| EcDnaN (367aa) vs. ScPCNA (341aa) | 58 | 34-85 | 0.119 |
| EcDnaB (471aa) vs. ScDna2 (1532aa) | 81 | 49-111 | 0.115 |
| EcPriA (732aa) vs. ScDna2 | 46 | 31-85 | 0.441 |
| EcPriA vs. EcDnaB | 56 | 42-96 | 0.533 |
| EcDnaG (582aa) vs. ScDnapα (1469aa) | 71 | 46-124 | 0.214 |
| ScDnapδ (1097aa) vs. BsPolC (1437aa) | 63 | 33-97 | 0.0995 |
| ScDnapε (2222aa) vs. BsPolC | 54 | 34-105 | 0.425 |
| EcPolA (928aa) vs. BsPolC | 54 | 32-117 | 0.641 |
| EcLigase (661aa) vs. ScLigase (756aa) | 37 | 29-86 | 0.852 |
| EcSSB (178aa) vs. ScRPA-1 (621aa) | 42 | 24-73 | 0.277 |
| ScRPA-2 (273aa) | 54 | 21-81 | 0.051 |
| ScRPA-3 (123aa) | 35 | 17-57 | 0.182 |
| ScRPA-1 vs. ScRPA-2 | 40 | 25-96 | 0.589 |
| ScRPA-2 vs. ScRPA-3 | 30 | 20-85 | 0.708 |
| ScRPA-1 vs. ScRPA-3 | 36 | 22-78 | 0.543 |
| EcPolA vs. ScFen-1 (382 aa) | 58 | 28-91 | 0.104 |

**Figure 4.1** Schematic representation of shuffled versus unshuffled alignment scores for two pairwise comparisons; *E. coli* DnaA vs. *S. cerevisiae* ORC6 and *E. coli* DnaX vs. *S. cerevisiae* RFC-4. Graph plots the number of occurrences of the shuffled alignments versus the Smith-Waterman alignment score. The unshuffled alignment score is indicated with a line. The P-value refers to the probability that the unshuffled alignment score will fall outside of the range of shuffled scores.


*E. coli* DnaA vs. *S. cerevisiae* ORC6
P-value=0.43924
unshuffled alignment
score=53
# of occurences
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
32 34 36 38 40 42 44 46 48 20 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92
Smith-Waterman alignment score


*E. coli* DnaX vs. *S. cerevisiae* RFC-4
P-value=1.63 x 10$^{-8}$
unshuffled alignment
score=227
# of occurrences
0 10 20 30 40 50 60 70 80 90 100 110 120
32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100 110 120 130 140 150 160 170 180 190 200 210 220 230
Smith-Waterman alignment score

### 1. Origin-binding proteins

One of the major differences between eukaryotic and eubacterial replication systems is the presence of many versus a single origin of replication. However, in both eukaryotic and eubacterial replication systems, an early step in the initiation of replication is the binding of the origin by an origin recognition protein. In *E. coli*, this function is performed by DnaA (Kornberg and Baker, 1992) and in eukaryotes, by a heteromeric complex of six Origin Recognition Complex (ORC) proteins (reviewed in Stillman, 1996; Diffley, 1997). The six *S. cerevisiae* ORC proteins only show a low level of similarity to each other at the amino acid level, and only two (ORC1 and ORC5) are significantly similar to the *E. coli* DnaA protein at the 90% confidence level (1 in 10 shuffled alignments will have the same alignment score as the non-shuffled alignment; table 4.2). This similarity might not be reflective of common ancestry, but of functional convergence; all three of these proteins bind ATP and all three possess sequences resembling nucleotide- and $Mg^{2+}$-binding pockets found in a wide variety of proteins (Walker et al., 1982; Bell et al., 1995; Koonin, 1997). In addition to little primary amino acid sequence similarity, both proteins exhibit functional differences; it is not clear that they perform analogous functions.

*In vivo* footprinting experiments showed that ORC proteins are constitutively bound to the ARS (autonomously replicating sequence) of yeast throughout the cell cycle and do not dissociate from the origin during initiation of replication, nor after replication has been initiated (Diffley et al., 1994). However, the size of the footprint increases at the onset of DNA replication in $G_1$ indicating a possible interaction between the CDC6 protein and ORC (Cocker et al., 1996). Both genetic and biochemical evidence suggests that the CDC6 protein interacts with ORC (Cocker et al., 1996; Heichman, 1996;

Piatti et al., 1996). It is possible that CDC6 and ORC form a complex at ARS sites and that modification of CDC6 protein, possibly by phosphorylation, positively promotes the initiation of replication (Heichman, 1996). In contrast, the *E. coli* DnaA protein cannot bind to the *E. coli* origin, *ori*C, as *ori*C is sequestered for much of the cell cycle by a membrane-associated protein, SeqA (Lu et al., 1994). In *E. coli,* newly replicated DNA is hemimethylated (methylated only on one strand) at GATC sites; SeqA binds to hemimethylated DNA around the *ori*C region, and in the promoter region of the *dna*A gene thus preventing its expression (Slater et al., 1995). *ori*C remains sequestered for approximately one-third of the *E. coli* cell cycle, at which point it is fully methylated on both strands and ready for another round of replication initiation (Campbell and Kleckner, 1990).

Although both DnaA and ORC require ATP for DNA-binding activity, key differences exist in the utilization of bound ATP molecules. Bell and colleagues have shown that binding of ORC to ARS consensus sequences is promoted by ATP and that one ORC subunit, ORC1, binds ATP strongly in the presence of ARS-containing DNA, but weakly in its absence, and does not hydrolyze ATP once bound to ARS-containing DNA (Klemm et al., 1997). This is unlike the utilization of ATP by the *E. coli* DnaA protein; it also requires ATP to be in an active form, but hydrolysis of DnaA-bound ATP negatively regulates replication initiation (Mizushima et al., 1997). The exact mechanism through which this negative regulation occurs is unknown but it is thought to involve a specific inactivation factor of DnaA, IdaA (Katayama and Crooke, 1995), which enhances the hydrolysis of ATP by DnaA (Mizushima et al., 1997).

2. Single-stranded binding proteins

Subsequent to origin binding by an origin recognition protein, one of the next steps in initiation of replication is binding of the origin region by a single-stranded-binding protein (Kornberg and Baker, 1992). In *E. coli,* this function is performed by SSB (single-stranded-binding protein) and in eukaryotes, by RP-A (Replication Protein-A). RP-A is a heterotrimeric complex (in *Homo sapiens* 70-, 34- and 11-kDa) with the ssDNA-binding activity residing in the largest and second largest subunits. All of these proteins have been described as homologs (Philipova et al., 1996), but published amino acid alignments are not compelling. Indeed, only one non-shuffled alignment, *E. coli* SSB versus *S. cerevisiae* RPA-2, is significantly better than shuffled alignments at the 95% confidence level (table 4.1). The *E. coli* SSB protein is 177 amino acids in length yet of those residues, only 19% are similar, not identical, to the (longer) eukaryotic second and third largest subunits (based on published alignments; Philipova et al., 1996). Furthermore, multiple gaps (indicating many independent insertion and deletion events in the evolution of these genes) must be introduced to align the largest RP-A subunit with the *E. coli* SSB protein in regions of the proteins thought to be essential for SSB activity. Three aromatic amino acids (Trp at position 40, Trp at position 54 and Phe at position 60) are known to be critical for SSB binding in the *E. coli* protein (Curth et al., 1993; Overman et al., 1988). To align similar, not identical, aromatic amino acid residues of the largest RP-A subunit with the *E. coli* SSB protein, a gap corresponding to 14 amino acids must be introduced between Trp-40 and Trp-54. Futhermore, the alignment of the amino acid sequences surrounding the SSB domain is also no better than random and contains many gaps. Based solely on these amino acid alignments, it is difficult to convincingly call the eukaryotic and eubacterial SSB proteins homologs.

Recently, the crystal structures of the *E. coli* SSB protein and the SSB-binding domain of the largest RP-A subunit bound to DNA were solved (Bochkarev et al., 1997; Raghunathan et al., 1997). Two additional structures of replication-associated SSB proteins have also been determined, that of the gene V protein of bacteriophage f1 (Skinner et al., 1994) and the gp32 protein of bacteriophage T4 (Shamoo et al., 1995). All of these proteins share a common structural fold, the OB (oligonucleotide/oligosaccharide binding)-fold, characterized by a five-stranded β-sheet coiled to form a closed β-barrel, which is in turn capped by an α-helix located between the third and fourth strands (Murzin, 1993). This OB-fold is also found in proteins with functions unrelated to DNA replication: staphylococcal nuclease, the anticodon-binding domain of aspartyl-tRNA synthetase, and the B-subunits of heat-labile enterotoxin and verotoxin-1. In all these proteins, the amino acid residues that form the OB-fold cannot be aligned with each other with any degree of confidence (Murzin, 1993).

Interestingly, the structure of the RP-A SSB domain is more similar to that of *S. cerevisiae* aspartyl-tRNA synthetase bound to tRNA than to any other replication-associated SSB protein (Bochkarev et al., 1997). Futhermore, if the structures of the *E. coli* SBB and RP-A largest subunit bound to ssDNA are superimposed, Trp-54 of *E. coli* SSB, a residue known to be involved in SSB-binding activity, is more than 14-Å away from ssDNA (Raghunathan et al., 1997). These findings add further evidence against eubacterial and eukaryotic replication-associated SSB proteins evolving by descent from a common ancestral DNA-binding protein; other explanations are equally likely. For instance, the ability to bind single-stranded nucleic acids might have evolved independently many times and been recruited into replication functions. Alternatively, the OB-fold common to proteins that bind single-stranded nucleic acids or oligosaccharides could have been shuffled between proteins that originally lacked the ability to bind single-stranded nucleic acids.

## 3. Processivity factors

DNA-dependent DNA polymerases are not intrinsically processive; they will dissociate from DNA templates unless prevented from doing so by protein factors (Kornberg and Baker, 1992). Such proteins are called processivity factors or sliding clamps. The eubacterial sliding clamp is coded for by the *dna*N gene (the protein product is often called Polβ), and in eukaryotes by Proliferating Cell Nuclear Antigen (PCNA). The two proteins have very similar biochemistry and can be considered functionally analogous (Kelman and O'Donnell, 1995); both PCNA and Polβ interact with the clamp loading complex (see chapter III) and both interact with the replicative DNA polymerase (DNA pol δ/ε in eukaryotes and the polIII holoenzyme complex in eubacteria). One structural difference between PCNA and Polβ is that PCNA forms a trimer whereas Polβ forms a dimer. The functional significance, if any, of this difference is not known. Amino acid alignments of these proteins are not convincing (table 4.2; Kelman and O'Donnell, 1995), but the crystal structures of both the eubacterial and eukaryotic proteins are almost identical and can be superimposed (Kelman and O'Donnell, 1995). Given that structural similarity extends over the entire length of the proteins rather than being confined to a particular functional domain, it is likely that these two proteins did indeed evolve from an ancestral 'sliding clamp' and have since diverged in sequence.

