

## **INFORMATION TO USERS**

**This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.**

**The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.**

**In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.**

**Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.**

**Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.**

**ProQuest Information and Learning  
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA  
800-521-0600**

**UMI<sup>®</sup>**



**STUDIES ON THE EVOLUTION OF ARCHAEAL  
AND EUKARYOTIC CHAPERONINS**

**by**

**John M. Archibald**

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

**at**

**Dalhousie University  
Halifax, Nova Scotia  
April, 2001**

**© Copyright by John M. Archibald, 2001**



**National Library  
of Canada**

**Acquisitions and  
Bibliographic Services**

**395 Wellington Street  
Ottawa ON K1A 0N4  
Canada**

**Bibliothèque nationale  
du Canada**

**Acquisitions et  
services bibliographiques**

**395, rue Wellington  
Ottawa ON K1A 0N4  
Canada**

*Your file Votre référence*

*Our file Notre référence*

**The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.**

**The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.**

**L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.**

**L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.**

0-612-66656-5

**Canada**



**DALHOUSIE UNIVERSITY**

**FACULTY OF GRADUATE STUDIES**

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Studies on the Evolution of Archaeal and Eukaryotic Chaperonins"

by John M. Archibald

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

Dated: May 18, 2001

External Examiner  
Research Supervisor  
Examining Committee



DALHOUSIE UNIVERSITY

DATE: May 29<sup>th</sup> / 2001

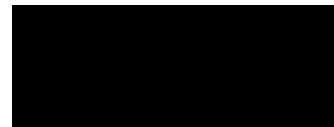
AUTHOR: John M. Archibald

TITLE: Studies on the evolution of archaeal and eukaryotic chaperonins

DEPARTMENT OR SCHOOL: Biochemistry and Molecular Biology

DEGREE: PhD CONVOCATION: October YEAR: 2001

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.



Signature of Author

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in this thesis (other than brief excerpts requiring only proper acknowledgment in scholarly writing), and that all such use is clearly acknowledged.

**Colin Turnbull (1961) took a pygmy friend, Kenge, out of the forest for the first time in his life, and they climbed a mountain together and looked out over the plains. Kenge saw some buffalo 'grazing lazily several miles away, far down below. He turned to me and said. "What insects are those?" ...At first I hardly understood, then I realized that in the forest vision is so limited that there is no great need to make an automatic allowance for distance when judging size. Out here in the plains, Kenge was looking for the first time over apparently unending miles of unfamiliar grasslands, with not a tree worth the name to give him any basis for comparison... When I told Kenge that the insects were buffalo, he roared with laughter and told me not to tell such stupid lies...' (pp. 227-228).**

**Richard Dawkins, The Extended Phenotype, 1982**

## TABLE OF CONTENTS

<b>Table of Contents</b>	v
<b>List of Illustrations</b>	vi
<b>Abstract</b>	viii
<b>Abbreviations and Symbols Used</b>	ix
<b>Acknowledgements</b>	x
<b>Introduction</b>	1
<b>Materials and Methods</b>	18
<b>Chapter I</b> Archaeal Chaperonin Evolution: Recurrent Paralogy	36
<b>Chapter II</b> Origin and Evolution of the Eukaryotic Chaperonin CCT	91
<b>Chapter III</b> Molecular Chaperones Encoded by a Reduced Nucleus—the Cryptomonad Nucleomorph	138
<b>Chapter IV</b> The Chaperonin Genes of 'Jakobid' Flagellates: Implications for Early Eukaryotic Evolution	170
<b>Appendix A</b> Supplementary Data	196
<b>Appendix B</b> Supplementary Alignments	197
<b>References</b>	201