4. DNA-dependent DNA polymerases

The replicative DNA polymerases of eubacteria (family A and C) and eukaryotes (family B) not only exhibit functional and biochemical differences (Kornberg and Baker, 1992), but are also not alignable with each other at the amino acid level (see tables 4.2); it is difficult to conclude from sequence comparisons that DNA-dependent DNA polymerases of eubacteria, archaebacteria, and eukaryotes evolved from a single ancestral DNA polymerase. However, other researchers have endeavored to find signature sequences common to catalytic regions of DNA polymerases in attempts to demonstrate homology (see for example Bernad et al., 1989; Blanco et al., 1991; Joyce and Steitz, 1994). These signature sequences are limited to three extremely short regions, A, B, and C, that mutational studies have identified as functioning in metal-binding (reviewed in Joyce and Steitz, 1994). Region A, analogous to a small portion of polymerase domain II of family B DNA polymerases (see figure 2.4, chapter II), has only a single identical amino acid, aspartate, in 14 residues (figure 5 of Wang et al., 1997). Likewise, region C, analogous to polymerase domain I of family B DNA polymerases, is characterized by the presence of two conserved aspartate residues, often separated by a single amino acid (Wang et al., 1997). The presence of three conserved aspartate residues in DNA polymerases, which are quite often over 1000 amino acids in length, surely cannot be considered as evidence supporting a common evolutionary origin for DNA polymerases. Rather, the presence of these conserved residues is perhaps better interpreted as evidence for mechanistic similarities in the biochemistry of catalysis.

Until very recently, the only crystal structure solved for a DNA-dependent DNA polymerase was that of *E. coli* DNA polymerase I (Klenow fragment), a family A polymerase (Ollis et al., 1985; Beese et al., 1993). The

Klenow fragment is composed of two domains, the 3′-5′ exonuclease domain and the polymerization domain. The structure of the polymerization domain has been likened to that of a U-shaped hand: the thumb and fingers pointing upwards forming a cleft, the palm, which functions in DNA binding. Conserved aspartate residues of regions A and C are located in the palm domain, consistent with mutational studies identifying these residues as functioning in catalysis. Even though there is no amino acid similarity between *E. coli* DNA polymerase I and HIV-1 reverse transcriptase, a RNA-dependent DNA polymerase, the crystal structure of these two proteins are remarkably similar as the catalytic residues are located in approximately the same location in the palm (Kohlstaedt et al., 1993).

It was expected, therefore, that the crystal structure of a family B DNA polymerase should appear similar to those of *E. coli* DNA polymerase I and HIV-1 reverse transcriptase. However, the recently solved structure of the family B DNA polymerase of bacteriophage RB69 exhibited a large number of structural differences (Wang et al., 1997). For instance, the finger domain, which includes conserved polymerase domains III and IV of family B DNA polymerases, is not homologous to the finger domain of any other polymerase. Polymerase domain III of family B DNA polymerases has been implicated in dNTP binding (Dong and Wang, 1995), but functionally analogous residues of *E. coli* DNA polymerase I are located in a different α-helix than those of the RB69 DNA polymerase. Wang and colleagues concluded that while the residues of RB69 and *E. coli* DNA polymerases involved in dNTP binding may play similar roles in DNA synthesis, it is unlikely that the similarities arose by divergence from a common ancestor but rather arose by convergence (Wang et al., 1997). However, the palm domain of RB69 DNA polymerase appears to be structurally similar to *E. coli*

DNA polymerase I. The only significant difference appears to be the location of one of the conserved aspartate residues, which lies in a different α-helix in the RB69 DNA polymerase than the functionally analogous residue of *E. coli* DNA polymerase I.

Family A and B DNA polymerases thus appear to be a mix of conserved and non-conserved structures and functionally analogous residues. The balance of evidence neither supports nor excludes the possibility that these DNA polymerases evolved by divergence from a common ancestor. It is tempting to speculate that the crystal structure of a family C DNA polymerase (typified by *E. coli* DNA polymerase III) would help in resolving issues of divergence or convergence, but it is possible that this polymerase structure would be different again from the family A and B structures.

### Archaebacterial genomes encode proteins most similar to eukaryotic replication-associated proteins

In grade of cellular organization, archaebacteria resemble eubacteria in many aspects; both have circular chromosomes with tightly packed genes, often overlapping and (sometimes identically) arranged operons, eubacteria (and presumably) archaebacteria use a single origin of replication, and both share common cell division components (Fts family of proteins; Margolin et al., 1996; Baumann and Jackson, 1997). The eukaryotic-style of genome organization, cell division mechanisms, and cell cycle control are so radically different from those of eubacteria and archaebacteria that it would be reasonable to assume that archaebacteria would appear eubacterial in terms of replication proteins.

**Table 4.3** Comparisons of archaebacterial, eukaryotic and eubacterial replication proteins. Archaebacterial ORFs identified by BLASTP and TBLASTN searches with eukaryotic replication proteins are presumed to have some function in replication. Optimal score, range and P-value are as for table 4.2. *Methanococcus jannaschii* is abbreviated to Mj. Mj0363, 0961, and 1498 refer to ORFs with sequence similarity to eukaryotic minichromosome maintenance (MCM) proteins. Only one comparison was performed between these ORFs and a single MCM protein, *S. cerevisiae* MCM2. MjYPZ1 and MjYPV1 refer to plasmid-encoded proteins that show similarity to *S. cerevisiae* ORC1 and CDC6. Only the comparison between these proteins and ScORC1 is shown.

| Comparison | Optimal score | Range | P-value |
|---|---|---|---|
| MjPCNA vs. ScPCNA | 322 | 24-91 | $3.09 \times 10^{-15}$ |
| MjPCNA vs. EcDnaN | 28 | 23-84 | 0.948 |
| MjRad2/Fen-1 vs. ScFen1 | 533 | 26-83 | $1.62 \times 10^{-29}$ |
| MjRad2/Fen-1 vs. EcPolA | 62 | 28-86 | 0.0745 |
| MjRFC-A vs. ScRFC-1 | 238 | 31-108 | $2.31 \times 10^{-10}$ |
| ScRFC-2 | 183 | 26-90 | $1.46 \times 10^{-9}$ |
| ScRFC-3 | 321 | 27-100 | $4.86 \times 10^{-15}$ |
| ScRFC-4 | 279 | 28-75 | $4.08 \times 10^{-15}$ |
| ScRFC-5 | 125 | 26-90 | $5.43 \times 10^{-6}$ |
| MjRFC-A vs. EcDnaX | 63 | 28-98 | 0.0628 |
| MjRFC-A vs. EcHolB | 64 | 25-93 | 0.00871 |
| MjDnap vs. ScDnapd | 313 | 35-113 | $1.38 \times 10^{-15}$ |
| MjDnap vs. ScDnape | 314 | 31-112 | $2.17 \times 10^{-13}$ |
| MjDnap vs. ScDnapa | 237 | 34-111 | $4.87 \times 10^{-11}$ |
| MjDnap vs. EcPolA | 46 | 31-100 | 0.656 |
| MjDnap vs. BsPolC | 49 | 36-130 | 0.698 |
| Mj0363 vs. ScMCM2 | 416 | 31-101 | $6.01 \times 10^{-22}$ |
| Mj0961 vs. ScMCM2 | 441 | 31-96 | $2.53 \times 10^{-22}$ |
| Mj1498 vs. ScMCM2 | 486 | 31-100 | $5.52 \times 10^{-27}$ |
| MtYPZ1 vs. ScORC-1 | 94 | 28-90 | $5.25 \times 10^{-3}$ |
| MtYPV1 vs. ScORC-1 | 94 | 28-84 | $2.65 \times 10^{-3}$ |

Even before the availability of genome sequence, it was clear from drug inhibition studies that archaebacteria resembled eukaryotes in terms of sensitivities to various DNA replication inhibitors (Forterre et al., 1984; Schinzel and Burger, 1984). Archaebacterial DNA-dependent DNA polymerases that were purified, biochemically characterized, and sequenced all revealed more primary sequence similarity to eukaryotic replicative family B homologs than to the non-replication functioning eubacterial homolog (reviewed in Forterre and Elie, 1993). In addition, the complete genome sequence of *Methanococcus jannashcii* has only a single family B DNA polymerase (Bult et al., 1995), making it likely that it is the replicative DNA polymerase.

Moreover, the *M. jannashcii* genome sequence, and individually sequenced archaebacterial replication proteins, shows many ORFs that are most clearly (or only) related in sequence to eukaryotic replication proteins (Bult et al., 1995; summarized in table 4.4). There are three general categories of results from sequence comparisons of replication proteins of eubacteria, archaebacteria, and eukaryotes. First, there are archaebacterial ORFs that are clearly homologous to eukaryotic replication proteins, while evidence that either the archaebacterial or eukaryotic protein is homologous to eubacterial proteins performing the same function is weak or absent (table 4.2). This category includes most of the replication-fork associated proteins.

Second are instances in which eubacterial and eukaryotic replication proteins are likely homologs, but the archaebacterial and eukaryotic versions are more similar in primary sequence. For instance, *M. jannaschii* possesses two homologs of the clamp loading complex that are more similar to the eukaryotic homologs (RF-C) than to the eubacterial homologs (γ complex; see table 4.2 and chapter III). Archaebacterial family B DNA polymerases are also

**Table 4.4** Summary of evidence for homology of eubacterial, archaebacterial, and eukaryotic replication proteins (based on Stillman, 1994). All eukaryotic and archaeal replication proteins share significant amino acid similarity. None of the bacterial replication proteins share signficant similarity with either eukaryotic or archaeal proteins performing analogous functions except those that are boxed.

[1] Archaeal proteins are from *M. jannaschii* unless indicated.

[2] See introduction for a discussion of homology.

[3] ? indicates that no predicted open reading frame from the *M. jannaschii* genome with significant similarity to known single-strand DNA-binding proteins was found.

[4] See table 1 for classification of family B DNA polymerases

[5] The eubacterial primase, DnaG, and eukaryotic DNA polymerase α are claimed to have homologous functional residues in conserved domains (fig. 4 of Prasartkaew et al., 1996). However, only 4 of 19 (21%) residues of *E. coli* DnaG and *Homo sapiens* DNA polymerase α are similar in motif A, none are identical. Of 16 residues of motif C, only 1 is identical (6%) while 3 are similar (18%). The proteins are not alignable outside of these domains.

[6] Identification of a eukaryotic replication fork-associated helicase has been problematic. Dna2, a yeast helicase, associates with the 5′-3′exo-endonuclease FEN1/Rad2(pombe) of yeast and is most likely involved in Okazaki fragment maturation (Budd and Campbell, 1997). It is not clear if Dna2 is associated with origin unwinding (as is PriA in *E. coli*), or with unwinding of the replication fork (as is DnaB in *E. coli*).

[7] The 5′-3′ exonuclease domain of eubacterial DNA polymerase I and the eukaryotic 5′-3′ exonuclease FEN-1/Rad2(pombe) have been classified as members of a homologous protein family based on amino acid alignments (reviewed in Leiber, 1997). However, the 5′-3′ exonuclease domain of *E. coli* DNA polymerase I (301 amino acids) and murine FEN1 (337 amino acids) are only 21% similar.

| Function at replication fork | Eubacterial protein (E. coli) | Eukaryotic protein (yeast/human) | Archaeal protein[1] | Evidence for homology [2] |
|---|---|---|---|---|
| origin recognition | DnaA | Origin recognition complex (ORC) proteins 1-6 | ORC1-like (plasmid encoded) M. thermoformicicum | little or no 1° sequence similarity between eub and euk/arch; all ATP-dependent |
| single-strand DNA-binding protein; loading of helicase, stimulates DNA polymerase | SSB | Replication protein A (RPA; 3 subunits) | ????[3] | little or no 1° sequence similarity between eub/euk; 3° structure similarity in SSB-binding domain |
| synthesis of primer | DnaG | DNA polymerase α<br>(Family B DNA polymerase)[4] | Family B DNA polymerase (many archaeal sequences) | little or no 1° sequence similarity between DnaG and DNA polymerase α[5] |
| helicase | DnaB (5'-3' helicase)<br>PriA (3'-5' helicase) | Dna2 (3'-5' helicase) | archaeal Dna2 | little or no 1° sequence similarity among helicases, expept in ATP-binding pocket; different activities and substrate specificities[6] |
| clamp loading complex; DNA-dependent ATPase, stimulates loading of DNA polymerase | γ complex (γδδ'χΨ)<br>**dnaX=γ subunit**<br>**holB=δ' subunit** | **Replication factor-C (RFC) 5 homologous subunits, RFC 1-5** | **archaeal RFC-1**<br>**archaeal RFC-3** | significant a.a. similarity between eub/euk/arch; similar biochemistry (see text) |
| processivity factor, 'sliding clamp' | Polβ (dnaN) | Proliferating cell nuclear antigen (PCNA) | archaeal PCNA | little or no 1° sequence similarity but crystal structures of eub & euk proteins are identical and superimposable; see text |
| synthesis of leading and lagging strands | DNA polymerase III core (αθε)<br>(Family C DNA polymerase) | DNA polymerase α/ε/δ (Family B DNA polymerase) | Family B DNA polymerase (many archaeal sequences) | little or no 1° sequence similarity between different families of polymerases; all have similar biochemical activities (3'-5' exo, polymerization) |
| ligation of fragments on lagging strand | DNA ligase (NAD-dependent) | DNA ligase (ATP-dependent) | DNA ligase (ATP-dependent)<br>M. jannaschii, D. ambivalens, M. thermoformicicum | little or no 1° sequence similarity and different nucleotide co-factors |
| removal of primers | DNA polymerase I (Family A DNA polymerase);<br>**Ribonuclease H** | FEN1/Rad2 (pombe);<br>**Ribonuclease H** | archaeal FEN1/Rad2;<br>**Ribonuclease H** | ribonuclease H's homologous; the 5'-3' exo domain of E. coli DNA polI and FEN1/Rad2 are claimed to be homologous but alignments are weak.[7] |

more similar in primary sequence to eukaryotic homologs than to the *E. coli* family B DNA polymerase (chapter II).

Third are comparisons in which replication functions are performed by a number of homologous proteins in eukaryotes, but that appear to be reduced in number in archaebacteria. Thus there are three replicative family B DNA polymerases in eukaryotes (α, δ and ε; Braithwaite and Ito, 1993) but only a single homolog in *M. jannaschii;* five clamp loading proteins in *S. cerevisiae* and *H. sapiens* (Cullmann et al., 1995), but only two in *M. jannaschii;* six MCM proteins in eukaryotes (Kearsey et al., 1996), but only three in *M. jannaschii.* Interestingly, the *M. jannaschii* MCM homologs are all more similar to each other than to eukaryotic homologs, suggesting that they result from a gene duplication event independent of the duplication event that gave rise to the eukaryotic homologs (chapter III). Whether or not this trend of reduced numbers of homologs of eukaryotic replication proteins will hold for other archaebacteria awaits further genome sequence; it is possible that M. *jannaschii* represents a case of genome reduction (*Sulfolobus solfataricus* P2 has three family B DNA polymerases whereas *M. jannaschii* has only one; Edgell et al., 1997).

### Gaps in the data

While *M. jannaschii* appears to encode a basic set of eukaryote-like replication proteins, there are a number of critical components that appear to be missing or that have not yet been identified from the complete genome sequence. For instance, no single-stranded DNA-binding protein was identified (Bult et al., 1995), yet this protein is essential for initiation of replication in both eubacteria and eukaryotes (Kornberg and Baker, 1992). Also critical for initiation of replication are origin-binding proteins, yet neither eubacterial nor eukaryotic homologs were reported in the initial publication (Bult et al., 1995). Subsequent work by other researchers identified a possible homolog of the eubacterial origin-binding protein DnaA (Koonin, 1997). However, it is unlikely that this protein is a true homolog of DnaA, as database searches with this ORF have low significance values, and the predicted protein is a member of the largest gene family (>20 genes) in the *M. jannaschii* genome. Two sequences in databases, not from *M. jannaschii* but from the closely related *Methanobacterium thermoformicicum* (Nolling et al., 1992), are possible homologs of the ORC1 and CDC6 proteins of eukaryotes (table 4.2). Curiously, these genes are present on plasmids and may be important for plasmid maintenance and replication. It is not clear what role, if any, these proteins might play in chromosomal replication.

## Chapter 5. General discussion and conclusions

### The last common ancestor of life had a DNA-based genome

Accumulation of genome sequencing data from representatives of the three major lineages has proved invaluable in determining the nature of the last common ancestor of life (Clayton et al., 1997). The presence of shared proteins in all three lineages with significant sequence similarity, similar biochemistry, and similar cellular functions surely provides compelling evidence against the idea that the common cellular ancestor from which archaebacteria and eubacteria evolved was a progenote, a primitive entity still in the process of refining the accuracy and efficacy of many cellular pathways found in extant organisms. Rather, evidence suggests that the last common ancestor (the cenancestor) was likely a sophisticated organism resembling a modern-day eubacterium or archaebacterium in aspects of genome organization and gene content. Evidence supporting the "advanced" nature of the last common ancestor can only become stronger as other genome sequences are completed and as detailed structure/function studies are undertaken for many archaebacterial homologs of eukaryotic and eubacterial proteins.

Yet, work presented in this thesis seems in fact to argue in favor of a common ancestor that was "unsettled" in the choice of DNA replication machinery as most replication proteins of eubacteria and archaebacteria/ eukaryotes show little or no sequence similarity. In this respect, Woese and Fox (1977a) were correct in predicting that [italics mine]

> "Thus, genome organization, control hierarchies, (some) repair mechanisms, *certain enzymes involved in DNA replication,*

should appear quite dissimilar in the two cases [prokaryote and eukaryote]."

However, the last common ancestor as reconstructed from genome sequencing data was perhaps not as primitive as Woese and Fox first envisioned for sufficient evidence exists to suggest that it possessed a DNA-based genome and the ability to replicate it: a proofreading family B DNA polymerase, some (but not all) clamp loading proteins, and a processivity factor all can be traced back to the last common ancestor using parsimony-based arguments (chapters III and IV).

Other authors have suggested that the lack of sequence similarity of replication proteins is due to the fact that the last common ancestor did not have a DNA-based genome, but one based on RNA (Mushegian and Koonin, 1996). Replication proteins would have thus evolved independently in the lineages leading to eubacteria and archaebacteria/eukaryotes after they split from a common ancestor. However, this is unlikely, for two reasons. First, despite the general paucity of significant primary sequence similarity between replication fork proteins of eubacteria and archaebacteria/eukaryotes, some proteins *are* homologs, as discussed in chapters II and III (see also O'Donnell et al. 1993; Stillman, 1994; Edgell and Doolittle, 1997).

Second, other components essential for replication, but not always situated at the replication fork, are also found in representatives of all three domains (Benner et al., 1989; Benner et al., 1993). Metabolic enzymes essential for the synthesis of dNTPs from rNTPS (ribonucleotide reductases), and for maintaining low levels of dUTP (deoxyuridine triphosphatase) to ensure that dUTP is not incorporated into DNA, are also found in representatives of the three domains (Riera et al., 1997; Tauer and Benner, 1997). In most cases it is clear that analogous functions are performed by homologous proteins. Thus, multiple components involved in DNA replication, both at the replication fork and elsewhere, can be traced back to the cenancestor. It is likely then that the cenancestor was a DNA-based organism with a working DNA replication apparatus of some sort but because of the lack of sequence similarity of eubacterial and archaebacterial/eukaryotic replication proteins, it cannot be confidently stated what kind of replication apparatus is was. The (not mutually exclusive) possibilities are:

- The above arguements notwithstanding, most eubacterial and archaeal/eukaryotic replication proteins are homologs (do descend from cenancestral proteins performing the same function) but have often been so radically changed in sequence as to be unrecognizable.

- The cenancestor contained both eubacterial and archaebacterial/eukaryotic type replication systems (perhaps one was for repair), and different components of these systems were lost in the eubacterial and archaebacterial/eukaryotic lineage after their divergence.

- "New" (nonhomologous) proteins have been recruited into a replication function in one or the other lineages, replacing cenancestral components.

Why are replication proteins so divergent?

Still, the lack of sequence similarity of eubacterial and archaebacterial/ eukaryotic replication proteins is both confusing and surprising given that other protein components of essential cellular processes (such as transcription and translation) are conserved in sequence and function.