## LIST OF ILLUSTRATIONS

<b>Figure 1</b>	Structural comparison of group I and group II chaperonins	9
<b>Figure 1.1</b>	Southern hybridizations of labeled chaperonin gene fragments to restricted <i>Sulfolobus</i> genomic DNAs	40
<b>Figure 1.2</b>	Alignment of archaeal chaperonin protein sequences	42
<b>Figure 1.3</b>	Phylogenetic analysis of archaeal chaperonins	48
<b>Figure 1.4</b>	An homologous insertion in the $\alpha$ and $\beta$ chaperonin subunits of <i>Pyrodictium occultum</i>	52
<b>Figure 1.5</b>	DNA sequence alignment of a putative gene conversion tract between the <i>P. occultum</i> $\alpha$ and $\beta$ subunit genes	53
<b>Figure 1.6</b>	Sliding window analyses of percent (%) identity shared between duplicate crenarchaeal chaperonin genes	56
<b>Figure 1.7</b>	Sliding window analyses of percent (%) identity shared between duplicate euryarchaeal chaperonin genes	57
<b>Figure 1.8</b>	Maximum likelihood (ML) sliding window analyses of crenarchaeal chaperonin DNA sequences	62
<b>Figure 1.9</b>	Maximum likelihood (ML) sliding window analyses of euryarchaeal chaperonin DNA sequences	65
<b>Figure 1.10</b>	Incongruent phylogenetic signal in the apical domain of crenarchaeal chaperonins	69
<b>Figure 1.11</b>	Testing the removal of the <i>P. occultum</i> and <i>A. pernix</i> sequences on the incongruent phylogenetic signal in crenarchaeal chaperonins	72
<b>Figure 1.12</b>	Putative gene conversion tracts in <i>P. occultum</i> and <i>A. pernix</i> map to the apical domain	74
<b>Figure 1.13</b>	Distribution of amino acid substitutions between the $\alpha$ and $\beta$ paralogs in crenarchaeotes and <i>T. acidophilum</i>	76
<b>Figure 1.14</b>	Evolution of archaeal chaperonin complex architecture	80
<b>Figure 1.15</b>	Archaeal chaperonin evolution by recurrent paralogy	83
<b>Figure 1.16</b>	Hypothetical gene duplication and gene conversion scenarios and their effects on the inference of a known phylogeny	84
<b>Figure 2.1</b>	Alignment of archaeal chaperonins and eukaryotic CCT protein sequences	97
<b>Figure 2.2</b>	Placement of <i>Trichomonas</i> and <i>Giardia</i> chaperonins into known CCT subunit families	107
<b>Figure 2.3</b>	Phylogenetic analysis of group I and group II chaperonin protein sequences	109
<b>Figure 2.4</b>	Phylogeny of group II chaperonin protein sequences	111
<b>Figure 2.5</b>	Testing the position of the root of the eukaryotic CCT tree	114
<b>Figure 2.6</b>	Phylogenetic analysis of CCTzeta protein sequences	116
<b>Figure 2.7</b>	Rates of evolution differ among CCT subunits	117
<b>Figure 2.8</b>	Regions of variability among the different CCT subunits	119
<b>Figure 2.9</b>	Site-rate analysis of CCT subunits	122
<b>Figure 2.10</b>	Subunit-specific 'signatures' in individual CCT apical domains	127

<b>Figure 2.11</b>	Subunit-specific signatures in individual CCT intermediate and equatorial domains	128
<b>Figure 3.1</b>	The <i>hsp70</i> gene from the <i>Guillardia theta</i> nucleomorph	143
<b>Figure 3.2</b>	Phylogenetic analysis of HSP70 protein sequences	145
<b>Figure 3.3</b>	The <i>hsp90</i> gene from the <i>Guillardia theta</i> nucleomorph	149
<b>Figure 3.4</b>	Phylogenetic analysis of HSP90 protein sequences	150
<b>Figure 3.5</b>	The heat shock transcription factor gene from the <i>Guillardia theta</i> nucleomorph	153
<b>Figure 3.6</b>	Phylogenetic analysis of heat shock transcription factor protein sequences	156
<b>Figure 3.7</b>	Phylogenetic analysis of archaeal and eukaryotic chaperonin protein sequences	158
<b>Figure 3.8</b>	Alignment of select CCTalpha protein sequences	160
<b>Figure 3.9</b>	Alignment of select CCTbeta protein sequences	161
<b>Figure 3.10</b>	Alignment of select CCTgamma protein sequences	162
<b>Figure 3.11</b>	Alignment of select CCTdelta protein sequences	163
<b>Figure 4.1</b>	Spliceosomal introns in the <i>Reclinomonas</i> and <i>Malawimonas</i> CCT genes	176
<b>Figure 4.2</b>	Alignment of CCTalpha protein sequences	177
<b>Figure 4.3</b>	Alignment of CCTdelta protein sequences	180
<b>Figure 4.4</b>	Phylogenetic analysis of CCTalpha protein sequences	184
<b>Figure 4.5</b>	CCTalpha phylogeny with the parabasalid sequences removed	186
<b>Figure 4.6</b>	Rooted analyses of CCTalpha	188
<b>Figure 4.7</b>	CCTalpha and CCTdelta intron phylogenies	189
<b>Figure A.1</b>	Rooting the archaeal chaperonin tree	196
<b>Figure B.1</b>	Alignment of select CCTepsilon protein sequences	197
<b>Figure B.2</b>	Alignment of select CCTzeta protein sequences	198
<b>Figure B.3</b>	Alignment of select CCTeta protein sequences	199
<b>Figure B.4</b>	Alignment of select CCTtheta protein sequences	200