Some of the lack of sequence conservation could be due to the organizational differences of eukaryotic and prokaryotic genomes and the tremendous changes that accompanied the evolution of eukaryotic-style chromosomes from prokaryotic-style chromosomes. However, the prokaryotic-eukaryotic transition cannot be the cause of these changes; they must have occurred prior to the divergence of archaebacteria and eukaryotes, not at the origin of eukaryotes, since archaebacteria appear to have a basic set of eukaryotic-style replication proteins. In this sense, archaebacterial replication proteins are best viewed not as a "primitive" set of eukaryotic-style replication proteins, but as the ancestral set of replication proteins that eukaryotes have built upon by gene duplications (chapters II and III), gene fusions (chapter III), and recruitment of additional proteins (chapter IV). This is not to imply that the function of replication proteins shared by archaebacteria and eukaryotes will be the same. There will, for instance, be differences in the mechanism(s) by which archaebacteria and eukaryotes control the initiation of DNA replication that will likely be reflected in the functions of archaebacterial and eukaryotic MCM homologs.

The lack of sequence similarity between eubacterial and archaebacterial/eukaryotic replication proteins is most confusing *even if* changes in replication machinery occurred before the archaebacterial/ eukaryotic split. As discussed above, it is possible that all replication proteins are homologs but have diverged too much in sequence to identify them as such. A test of this theory would be to examine *within*-domain rates of sequence evolution of replication proteins and compare these rates to within-domain rates of proteins functioning in other cellular pathways. If replication proteins did in fact diverge from a single ancestral set of proteins, one might expect the rate of sequence evolution to be high between (for instance) two divergent eubacteria as compared to proteins involved in translation or transcription from the same two eubacteria. This high rate of sequence evolution might reflect differences in functional constraints on replication proteins versus constraints on other proteins. If, however, rates of sequence evolution are no higher than those of (for example) eubacterial transcription or translation proteins, this might be interpreted as evidence in favour of the last common ancestor of life possessing two non-homologous sets of replication proteins. The rate of sequence evolution would be no greater than any other protein since the replication apparatus was already "settled" in the last common ancestor and did not require high rates of sequence evolution to adapt to new replication fork functions and functional constraints after the divergence of the eubacterial and archaebacterial/ eukaryotic lineages. One of these ancestral sets of replication proteins would have been retained along the eubacterial lineage (typified by family A and C replicative polymerases), and the other retained in the archaebacterial/ eukaryotic lineage (typified by multiple family B replicative polymerases).

It might also be possible to detect remnants of one system in complete genome sequences even though the protein(s) might no longer function in replication. The presence of a family B DNA polymerase in *E. coli* that can interact with accessory proteins necessary for replication (Hughes et al., 1991; Bonner et al., 1992), and that can replicate chromosomal DNA in certain genetic backgrounds (Rangarajan et al., 1997), could possibly be interpreted as evidence supporting this hypothesis.

The same rate comparison, but performed on eukaryotic replication proteins, might reveal an unexpectedly high rate of sequence evolution compared to proteins with other cellular function because many eukaryotic replication proteins have evolved by gene duplications. In eukaryotes, DNA replication is tightly linked to the cell cycle by various checkpoints that insure completion of replication (DNA polymerase ε; Navas et al., 1995), that insure that initiation is limited to defined periods (CDC6 and MCM proteins; Kearsey et al., 1996), and that insure that replication only initiates once from an origin (ORC and MCM proteins; Diffley, 1997). All of the proteins involved in these checkpoints seem logical points at which to control DNA replication; a high rate of sequence evolution of these proteins within eukaryotes might be reflective of duplicates accumulating non-synonymous nucleotide substitutions and being selected for regulatory function(s), perhaps through protein-protein interactions with other regulatory proteins.

## TIGR and the honey pot

After the completion of all of the phylogenetic analyses presented in this thesis and most of the writing, The Institute for Genomic Research (TIGR) released 2 complete and 7 partial sequences of prokaryotic genomes (but not annotated). Among those released were the completely sequenced genome of the euryarchaeote *Archaeoglobus fulgidus* and the partially completed sequence of the eubacterium *Thermotoga maritima*. Both of these organisms occupy phylogenetic positions such that knowledge of their gene content would be extremely useful, *A. fulgidus* because it is unrelated to *M. jannaschii* and sometimes branches near the base of euryarchaeotes close to crenarchaeotes in 16S rRNA phylogenies (Olsen et al., 1994; Pace, 1997), and *T. maritima* because it is a deeply diverging eubacterium (see Eisen, 1995 and references therein) and phylogenetically distinct from previously complete eubacterial genomes. The completely sequenced genome of *Methanobacterium thermoautotrophicum* also became available at approximately the same time as the TIGR data (http://www.cric.com/

htdocs/sequences/methanobacter/ abstract.html). However, *M. thermoautotrophicum* will most likely ressemble *M. jannaschii* in gene content since both are similar in genome size and biochemistry (methanogenesis).

TBLASTN searches (Altschul et al., 1997) performed on the *A. fulgidus* sequence with replication proteins of *M. jannaschii, S. solfataricus* P2 and *S. cerevisiae* produced a number of interesting results:

- *A. fulgidus* appears to possess two family B DNA polymerases unlike *M. jannaschii* which only has one (Bult et al., 1996), but similar to crenarchaeotes, which possess two (*P. occultum;* Uemori et al., 1995) or three (*S. solfataricus* P2; Edgell et al., 1997).
- *A. fulgidus* appears to possess a single MCM homolog that is more similar in sequence to the *S. solfataricus* P2 MCM protein than to any of the three chromosomally-encoded *M. jannaschii* MCM proteins.
- *A. fulgidus* appears to possess two RF-C proteins as does *M. jannaschii.* One appears more similar in sequence to the *M. jannaschii* B paralog, which is likely an ortholog of eukaryotic RFC-1 proteins. Neither of the *A. fulgidus* RF-C paralogs appears to possess the DNA ligase domain common to eukaryotic RFC-1 proteins.

It is difficult to draw any firm conclusions on the presence of two family B DNA polymerases in *A. fulgidus.* Explanations that revolve around gene loss to account for the distribution of DNA polymerases in euryarchaeotes depend on the phylogenetic position of *A. fulgidus* relative to *M. jannaschii* and other euryarchaeotes. Unfortunately, the phylogenetic position of *A. fulgidus* within euryarchaeotes is unresolved as it sometimes branches closer to the base of euryarchaeotes than does *M. jannaschii,* and sometimes closer to terminal groups (Olsen et al., 1994). Thus, without detailed phylogenetic analyses of family B DNA polymerases, it is impossible to decide whether or not *M. jannaschii* lost one polymerase (assuming *M. jannaschii* is more basal than *A. fulgidus*), or whether a gene duplication event occurred after the divergence of *A. fulgidus* and *M. jannaschii.*

TBLASTN searches with the *E. coli* family B DNA polymerase against sequences from 7 of 8 eubacterial genomes released by TIGR failed to produce any significant results. However, one partially completed genome sequence, that of *Vibrio cholerae,* possessed a sequence which significantly matched the *E. coli* family B DNA polymerase in three unlinked contigs, all with expected P-values below $10^{-50}$. Inspection of the *V. cholerae* sequence confirmed that they contained exonuclease domain I, and polymerization domains I, II, V, and VII; all of these domains are found in family B DNA polymerases.

The finding of another family B DNA polymerase in a eubacterial genome, although from an organism closely related to *E. coli* (both are members of the $\gamma$-subdivision of proteobacteria), only adds confusion to questions raised in chapter II concerning the evolutionary history of eubacterial family B DNA polymerases. Foremost was whether or not the presence of this type of DNA polymerase in the *E. coli* genome, but not in any other eubacterial genome, was due (a) to a lateral transfer of this gene from a non-eubacterial source or (b) to multiple independent deletion events from eubacterial lineages except that leading to *E. coli.* Parsimony-based arguments suggest that multiple independent losses of a polymerase, present in the common ancestor of all eubacteria, from eubacterial lineages except that leading to *E. coli* and *V. cholerae* are unlikely. Since *V. cholerae* is closely related to *E. coli*, it is perhaps more parsimonious to suggest that this gene was acquired from a non-eubacterial, or a divergent eubacterial source, in the common ancestor of *V. cholerae* and *E. coli.* Yet, the amino acid sequence of the *E. coli* family B DNA polymerase is not specifically close to any eukaryotic or archaebacterial sequence as would be expected if *E. coli* acquired this gene by lateral transfer.

### Archaebacterial, eukaryotic, and eubacterial specific genes and gene functions

More than half of the predicted ORFs in the *M. jannaschii* genome sequence do not match any sequence in public databases (Bult et al., 1996). This apparent difference between the *M. jannaschii* genome and those of *Mycoplasma genitalium* and *Haemophilus influenzae* (Fleischmann et al., 1995; Fraser et al., 1995), of which over 80% matches database sequences, was heralded as support for the biological uniqueness and phylogenetic coherence of archaebacteria (Morell, 1996).

Surely much of this difference must be due to biases in representation of sequences in databases. A large percentage of the bacterial (prokaryotic) subset of GenBank consists of proteobacterial sequences (such as *E. coli* and *Salmonella typhimurium*). It is not surprising then that many of the ORFs of another proteobacterium, *H. influenzae,* match those existing proteobacterial sequences (Fleischmann et al., 1995). As more archaebacterial genes and genomes are sequenced, the number of unidentified ORFs in *M. jannaschii* (and other archaebacteria) will decrease as similar sequences are found in other archaebacteria. The function of these ORFs may still be unknown, since function can only be assigned through detailed biochemical and genetic studies.

Yet there will be a number of ORFs in archaebacteria, eubacteria, and eukaryotes for which one can never find similar sequences in databases. Since much of the basic biochemistry of metabolism and genetic processes is shared between archaebacteria, eukaryotes and eubacteria, it would be surprising to find that many new proteins had been invented *de novo* in either archaebacterial or eubacterial lineages after their divergence from a biochemically and structurally advanced last common ancestor. The current gene content of extant archaebacterial, eubacterial and eukaryotic genomes *must* have evolved from the gene content of the last common ancestor of all three groups. Thus, there may be few archaebacterial-, eubacterial- or eukaryotic-specific genes, although there will certainly be archaebacterial-, eubacterial- and eukaryotic-specific gene functions.

There are also a number of ORFs in archaebacteria that are homologous to eukaryotic proteins, which I classify into two broad categories. First, are archaebacterial proteins homologous to eukaryotic proteins that function in cellular processes that we typically think of as "eukaryotic", such as meiosis (see Ragan et al., 1996 for an example). The function of these archaebacterial homologs is unlikely to be analogous to their eukaryotic counterparts for archaebacteria lack many of the cellular processes of eukaryotes; these eukaryote-specific functions must have been added to the proteins after the split of archaebacteria and eukaryotes from a common ancestor. Second are archaebacterial proteins that are homologous to eukaryotic proteins, both of which are probably involved in the same cellular process but not necessarily with the same biochemistry or analogous function(s). The multiple family B DNA polymerases of crenarchaeotes (Uemori et al., 1995; Edgell et al., 1997) and eukaryotes (Braithwaite and Ito, 1993) are an excellent example of this category of homolog.

What to do next?

It is in the second category of archaebacterial and eukaryotic homologs that I think excellent research opportunities lie for, in spite of the availability of the complete genome sequence of *M. jannaschii,* little is still known about the function of archaebacterial DNA replication proteins. Some likely projects are:

- detailed studies on the biochemistry, function, and evolution of the divergent family B DNA polymerases of the *Sulfolobales.*
- function(s) of archaebacterial homologs of eukaryotic DNA replication proteins, such as MCM proteins, in halophilic archaebacteria where a working genetic system exists. A parallel line of investigation could focus on issues such as control of initiation of replication, chromosome segregation, and replication origins.
- a computer-based approach concerning the rates of sequence evolution of eukaryotic, archaebacterial and eubacterial DNA replication proteins.

These proposed studies could not only further understanding of archaebacterial DNA replication, but also help in understanding how the vast differences in eubacterial and eukaryotic/archaebacterial DNA replication systems arose.

**Appendix 1:** Sequences of degenerate and exact match PCR primers.

<u>degenerate DNA polymerase primers</u>

<u>α-specific</u>

| | |
|---|---|
| 1. DPDV(I)IV(I)GH | 5′ GAYCCNGAYRTNATTHRTNGGNC 3′ |
| 2. DFNSLYPS | 5′ GAYTTYAATWSNCTNTAYCCNTG 3′ |
| 3. KKKYAA | 5′ AARAARAARTAYGCNGC 3′ |

<u>δ-specific</u>

| | |
|---|---|
| 1. FDIEC (+ *Eco*RI site) | 5′ GGAATTCTTYGATATHGARTGC 3′ |
| 2. YGFTGA | 5′ TATYGGNTTYTAYGGNGC 3′ |
| 3. DTDSVM (+ *Eco*RI site) | 5′ CGGGATCCATNACNGARTCNGTRTC 3′ |
| 4. DCPIFY | 5′ GTARAADAGNGGRCARTC 3′ |

<u>ε-specific</u>

| | |
|---|---|
| 1. QIMMISY | 5′ CAGATYATGATGATYTCNTAC 3′ |
| 2. NGDFFDWPF | 5′ AAYGGNGAYTTYTTYGATTGGCCNTT 3′ |
| 3. MYPNI (+ *Eco*RI site) | 5′ GAATTCDATRTTNGGRTACAT 3′ |
| 4. ELDTDG | 5′ CCRTCNGTRTCNAGG 3′ |

<u>α and δ</u>

| | |
|---|---|
| 1. SLYPSI (+ *Eco*RI site) | 5′ GGAATTCNCTSTAYCCNTC 3′ |
| 2. YGDTDS (+ *Eco*RI site) | 5′ GGAATTCNGTRTCNCCSTA 3′ |

<u>direct-match primers for amplication of *S. shibatae* and *S. solfataricus* P2 B3 and *S. solfataricus* B2 P2 DNA polymerase</u>

| | |
|---|---|
| 1. Shib B3-1 | 5′ TAGCCATGTTTATGTTC 3′ |
| 2. Shib B3-2 | 5′ TTGACTAGAGTATCTGG 3′ |
| 3. UPS-1 | 5′ AGAGGGCACATAGTCATAGC 3′ |
| 4. DST-1 | 5′ CGTTCTCTATGATAATAATTGG 3′ |

B=C/T/G; D=A/T/G; H=A/C/T; K=T/G; M=A/C; N=A/C/T/G; R=A/G; S=C/G; V=A/C/G; W=A/T; Y=C/T

**References**

Adachi, J., and Hasegawa, M. (1992). MOLPHY, Programs for molecular phylogenetics, version 2.2. Institute of Statistical Mathematics, Tokyo.

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.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST - A New Generation of Protein Database Search Programs. Nucleic Acids Res. *in press*

Amils, R., Cammarano, P., and Londei, P. (1993). Translation in archaea. *In* In The Biochemistry of Archaea (Archaebacteria), M. Kates, D. J. Kushner, and A. T. Matheson, eds. (Amsterdam, Elsevier), pp. 393-432.

Araki, H., Ropp, P. A., Johnson, A. L., Johnson, L. H., Morrison, A., and Sugino, A. (1992). DNA polymerase II, the probable homolog of mammalian DNA polymerase ε replicates chromosomal DNA in the yeast *Saccharomyces cerevisiae*. EMBO J. **11**, 733-740.

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 of life and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA **93**, 7749-7754.

Baumann, P., and Jackson, S. P. (1996). An archaebacterial homologue of the essential eubacterial cell division protein FtsZ. Proc. Natl. Acad. Sci. USA **93**, 6726-6730

Beese, L. S., and Steitz, T. A. (1991). Structural basis for the 3'-5' exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. EMBO J. **10**, 733-740.

Beese, L. S., Derbyshire, V., and Steitz, T. A. (1993). Structure of DNA polymerase I Klenow fragment bound to duplex DNA. Science **260**, 352-355.

Bell, S. P., Mitchell, J., Leber, J., Kobayashi, R., and Stillman, B. (1995). The multidomain structure of Orc1p reveals similarity to regulators of DNA replication and transcriptional silencing. Cell **83**, 563-568.

Benner, S. A., Elington, A. D., and Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. Proc. Natl. Acad. Sci. USA **86**, 7054-7058.

Benner, S. A., Cohen, M. A., Gonnet, G. H., Berkowitz, D. B., and Johnsson, K. P. (1993). Reading the palimpsest: contempory biochemical data and the RNA world. *In* The RNA World, R. F. Gasteland and J. F. Atkins, eds. (Cold Spring Harbour Laboratory Press), pp. 27-70.

Bernad, A., Blanco, L., Lázaro, J. M., Martín, G., and Salas, M. (1990). A conserved 3′-5′ exonuclease active site in prokaryotic and eukaryotic DNA polymerases. Cell **59**, 219-228.

Blanco, L., Bernad, A., Blasco, M. A., and Salas, M. (1991). A general structure for DNA-dependent DNA polymerases. Gene **100**, 27-38.

Blow, J. J., and Laskey, R. A. (1988). A role for the nuclear envelope in controlling DNA replication within the cell cycle. Nature **332**, 546-548

Bochkarev, A., Pfuetzner, R. A., Edwards, A. M., and Frappier, L. (1997). Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA. Nature **385**, 176-181.

Bonner, C. A., Hays, S., McEntee, K., and Goodman, M. F. (1990). DNA polymerase II is encoded by the DNA damage-inducible *din*A gene of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **87**, 7663-7667.

Bonner, C. A., Stukenberg, P. T., Rajagopalan, M., Eritja, R., O'Donnell, M., McEntee, K., Echols, H., and Goodman M. F. (1992). Processive DNA synthesis by DNA polymerase II mediated by DNA polymerase III accessory proteins. J. Biol. Chem. **267**, 11431-11438.

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

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

Budd, M. E., and Campbell, J. L. (1993). DNA polymerases delta and epsilon are required for chromosomal replication in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **13**, 496-505.

Budd, M. E., and Campbell, J. L. (1997). A yeast replicative DNA helicae, Dna2 helicase, interacts with yeast FEN-1 in carrying out its essential function. Mol. Cell. Biol. **17**, 2136-2142.

Bult, C. J., et al. (1996). The complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii.* Science **273**, 1066-1073.

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

Burbelo, P. D., Utani, A., Pan, Z-Q., and Yamada, Y. (1993). Cloning of the large subunit of activator 1 (replication factor C) reveals homology with bacterial DNA ligases. Proc. Natl. Acad. Sci. USA **90**, 11543-11547.

Campbell, J. L., and Kleckner, N. (1990). *E. coli oriC* and the *dnaA* gene promoter are sequestered from *dam* methyltransferase following the passage of the chromosomal replication fork. Cell **62**, 967-979.

Carter, J. R., Franden, M. A., Aebersold, R., and McHenry, C. S. (1993). Identification, isolation, and characterization of the structural gene encoding the delta' subunit of *Escherichia coli* DNA polymerase III holoenzyme. J. Bacteriol. **175**, 3812-3822.

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

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

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

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

Chen, H., Lawrence, C. B., Bryan, S. K. and Moses, R. E. (1990). Aphidicolin inhibits DNA polymerase II of *Escherichia coli*, an α-like DNA polymerase. Nucleic Acids Res. **18**, 7185-7186.

Chen, Y., Hennessy, K., Botstein, D., and Tye, B. K. (1992). CDC46/MCM5, a yeast protein whose subcellular localization is cell cycle-regulated, is involved in DNA replication at autonomously replicating sequences. Proc. Natl. Acad. Sci. USA **89**, 10459-10463.

Clayton, R. A., White, O., Ketchum, K. A., and Venter, J. C. (1997). The first genome from the third domain of life. Nature **387**, 459-462.

Cocker, J. H., Piatti, S., Santocanale, C., Nasymth, K., and Diffley, J. F. X. (1996). An essential role for the Cdc6 protein in forming the pre-replicative complexes of budding yeast. Nature **379**, 180-182.

Copeland, W. C., and Wang, T. S.-F. (1993). Mutational analysis of the human DNA polymerase α. J. Biol. Chem. **268**, 11028-11040.

Cullmann, G., Fien, K., Kobayashi, R., and Stillman, B. (1995). Characterization of the five replication factor C genes of *Saccharomyces cerevisiae*. Mol. Cell. Biol. **15**, 4661-4671.

Curth, U., Greipel, J., Urbanke, C., and Maass, G. (1993). Multiple binding modes of the single-stranded DNA binding protein from *Escherichia coli* as detected by tryptophan fluorescence and site-directed mutagenesis. Biochemistry **32**, 2585-2591.

Darnell, J. E. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. Science **202**, 1257-1260.