## ABSTRACT

Chaperone-assisted protein folding is a universal cellular process. The chaperonins are a class of evolutionarily related molecular chaperones that form multisubunit double-ring complexes and facilitate protein folding through the hydrolysis of ATP. This thesis examines (primarily) the role of gene duplication in the origin and evolution of chaperonins in Archaea and the eukaryotic cytosol.

The polymerase chain reaction (PCR) was used to isolate chaperonin genes from thermophilic archaea. Phylogenetic analyses of archaeal chaperonins reveal a complex pattern of gene duplication, gene conversion and gene loss, and suggest that hetero-oligomeric chaperonin complexes have evolved multiple times independently during the history of this group. A novel chaperonin subunit-encoding gene was also isolated from two species of *Sulfolobus*.

Eukaryotic cytosolic chaperonin genes were isolated from two putative early-diverging amitochondriate protists, *Trichomonas vaginalis* and *Giardia lamblia*. In contrast to the lineage-specific pattern of gene duplication observed in archaea, numerous duplications took place early in eukaryotic evolution and produced eight distinct chaperonin paralogs, prior to the diversification of all eukaryotes under investigation. Further studies revealed significant differences in the rates of amino acid sequence evolution of eukaryotic chaperonins compared to those in Archaea, and among the different eukaryotic chaperonins themselves.

Molecular evolutionary studies were also performed on chaperones encoded in the 'nucleomorph' genome of the cryptomonad alga, *Guillardia theta*. The *G. theta* nucleomorph—the remnant nucleus of a photosynthetic eukaryotic endosymbiont—was found to retain genes encoding cytosolic forms of HSP70, HSP90 and eight cytosolic chaperonin subunits. Striking differences in the degree of conservation of the various nucleomorph-encoded molecular chaperones were observed, suggesting reduced (or different) evolutionary constraints on the functions of the chaperones in this unusual cellular context.

Finally, the taxonomic sampling of eukaryotic cytosolic chaperonin genes was expanded to include the 'jakobid' flagellates, an enigmatic group of free-living, mitochondriate protists. Unlike most protein-coding genes in protists, the chaperonin genes of two jakobids, *Reclinomonas americana* and *Malawimonas jakobiformis*, were found to possess numerous spliceosomal introns. An analysis of the intron positions in these genes from a wide diversity of eukaryotes suggests that many of the intron-sparse or intron-lacking protist lineages have lost spliceosomal introns during their evolutionary history.

## ABBREVIATIONS AND SYMBOLS USED

<b>Å</b>	<b>Ångstrom</b>
<b>BLAST</b>	<b>basic local alignment search tool</b>
<b>bp</b>	<b>base pair</b>
<b>CCT</b>	<b>chaperonin-containing TCP-1</b>
<b>cDNA</b>	<b>complementary DNA</b>
<b>CTAB</b>	<b>cetyltrimethylammonium bromide</b>
<b>EDTA</b>	<b>ethylene-diamine-tetra-acetic acid</b>
<b>ER</b>	<b>endoplasmic reticulum</b>
<b>EST</b>	<b>expressed sequence tag</b>
<b>gDNA</b>	<b>genomic DNA</b>
<b>HSP</b>	<b>heat shock protein</b>
<b>kb</b>	<b>kilobase pair</b>
<b>kDa</b>	<b>kilodalton</b>
<b>LB</b>	<b>Luria Bertani</b>
<b>ORF</b>	<b>open reading frame</b>
<b>PCR</b>	<b>Polymerase chain reaction</b>
<b>pI</b>	<b>isoelectric point</b>
<b>SDS</b>	<b>sodium dodecyl sulfate</b>
<b>SSUrRNA</b>	<b>small subunit ribosomal RNA</b>
<b>TCP-1</b>	<b>t-complex polypeptide 1</b>



## ACKNOWLEDGEMENTS

There are many people I want to thank for their personal and professional help during the past four years.