De Vega, M., Lazaro, J., Salas, M., and Blanco, L. (1996). Primer-terminus stabilization at the 3'-5' exonuclease active site of Φ29 DNA polymerase. Involvement of two amino acid residues highly conserved in proofreading DNA polymerases. EMBO J. **15**, 1182-1192.

Diffley, J. F. X. (1997). Once and only once upon a time, specifying and regulating origins of DNA replication in eukaryotic cells. Genes Dev. **10**, 2819-2830.

Dong, Q., and Wang, T. S.-F. (1995). Mutational studies of human DNA polymerase α, Lysine 950 in the third most conserved region of α-like DNA polymerases is involved in binding the deoxynucleotide triphosphate. J. Biol. Chem. **270**, 21563-21570.

Dong, Z., Onrust, R., Skangalis, M., and O'Donnell, M. (1993). DNA polymerase III accessory proteins. I. holA and holB encoding delta and delta'. J. Biol. Chem. **268**, 11758-11765.

Donovan, S., Harwood, J., Drury, L. S., and Diffley, J. F. X. (1997). Cdc6p-dependent loading of Mcm proteins onto pre-replicative chromatin in budding yeast. Proc. Natl. Acad. Sci. USA **94**, 5611-5616.

Doolittle, R. F. (1986). Of URFs and ORFs, A primer on how to analyze derived amino acid sequences (Mill Valley, California, University Science Books).

Doolittle, R. F. (1995). The origins and evolution of eukaryotic proteins. Philos. Trans. R. Soc. (Lond). B **349**, 235-240.

Doolittle, W. F. (1978). Genes in pieces, were they ever together? Nature **272**, 581-582.

Doolittle, W. F. (1980). Revolutionary concepts in evolutionary cell biology. TIBS **5**, 146-149.

Doolittle, W. F. (1989). Whatever happened to the progenote? *In* The Hierarchy of Life, B. Fernholm, K. Bremer, and H. Jörnvall, eds. (Elseveir Science Publishers), pp. 65-72.

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

Drouin, G., Moniz de Sa, M., and Zucker, M. (1995). The *Giardia lamblia* actin gene and the phylogeny of eukaryotes. J. Mol. Evol. **41**, 841-849.

Edgell, D. R., Klenk, H-P., and Doolittle, W. F. (1997). Gene duplications in evolution of archaeal family B DNA polymeraes. J. Bacteriol. **179**, 2632-2640.

Edgell, D. R. and Doolittle, W. F. (1997). The origin(s) of archaeal DNA replication proteins. Cell **89**, 995-998.

Elie, C., De Recondo, A. M., and Forterre, P. (1989). Thermostable DNA polymerase from the archaebacterium *Sulfolobus acidocaldarius*. Eur. J. Biochem. **178**, 619-626.

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

Endo, T., Ikeo, K., and Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**, 685-690.

Escarceller, M., Hicks, J., Gudmundsson, G., Trump, G., Touati, D., Lovett, S., Foster, P. L., McEntee, K., and Goodman, M. F. (1994). Involvement of *Escherichia coli* DNA polymerase II in response to oxidative damage and adaptive mutation. J. Bacteriol. **176**, 6221-6228.

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

Felsenstein, J. (1996). PHYLIP (Phylogeny Inference Package) version 3.57c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

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

Fitch, W. S. (1970). Distinguishing homologous from analogous proteins. Sys. Zool. **19**, 99-113.

Forterre, P. (1992). The DNA polymerase from the archaebacterium *Pyrococcus furiosus* does not testify for a specific relationship betweeen archaebacteria and eukaryotes. Nucleic Acids Res. **20**, 1811.

Forterre, P., and Elie, C. (1993). Chromosome structure, DNA topoisomerases, and DNA polymerases in archaebacteria (archaea) In The Biochemistry of Archaea (Archaebacteria), M. Kates, D. J. Kushner, and A. T. Matheson, eds. (Amsterdam, Elsevier), pp. 325-361.

Forterre, P., Elie, C., and Kohiyama, M. (1984). Aphidicolin inhibits growth and DNA synthesis in halophilic archaebacteria. J. Bacteriol. **159**, 800-802.

Fox, G. E., et al. (1980). The phylogeny of prokaryotes. Science **209**, 457-463.

Fraser, C. M., et al. (1995). The minimal gene comlement of *Mycoplasma genitalium*. Science **270**, 397-403.

Fuchs, T., Huber, H., Burggraf, S., and Stetter, K. O. (1996). 16S rDNA-based phylogeny of the archael order *Sulfolobales* and reclassification of *Desulfurolobus amibvalens* as *Acidianus ambivalens* comb. nov.. System. Appl. Microbiol. **19**, 56-60.

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

Gibson, S., Surosky, R. T., and Tye, B. K. (1991). The phenotype of minichromosome maintenance mutant *mcm3* is characteristic of mutants defective in DNA replication. Mol. Cell. Biol. **10**, 5707-5720.

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

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

Gogarten, J. P., Hilaro, H., and Olendzenski, L. (1996). Gene duplications and horizontal gene transfer during early evolution. *In* Evolution of Microbial Life, D. McL. Roberts, P. Sharp, G. Alderson, and M. A. Collins, eds. Cambridge University Press, pp. 267-292.

Goodman, M., Moore, G. W., and Matsuda, G. (1975). Darwinian evolution in the genealogy of haemoglobin. Nature **253**, 603-608.

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

Gropp, F., Reiter, W. D., Sentenac, A., Zillig, W., Schnabel, R., Thomm, M., and Stetter, K. O. (1986). Homologies of components of DNA-dependent RNA polymerases of archaebacteria, eukaryotes and eubacteria. System. Appl. Microbiol. **7**, 95-101.

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.

Hain, J., Reiter, W-D., Hdepohl, U., and Zillig, W. (1993). Elements of an archael promoter defined by mutational analysis. Nucleic Acids Res. **20**, 5423-5428.

Hamal, A., Forterre, P., and Elie, C. (1990). Purification and characterization of a DNA polymerase from the archaebacterium *Thermoplasma acidophilum*. Eur. J. Biochem. **190**, 517-521.

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

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

Heichman, K. A. (1996). Cdc6 and DNA replication: limited to humble origins. BioEssays **18**, 859-862.

Hilario, E., and Gogarten, J. P. (1993). Horizontal transfer of ATPase genes-the tree of life becomes a net of life. Biosystems **31**, 111-119.

Himmelreich, R., Hilber, H., Plagens, H., Pirkl, E., Li, B. C., and Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. **24**, 4420-4449.

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

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

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. Syst. Biol. **44**, 17-48.

Huelsenbeck, J. P. (1997). Is the Felenstein zone a fly trap? Syst. Biol. **46**, 69-74.

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.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. Philos. Trans. R. Soc. Lond. B, Biol. Sci. **346**, 359-366.

Hughes, A. J. Jr., Bryan, S. K., Chen, H., Moses, R. E., and McHenry, C. S. (1991). *Escherichia coli* DNA polymerase II is stimulated by DNA polymerase II holoenzyme auxiliary subunits. J. Biol. Chem. **266**, 4568-4573.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, J. (1989). Evolutionary relationships of archaebacteria, eubacteria, and euakaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**, 9355-9399.

Iwasaki, H., Nakata, A., Walker, G. C., and Shinagawa, H. (1990). The *Escherichia coli polB* gene, which encodes DNA polymerase II, is regulated by the SOS system. J. Bacteriol. **172**, 6268-6273.

Iwasaki, H., Ishino, Y., Toh, H., Nakata, A., and Shinagawa, H. (1991). *Escherichia coli* DNA polymerase II is homologous to α-like DNA polymerases. Mol. Gen. Genet. **226**, 24-33.

Jensen, R. A. (1976). Enzyme recruitment in the evolution of new function. Ann. Rev. Microbiol. **30**, 409-425.

Joshi, P., and Dennis, P. P. (1993). Structure, function, and evolution of the family of superoxide dimutase proteins from halophilic archaebacteria. J. Bacteriol. **175**, 1572-1579.

Joyce, C. M., and Steitz, T. A. (1994). Function and structure relationships in DNA polymerases. Annu. Rev. Biochem. **63**, 777-822.

Kaneko, T., et al. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3**, 109-136.

Katayama, T., and Crooke, E. (1995). DnaA protein is sensitive to a soluble factor and is specifically inactivated for initation of *in vitro* replication of the *Escherichia coli* minichromosome. J. Biol. Chem. **270**, 9265-9271.

Kearsey, S. E., Maiorano, D., Holmes, E. C., and Todorov, I. T. (1996). The role of MCM proteins in the cell cycle control of genome duplication. BioEssays **18**, 183-190.

Keeling, P. J., and Doolittle, W. F. (1996a). A non-canonical genetic code in an early diverging eukaryotic lineage. EMBO J. **15**, 2285-2290.

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

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

Kelman, Z., and O'Donnell, M. (1995). Structural and functional similarties of prokaryotic and eukaryotic DNA polymerase sliding clamps. Nucleic Acids Res. **23**, 3613-3620.

Kesti, T., Frantti, H., and Syvaoja, J. E. (1993). Molecular cloning of the cDNA for the catalytic subunit of human DNA polymerase ε. J. Biol. Chem. **268**, 10238-10245.

Klemm, R. D., Austin, R. J., and Bell, S. P. (1997). Coordinate binding of ATP and origin DNA regulates ATPase activity of the origin recognition complex. Cell **88**, 493-502.

Klimczak, L. J., Grummt, F., and Burger, K. J. (1985). Purification and characterization of DNA polymerase from the archaebacterium Sulfolobus acidocaldarius. Nucleic Acids Res. **13**, 5269-5281.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A., and Steitz, T. A. (1992). Crystal structure of at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science **256**, 1783-1790.

Kong, H., R. B. Kucera, and Jack, W. E. (1993). Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis*. J. Biol. Chem. **268**, 1965-1975.

Koonin, E. V. (1997). Evidence for a family of archaeal ATPases. Science **275**, 1489-1490.

Kornberg, A., and Baker, T. A. (1992). DNA replication, 2nd. Ed. (New York, W. H. Freeman and Company).

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

Kumar, S., Tamura, K., and Nei, M. (1993). MEGA, Molecular Evolutionary Genetics Analysis (The Pennsylvania State University, University Park, PA). Version 1.0.

Lake, J. A., Henderson, E., Oakes, M. I., and Clark, M. W. (1984). Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci USA **81**, 3786-3790.

Lake, J. A., Henderson, E., Clark, M. W., Scheinman, A., and Oakes, M. I. (1986). Mapping evolution with three dimensional ribosome structure. System. Appl. Microbiol. **7**, 131-136.