First, I would like to thank Ford for giving me the opportunity to do graduate school in his laboratory. The Doolittle lab is special—Ford is friends with all his students and postdocs, and this fosters a relaxed, fun and freewheeling research environment. I always enjoyed debating scientific and philosophical issues with Ford, and his light-hearted approach to day-to-day life made the tedium of graduate school more bearable.

Ford's lab is filled with smart, interesting and enthusiastic people, and I have benefited enormously from the guidance and advice of a large number of graduate students and postdocs. I especially want to thank Naomi Fast and John Logsdon who were extremely generous with their time during the first two years I was in the lab, struggling to carve out a niche. Naomi helped me learn basic molecular biology and molecular evolution, and was a great teacher, colleague and friend. John taught me a huge amount about phylogenetic reconstruction, and was incredibly patient and generous with his time when I struggled to understand difficult concepts. John was also an excellent collaborator on much of the work presented in this thesis, and was the one who convinced me to pursue the chaperonin work in the first place! His extraordinary attention to detail re-defined for me the concept of a 'rigorous analysis'. Thanks John. David Edgell and David Faguy were always willing to give advice and were excellent lab companions in the early days. Sandie Baldauf and Arlin Stoltzfus were generous with time and advice, and I learned a great deal from watching them 'do science'. Oisín Feeley was a good friend and lab companion who was always willing to listen to, and try to understand, almost any scientific issue.

I also want to thank Andrew Roger for inspiring me during my first two years of graduate school—he now continues to do so from the lab next door! Andrew’s knowledge of protistology and phylogenetic inference is extraordinary, and the research described in this thesis benefited greatly from his input. I am especially grateful for his help and advice during the decision-filled final year of my thesis. Jan Andersson, Yan Boucher, Joel Dacks, Yugi Inagaki, Camilla Nesbø, Alastair Simpson and Jeff Silberman were excellent lab colleagues, and I learned from them individually and as a group. I’ll miss our daily coffee runs. Geoff Morris worked with me on the jakobid chaperonin project as a summer student, and could not have been a better trainee. He worked hard on what turned out to be an extremely difficult project. Paul Briggs was always willing to help with my never-ending stream of computer problems, and Roisin McDevitt-MacKenzie and Chris Gerbert were enormously helpful when it came to administrative details.

I thank Charley O’Kelly, Sue Douglas, Michel Leroux and Christian Blouin for being great collaborators. I learned a lot from Charley and Sue about protistology, and Michel was instrumental in helping me elucidate and sharpen some of the more theoretical aspects of this thesis. What began as a no-holds-barred email debate about chaperonin evolution developed into a collaboration. I benefited from having gone through the thesis writing / defense process shortly after Christian, and he was an invaluable colleague when it came to the protein structure/function aspects of my research.

Mike Charette has been a great friend and roommate over the years. We were stricken by the ‘science bug’ at about the same time as undergraduates, and decided to move across campus and do PhDs in adjacent labs! Mike has always been generous with his time, listening to me rant about lab problems or

participating in 'kitchen talks' on all sorts of issues (scientific and otherwise), often until the wee hours of the morning. Thanks Mike.

Finally, I thank my family for their support and advice over the years. My parents never judged my (sometimes dubious) decisions about how to live my life, but simply encouraged me, no matter what. My sisters have also been supportive, despite our differences. Most of all, I thank Shauna for her love, patience and support through the ups and downs that have defined my existence since we met. I tried to keep the hard work and long hours of graduate school in perspective by reminding myself of the fact that most of the time she was working harder than I was—and smiling about it! This thesis is for her.

## INTRODUCTION

In the 1950s and 1960s, Christian B. Anfinsen performed landmark experiments demonstrating that some pure denatured proteins can refold spontaneously into their biologically active three-dimensional conformations upon the removal of denaturant and in the absence of additional macromolecules or energy. These experiments contributed to the formulation of the principle of protein 'self-assembly'—the idea that all of the information necessary for a protein to fold into its native structure resides within the linear sequence of amino acids. The notion of self-assembly provided the intellectual framework for much of the research devoted to the issue of protein folding for the next twenty years.

The universality of the principle of self-assembly has been challenged by the discovery of a ubiquitous class of proteins, the *molecular chaperones*, upon which the proper folding of many proteins *in vivo* appears to be completely dependent. The term 'molecular chaperone' was first used by Laskey *et al.* (1978) in reference to the biochemical properties of nucleoplasmin, a protein required for the proper formation of nucleosome cores from DNA and histones in *Xenopus* eggs. While this protein is clearly essential for nucleosome *assembly*, it does not form part of the nucleosome itself. Nucleoplasmin is thought to bind to the histone proteins and neutralize their positive charges (Laskey *et al.* 1978). This binding promotes histone-histone interactions by reducing the electrostatic repulsion between them and minimizes the formation of insoluble aggregates by competing with the electrostatic attraction between the histones and DNA (Ellis and Hemmingsen 1989).

The term 'molecular chaperone' was later used by Ellis (1987) to describe the function of a protein involved in the assembly of ribulose 1,5-bisphosphate

carboxylase-oxygenase (Rubisco), the multi-subunit carbon dioxide-fixing enzyme in chloroplasts and cyanobacteria. This molecule, referred to as the Rubisco subunit binding protein (RsuBP or RBP), appeared to mediate the assembly of the Rubisco complex by binding transiently and noncovalently to the Rubisco large subunit (Barraclough and Ellis 1980). Molecular chaperones were thus originally defined "...as a class of cellular proteins whose function is to ensure that the folding of certain other polypeptide chains and their assembly into oligomeric structures occur correctly" (Ellis 1987). "The term 'molecular chaperone' seems appropriate because the traditional role of the human chaperone, if described in biochemical terms, is to prevent improper interactions between potentially complementary surfaces and to disrupt any improper liaisons that may occur" (Ellis and Hemmingsen 1989). More than 20 different protein families now appear to fit this broad definition, and others are continually being discovered (Ellis 1997). Nevertheless, it is only quite recently that the apparent 'chaperone' activities of nucleoplasmin and RBP were seen as anything but curious exceptions to the general rule of protein self-assembly, largely due to the prevailing (and, in hindsight, incorrect) view that protein folding *in vitro* mimics protein folding *in vivo*.

### **Why do cells need molecular chaperones?**

It was clear from Anfinsen's work (1973) that a successful *in vitro* refolding experiment required that the concentration of protein in the reaction be relatively low. Completely or partially unfolded proteins are inherently insoluble, due to the fact that they expose hydrophobic surfaces that in the native state are buried inside the protein. At increasing concentrations, the probability of inappropriate interactions between exposed hydrophobic surfaces increases, and aggregation becomes likely (Martin, Mayhew and Hartl 1996).

The concentration of macromolecules inside living cells is, of course, extraordinarily high. It follows that during and after translation and translocation, potentially interactive surfaces on unfolded polypeptide chains are transiently exposed to the intracellular environment and can interact non-productively with a large number of proteins. Molecular chaperones are thought to prevent this from occurring. Most chaperones appear to bind their substrates through hydrophobic-hydrophobic interactions, although the exact mechanism by which they perform this function is, according to the original definition, irrelevant.

Some molecular chaperones are classified as 'stress' or 'heat shock' proteins, due to the fact that they undergo a dramatic increase in expression when cells are exposed to increased temperature. In this context, molecular chaperones are thought to bind to and assist the refolding of heat-denatured proteins. While these proteins clearly play an important role in the heat shock response, it is also clear that they perform essential functions under normal growth conditions (Hendrick and Hartl 1993).

### **The chaperonins: a unique class of chaperone**

The discovery of the *chaperonins* serves as an excellent example of how science often proceeds. The elucidation of this particular class of molecular chaperone was the result of a merger between two independent lines of investigation—one, genetic experiments on bacteriophage  $\lambda$  morphogenesis, the other, biochemical studies on the synthesis of Rubisco in plant chloroplasts. Ellis (1996) and, more recently, Lorimer (2001) provide excellent personal accounts of the early history of the field.