Lawson, F. S., Charlebois, R. L., and Dillon, J-A. R. (1996). Phylogenetic analysis of carbmoylphosphate synthetase genes, complex evolutionary history includes an interal duplication within a gene which can root the tree of life. Mol. Biol. Evol. **13**, 970-977.

Leiber, M. R. (1997). The FEN-1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. BioEssays **19**, 233-240.

Li, W-H. (1997). Molecular Evolution. (Sinauer and Associates, Inc., Publishers).

Liang, C., Weinreich, M, and Stillman, B. (1995). ORC and Cdc6p interact and determine the frequency of initation of DNA replication in the genome. Cell **81**, 667-676.

Lu, M., Campbell, J. L., Boye, E., and Kleckner, N. (1994). SeqA: A negative modulator of replication initiation in *E. coli.* Cell **77**, 413-426.

Maine, G. T., Sinha, P., and Tye, B.-K. (1984). Mutants of *S. cerevisiae* defective in the maintenance of minichromosomes. Genetics **106**, 365-385.

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

Mathur, E. J., Adams, M. W., Callen, W. N., and Cline, J. M. (1991) . The DNA polymerase gene from the hyperthermophilic marine archaebacterium, *Pyrococcus furiosus,* shows sequence homology with α-like DNA polymerases. Nucleic Acids Res. **19**, 6952.

McGehee, R. E. Jr., and Habener, J. F. (1995). Differentiation-specific element binding protein (DSEB) binds to a defined element in the promoter of the angiotensinogen gene required for the irreversible induction of gene expression during differentiation of 3T3-L1 adipoblasts to adipocytes. Mol. Endocrinol. **9**, 487-501.

Mizushima, T., Nishia, S., Kurokawa, K., Katayama, T., Miki, T., and Sekimizu, K. (1997). Negative control of DNA replication by hydrolysis of ATP bound to DnaA protein, the initiator of chromosomal DNA replication in *Escherichia coli.* EMBO J. **16**, 3724-3730.

Morell, V. (1996). Life's last domain. Science **273**, 1043-1045.

Morrison, A., Christensen, R. B., Alley, J., Beck, A. K., Bernstine, E. G., Lemontt, J. F., and Lawrence, C. W. (1989). *REV3*, a *Saccharomyces cerevisiae* gene whose function is required for mutagenesis, is predicted to encode a nonessential DNA polymerase. J. Bacteriol. **171**, 5659-5667.

Morrison, A., Araki, H., Clark, A. B., Hamatake, R. K., and Sugino, A. (1990). A third essential DNA polymerase in *S. cerevisiae.* Cell **62**, 1143-1151.

Mushegian, A. R., and Kooin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc. Natl. Acad. Sci. USA **93**, 10268-10273.

Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. **12**, 861-867.

Navas, T. A., Zhou, Z., and Elledge, S. J. (1995). DNA polymerase epsilon links the DNA replication machinery to the S-phase checkpoint. Cell **80**, 29-39.

Nakayama, N., and Kohiymama, M. (1985). An α-like DNA polymerase from *Halobacterium halobium*. Eur. J. Biochem. **152**, 293-297.

Naktinis, V., Turner, J., and O'Donnell, M. (1996). A molecular switch in a replication machine defined by an internal competition for protein rings. Cell **84**, 137-145.

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

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Mol. Biol. Evol. **3**, 418-426.

Nolling, J., van Eeden, F. J., Eggen, R. I., and de Vos, W. M. (1992). Modular organization of related Archaeal plasmids encoding different restriction-modification systems in *Methanobacterium thermoformicicum*. Nucleic Acids Res. **20**, 6501-6507.

Ochman, H., Gerber, A. S., and Hartl, D. L. (1988). Genetic applications of an inverse polymerase chain reaction. Genetics **120**, 621-623.

O'Donnell, M., Onrust, R., Dean, F. B., Chen, M., Hurwitz, J. (1993). Homology in accessory proteins of replicative polymerases--*E. coli* to humans. Nucleic Acids Res. **21**, 1-3.

Ohno, S. (1970). Evolution by gene duplication. (New York, Springer-Verlag).

Ohno, S. (1973). Ancient linkage groups and frozen accidents. Nature **244**, 259-262.

Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G., and Steitz, T. A. (1985). Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. Nature **313**, 762-766.

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

Overman, L. B., Bujalowski, W., and Lohman, T. M. (1988). Equilibrium binding of *Escherichia coli* single-strand binding protein to single-stranded nucleic acids in the (SSB)65 binding mode. Cation and anion effects and polynucleotide specificity. Biochemistry **27**, 456-471.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. Science **276**, 734-740.

Palmer, J. D. (1997). The mitochondrion that time forgot. Nature **387**, 454-455.

Pearson, W. R. (1990). Rapid and sensitive sequence comparisons with FASTP and FASTA. Methods in Enzymology **183**, 63-98.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U S A **85**, 2444-2448.

Pesole G., Gissi C., Lanave C., and Saccone C. (1995). Glutamine synthetase gene evolution in bacteria. Mol. Biol. Evol. **12**, 189-197.

Piatti, S., Bohm, T., Cocker, J. H., Diffley, J. F. X., and Nasymth, K. (1996). Activation of S-phase-promoting CDKs in late $G_1$ defines a 'point of no return' after which Cdc6 synthesis cannot promote DNA replication in yeast. Genes Dev. **10**, 1516-1531.

Pisani, F. M., De Martino, C., and Rossi, M. (1992). A DNA polymerase from the archaeon *Sulfolobus solfataricus* shows sequence similarity to family B DNA polymerases. Nucleic Acids Res. **20**, 2711-2716.

Philipova, D., Mullen, J. R., and Maniar, H. S., Lu, J., Gu, C., and Brill, S. J. (1996). A hierarchy of SSB protomers in replication protein A. Genes Dev. **10**, 2222-2233.

Prangishvili, D. A. (1986). DNA-dependent DNA polymerases of the thermoacidophilic archaebacterium *Sulfolobus acidocaldarius*. Mol. Biol. USSR **20**, 477-488.

Prangishvili, D., and Klenk, H-P. (1993). Nucleotide sequence of the gene for a 74 kDa DNA polymerase from the archaeon *Sulfolobus solfataricus*. Nucleic Acids Res. **21**, 2768

Prangishvili, D. A., and Klenk, H-P. (1994). The gene for a 74-kDa DNA polymerase from the archaeon *Sulfolobus solfataricus*. Syst. Appl. Microbiol. **16**, 665-671.

Prasartkaew, S., Zijlstra, N. M., Wilairat, P., Prosper Overdulve, J., and de Vries, E. (1996). Molecular cloning of a *Plasmodium falciparum* gene interrupted by 15 introns encoding a functional primase 53 kDa sunbunit as demonstrated by expression in a baculovirus system. Nucleic Acids Res. **24**3934-3941.

Ragan, M. A., Logsdon, J. M. Jr., Sensen, C. W., Charlebois, R. L., and Doolittle, W. F. (1996). An archaebacterial homolog of pelota, a meiotic cell division protein in eukaryotes. FEMS Microbiol. Lett. **144**, 151-155.

Raghunathan, S., Ricard, C. S., Lohman, T. M., and Waksman, G. (1997). Crystal structure of the homo-tetrameric DNA binding domain of *Escherichia coli* single-stranded DNA-binding protein determined by multiwavelength X-ray diffraction on the selenomethionyl protein at 2.9-Å resolution. Proc. Natl. Acad. Sci. USA **94**, 6652-6657.

Rameriz, C., Kopke, A. K. E., Yang, D-C., Boeckh, T., and Matheson, A. T. (1993). The structure, function and evolution of archaeal ribosomes, *In* M. Kates D. J. Kushner, and A. T. Matheson (ed.), The biochemistry of archaea (archaebacteria). (Elsevier Science Publishers, Amsterdam, The Netherlands). pp. 439-466.

Rangarajan, S., Gudmundsson, G., Qiu, Z., Foster, P. L., and Goodman, M. F. (1997). *Escherichia coli* DNA polymerase II catalyzes chromosomal and episomal DNA synthesis *in vivo*. Proc. Natl. Acad. Sci. USA **94**, 946-951

Reysenbach, A. L., Giver, L. J., Wickham, G. S., and Pace, N. R. (1992). Differential amplification of rRNA genes by polymerase chain reaction. Appl. Envrion. Microbiol. **58**, 3417-3418.

Reeck, G. R., de Haën, C., Teller, D. C., Doolittle, R. F., Fitch, W. F., Dickerson, R. E., Chambon, P., McLachlan, A. D., Margoliash, E., Jukes, T. H., and Zuckerkandl, E. (1987). "Homology" in proteins and nucleic acids, a terminology muddle and a way out of it. Cell **50**, 667.

Riera, J., Robb, F. T., Weiss, R., and Fontecave, M. (1997). Ribonucleotide reductase in the archaeon *Pyrococcus furiosus*, a critical enzyme in the evolution of DNA genomes? Proc. Natl. Acad. Sci. USA **94**, 475-478.

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

Roger, A. J. (1996). Studies on the phylogeny and gene structure of early-branching eukaryotes. PhD thesis, Dalhousie University.

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

Rossi, M., Rella, R., Pensa, S., Bartolucci, S., De Rosa, M., Gambacorta, A., Raia, C. A., Dell'Aversano Orabona, N. (1986). Structure and properties of a thermophilic and thermostable DNA polymerase isolated from *Sulfolobus solfataricus*. Syst. Appl. Microbiol. **16**, 337-341.

Rowles, A., and Blow, J. J. (1997). Chromatin proteins invovled in the initiation of DNA replication. Curr. Opin. Genet. Devel. **7**, 152-157.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular cloning, a laboratory manual. Cold Spring Harbour, New York, Cold Spring Harbour Laboratory Press.

Sensen, C. W., Klenk, H.-P., Singh, R. K., Allard, G., Chan, C. C.-Y., Liu, Q. Y., Penny, S. L., Young, F., Schenk, M., Gaasterland, T., Doolittle, W. F., Ragan, M. A., and Charlebois, R. (1996). Organizational characteristics and information content of an archaeal genome, 156 kbp of sequence from *Sulfolobus solfataricus* P2. Mol. Microbiol. **22**, 175-191.

Schinzel, R., and Burger, K. J. (1984). Sensitivity of halobacteria to aphidicolin, an inhibitor of eukaryotic α-type DNA polymerases. FEMS Microbiol. Letters. **25**, 187-190.