In the 1970s, Georgopoulos and co-workers were studying temperature-sensitive mutant strains of *Escherichia coli* that were unable to support the growth

of  $\lambda$  and several other bacteriophages. Using genetic selections designed to detect interactions between host and bacteriophage-encoded gene products, numerous *E. coli* genes were identified as being essential for phage  $\lambda$  growth at all temperatures and for bacterial growth at high temperatures (Friedman *et al.* 1984). One of the genes identified in these screens was given the name *groE*, due to the fact that the  $\lambda$  growth defect could be overcome by a mutation in the phage-encoded head protein gpE. The *groE* gene was found to encode a protein of approximately 65 kDa that possessed weak ATPase activity and formed unusual oligomeric structures exhibiting seven-fold rotational symmetry (Georgopoulos and Hohn 1978; Hendrix 1979; Hohn *et al.* 1979). Another *groE* host factor proved to be a much smaller protein of about 15 kDa: it was given the name GroES, and the larger *groE* gene product was named GroEL. These two proteins were eventually shown to interact *in vitro* and *in vivo*, and like GroEL, the GroES protein formed oligomeric structures (Chandrasekhar *et al.* 1986). For the most part, research was focussed on elucidating the exact role of the GroEL and GroES proteins in  $\lambda$  phage assembly. Little thought was given to what these proteins might be doing in the cytoplasm of uninfected cells (Ellis 1996).

At about the same time (and as mentioned above), ongoing experiments by Ellis and colleagues pointed to RBP as an important protein in the assembly of the chloroplast Rubisco complex. 'Higher' plant Rubisco is composed of eight large subunits, synthesized in the chloroplast, and eight small subunits, made in the cytosol and imported into the organelle. Barraclough and Ellis (1980) proposed that the transient binding of an oligomeric form of RBP to the Rubisco large subunit might be an obligatory step in the assembly process, keeping the newly synthesized proteins soluble until they could be incorporated into the Rubisco holoenzyme. This was based in part on the observation that purified Rubisco large subunit had a tendency to form insoluble aggregates. The formal

suggestion by Ellis in 1987 that RBP might act as a molecular chaperone, in a manner similar to nucleoplasmin, was met with limited enthusiasm. It wasn't until the chloroplast RBP and *E. coli* GroEL were sequenced—and found to be 50% identical at the amino acid level—that their respective biochemical roles were suggested to exemplify a more general cellular process. A Canadian biochemist, Sean Hemmingsen, proposed that this class of evolutionarily related molecular chaperones be called 'chaperonins' (Hemmingsen *et al.* 1988).

Goloubinoff *et al.* (1989) were the first to confirm the proposed activity of chaperonins by performing so-called 'neo-Anfinsen' *in vitro* protein refolding experiments. These were identical to the experiments of Anfinsen except that purified chaperonin (GroEL and GroES) was added to the refolding buffer. Goloubinoff *et al.* successfully demonstrated that the omission of GroEL, GroES or MgATP from the reaction resulted in a failure to reconstitute Rubisco activity from chemically denatured substrate. Since then, dozens of other proteins have been assayed in this manner and shown to be similarly dependent on the chaperonins for proper folding.

In the last decade, exhaustive structure/function studies have elucidated the basic features of the GroEL/ES system. As originally shown by Hendrix (1979) and Hohn *et al.* (1979), the *E. coli* GroEL oligomer is a double ring structure, with each ring containing seven identical GroEL monomers. Protein folding occurs inside the central cavity, and the GroES oligomer (a single ring with seven subunits) acts as a cap or lid, shielding the encapsulated substrate from the intracellular environment (Sigler *et al.* 1998). ATP binding and hydrolysis is an essential feature of the protein folding reaction. The first crystal structure of GroEL was resolved to 2.8Å by Braig *et al.* (1994), and the structure of the GroEL/ES complex was later determined (Xu, Horwich and Sigler 1997). Mutagenesis studies have identified numerous amino acid residues in GroEL



that are essential for substrate binding, GroES binding and ATP hydrolysis (Fenton *et al.* 1994).

### **Group II chaperonins**

An exciting development in the study of chaperonins has been the discovery that there are two distinct sub-classes ('groups'): GroEL and its eukaryotic organellar relatives (group I), and a collection of distinct but clearly homologous proteins in Archaea (Archaeobacteria) and the cytosol of eukaryotes (group II). The existence of group II chaperonins was first predicted on the basis of protein sequence similarity. GroEL, RBP and mitochondrial heat shock protein 60 (hsp60 or cpn60) were found to share significant amino acid sequence identity with a eukaryotic cytosolic protein of unknown function called *t*-complex polypeptide-1, or TCP-1 (Ahmad and Gupta 1990; Ellis 1990; Gupta 1990). At about the same time, Phipps *et al.* (1991, 1993) described an abundant double-ring protein complex present in the cytoplasm of the hyperthermophilic archaeon *Pyrodictium* that looked suspiciously similar to GroEL. Like GroEL, the complex accumulated to extremely high levels upon heat shock (in this case, a shift from 102C to 108C!) and possessed ATPase activity. It was not, however, a homo-oligomer, but was composed of two distinct subunit species of 56 and 59 kDa.

Trent and co-workers were also studying the heat-shock response in archaea. They observed that a polypeptide of 55 kDa was virtually the only protein expressed in *Sulfolobus* cells at extreme temperatures (Trent, Osipiuk and Pinkau 1990) and showed that it was present *in vivo* as part of a multi-subunit complex similar to those observed in *Pyrodictium* (Trent *et al.* 1991). Not surprisingly, the sequence of the *Sulfolobus* 'heat shock' protein confirmed it as a member of the chaperonins. What was surprising was that it appeared to be only

distantly related to GroEL. It was far more similar (approximately 40% identical) to TCP-1 (Trent *et al.* 1991).

By the late 1970s, TCP-1 was known to be an abundant protein in mouse testis, with its gene mapping to the t-complex locus on chromosome 17 (Silver, Artzt and Bennett 1979; Willison, Dudley and Potter 1986). Its presence had also been confirmed in human, *Drosophila* and *Saccharomyces* (Ursic and Culbertson 1991; Ursic and Ganetzky 1988; Willison *et al.* 1987), and mutations in the yeast TCP-1 gene seemed to be associated with a variety of microtubule-related abnormalities (Ursic and Culbertson 1991). The possibility that TCP-1 might be the cytosolic equivalent of *E. coli* GroEL sparked a flurry of investigation in the early 1990s. A report in *Nature* by Lewis *et al.* (1992) showed that TCP-1 was indeed part of a chaperonin-like 800-950 kDa oligomeric particle in mouse and human cytosol. In the very same issue, a paper by Yaffe *et al.* (1992) showed that (what was very likely) the same particle was involved in tubulin biogenesis, and Gao *et al.* (1992) reported (in *Cell*) that the complex also catalyzed the folding of  $\beta$ -actin. The eukaryotic cytosolic chaperonin had been identified, and a role in the folding of cytoskeletal proteins had been established.

The story got even more interesting in the mid-1990s. Lewis *et al.* (1992) had shown previously that, in addition to TCP-1, the chaperonin particle contained at least four other unidentified proteins of similar size. Franz-Ulrich Hartl and colleagues had also identified TCP-1 as part of a hetero-oligomeric chaperonin complex, which they referred to as 'TRiC' (for TCP-1 ring Complex). However, they went one step further and showed (by partial sequence analysis) that at least three of the 'unidentified' proteins were homologous to TCP-1 (Frydman *et al.* 1992). Ultimately, the eukaryotic cytosolic chaperonin complex proved to contain eight distinct TCP-1-like proteins (Kubota *et al.* 1994; Kubota, Hynes and Willison 1995b; Miklos *et al.* 1994; Rommelaere *et al.* 1993). The