Shamoo, Y., Friedman, A. M., Parsons, M. R., Konigsberg, W. H., and Steitz, T. A. (1995). Crystal structure of a replication fork single-stranded DNA binding protein (T4 gp32) complexed to DNA. Nature **376**, 362-366.

Sheaff, R., Ilsely, D., and Kuchta, R. (1991). Mechanism of DNA polymerase α inhibition by aphidicolin. Biochemistry. **30**, 8590-8597.

Shen, W. Y., and Waye, M. M. (1989). A novel method for generating a nested set of unidirectional deletion mutants using mixed oligodeoxynucleotides. Gene **70**, 205-211.

Skinner, M. M., Zhang, H., Leschnitzer, D. H., Guan, Y., Bellamy, H., Sweet, R. M., Gray, C. W., Konings, R. N., Wang, A. H., and Terwilliger, T. C. (1994). Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. Proc. Natl. Acad. Sci. USA **91**, 2071-2075.

Slater, S., Wold, S., Lu, M., Boye ,E., Skarstad, K., and Kleckner, N. (1995). *E. coli* SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. Cell **86**, 927-936.

Stanier, R. A. (1970). Some aspects of the biology of cells and their possible evolutionary significance. *In* Organization and Control of Prokaryotic and Eukaryotic Cells: 20th Symposium of the Society for General Microbioloy, H. P. Charles, and B. C. J. G. Knight eds., (Cambridge University Press, Cambridge). pp. 1-38.

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

Stillman, B. (1994). Smart machines at the replication fork. Cell **78**, 725-728.

Stillman, B. (1996). Cell cycle control of DNA replication. Science **274**, 1659-1664.

Stöffler, G., and Stöffler-Meilicke, M. (1986). Electron microscopy of archaebacterial ribosomes. System. Appl. Microbiol. **7**, 123-130.

Stukenberg, P. T., Studwell-Vaughan, P. S., and O'Donnell, M. (1991). Mechanism of the sliding β-clamp of DNA polymerase III holoenzyme. J. Biol. Chem. **266**, 11328-11334.

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

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

Sorokine, I., Ben-Mahrez, K., Nakayama, M., and Kohiyama, M. (1991). Exonuclease activity associated with DNA polymerases α and β of the archaebacterium *Halobacterium halobium*. Eur. J. Biochem. **197**, 781-784.

Swofford, D. L. (1993). PAUP: phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. (1993) *In* Molecular Systematics, 2nd edition, p. 407-515. D. M. Hillis, G. Moritz, B. K. Mable (ed.) Sinauer Associates, USA.

Tauer, A., and Benner, S. A. (1997). The $B_{12}$-dependent ribonucleotide reductase from the archaebacterium *Thermoplasma acidophila*, an evolutionary solution to the ribonucleotide reductase conundrum. Proc. Natl. Acad. Sci. USA **94**, 53-58.

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

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTALW, improving the sensitivity of progressive multiple sequence alignments through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673-4680.

Tsurimoto, T., and Stillman, B. (1989). Purification of a cellular replication factor, RF-C, that is required for coordinated synthesis of leading and lagging strands during simian virus 40 DNA replication in vitro. Mol. Cell. Biol. **9,** 609-619

Tsurimoto, T., Melendy, T., and Stillman, B. (1990). Sequential initiation of lagging and leading strand synthesis by two different polymerase complexes at the SV40 DNA replication origin. Nature **346,** 534-539.

Uemori, T., Ishino, Y., Toh, K., Adada, I., and Kato, I. (1993). Organization and nucleotide sequence of the DNA polymerase gene from the archaeon *Pyrococcus furiosus*. Nucleic Acids Res. **21,** 259-265.

Uemori, T., Ishino, Y., Doi, H., and Kat, I. (1995). The hyperthermophilic archaeon *Pyrodictium occultum* has two α-like DNA polymerases. J. Bacteriol. **177,** 2164-2177.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. (1982). Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. **1,** 945-951.

Waga, S., and Stillman, B. (1994). Anatomy of a DNA replication fork revealed by reconstitution of SV40 DNA replication *in vitro*. Nature **369,** 207-212.

Wang, J., Sattat, A. K. M. A., Wang, C. C., Karam, J. D., Konigsberg, W. H., and Steitz, T. A. (1997). Crystal structure of a pol α family replication DNA polymerase from bacteriophage RB69. Cell **89,** 1087-1099.

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

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

Woese, C. R. and Fox, G. E. (1978). Archaebacteria. J. Mol. Evol. **11,** 245-252.

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

Wong, S. W., Wahl, A. F., Yuan, P.-M., Arai, N., Pearson, B. E., Arai, K., Korn, D., Hunkapiller, M. W., and Wang, T. S.-F. (1988). Human DNA polymerase α gene expression is cell proliferation dependent and its primary structure is similar to both prokaryotic and eukaryotic replicative DNA polymerases. EMBO J. 7, 37-47.

Xiao, H., Crombie, R., Dong, Z., Onrust, R., O'Donnell, M. (1993). DNA polymerase III accessory proteins. III. holC and holD encoding chi and psi. J. Biol. Chem. **268**, 11773-11778.

Yan, H., Gibson, S., and Tye, B. K. (1991). MCM2 and MCM3, two proteins important for ARS activity, are related in structure and function. Genes Dev. 5, 944-957.

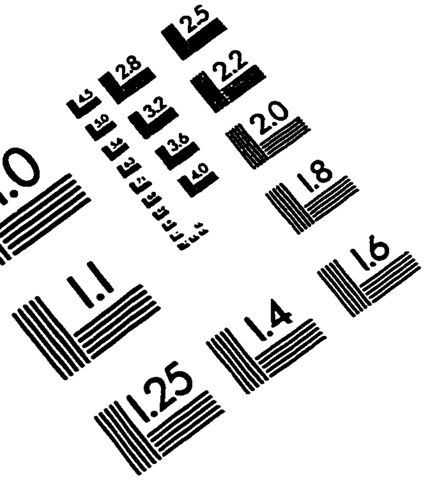
Zabel, H.-P., Fischer, H., Holler E., and Winter, J. (1985). *In vivo* and *in vitro* evidence for eucaryotic α-type DNA polymerases in methanogens. Purification of the DNA polymerase of *Methanococcus vaneilii*. System. Appl. Microbiol. **6**, 111-118.

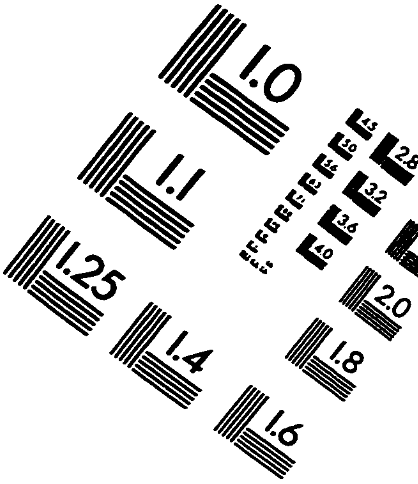
Zabel, H.-P., Holler E., and Winter, J. (1987). Mode of inhibition of the DNA polymerase of *Methanococcus vannielli* by aphidicolin. Eur. J. Biochem. **165**, 171-175.

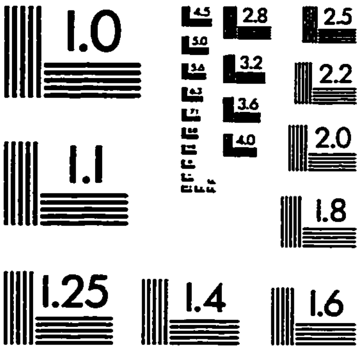
Zillig, W., Palm, P., and Klenk, H-P. (1992). A model of the early evolution of organisms: the arisal of the three domains of life from the common ancestor. *In* The Origin and Evolution of the Cell, H. Hartman, and K. Matsuno, eds., (World Scientific, Singapore). pp. 47-78.

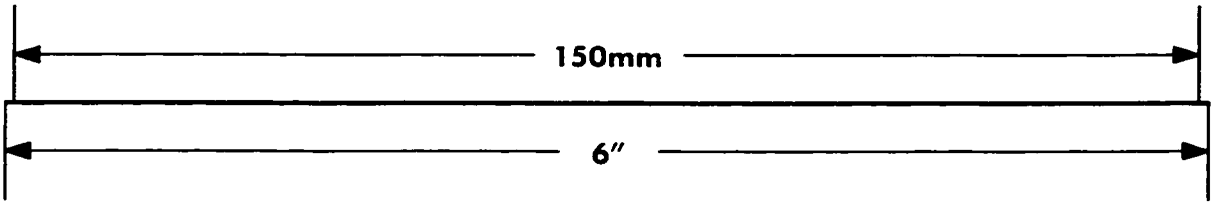
Zillig, W., Palm, P., Klenk, H-P., Langer, D., Hüdepohl, U., Hain, J., Lanzendörfer, and Holz, I. (1993). Transcription in archaea. *In* The Biochemistry of Archaea (Archaebacteria), M. Kates, D. J. Kushner, and A. T. Matheson, eds. (Amsterdam, Elsevier), pp. 367-386.

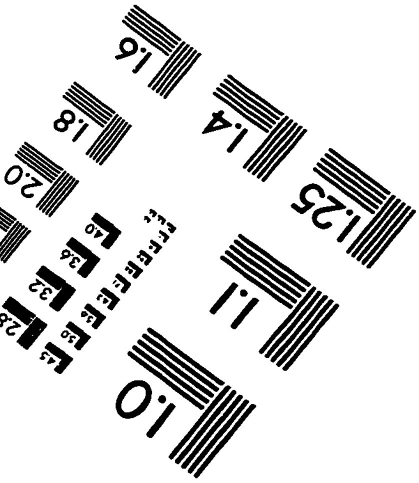
# IMAGE EVALUATION TEST TARGET (QA-3)

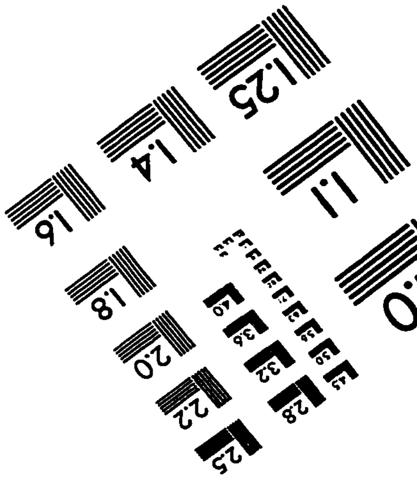












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989