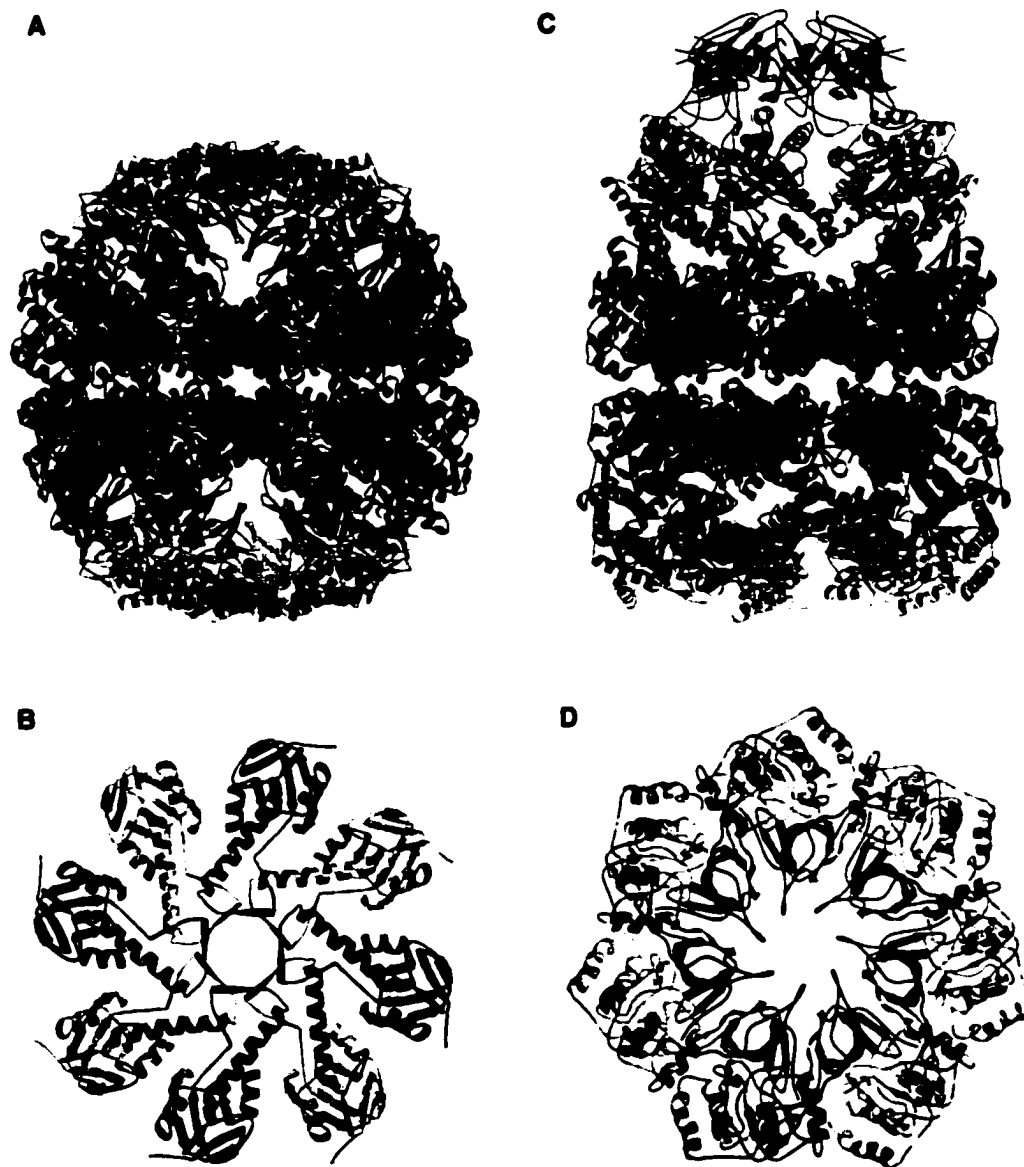
particle is now most often referred to as CCT (for chaperonin-containing TCP-1; Kubota *et al.* 1994), and the original TCP-1 protein is designated the CCT $\alpha$  subunit.

Conspicuously absent from the archaeal and eukaryotic (i.e., group II) chaperonin complexes was any sign of a homolog of the co-chaperonin GroES, the single-ring oligomer that seals off the folding chamber in the bacterial system. The crystal structure of a group II chaperonin suggested a possible reason why. The chaperonin complex from the archaeon *Thermoplasma acidophilum* is composed of eight-membered rings with alternating  $\alpha$  and  $\beta$  subunits, each of which possesses an extended 'helical protrusion' (Ditzel *et al.* 1998; Klumpp, Baumeister and Essen 1997; Nitsch *et al.* 1997). Interestingly, the protrusions exist in precisely the region of the molecule where the co-chaperonin GroES interacts with GroEL. Thus the group II chaperonin complex appears to possess a 'built-in lid' that is thought to control access to the central cavity in a manner analogous to the detachable GroES (Horwich and Saibil 1998). This important difference aside, the two classes of chaperonin have proved to be extraordinarily similar (Figure 1).

### **The tree of life**

The discovery that archaeal chaperonins were most closely related to a component of the eukaryotic cytosol came as a surprise to many. From an evolutionary perspective, it made perfect sense.

Prior to the late 1970's, the living world was viewed by most cell biologists as a dichotomous one: cells were either prokaryotic or eukaryotic. Prokaryotes were characterized primarily by the absence of eukaryote-specific features such as a nucleus, membrane-bound organelles and a complex internal cytoskeleton (Stanier 1970; Stanier and van Niel 1962). The advent of widespread small



**Figure 1** Structural comparison of group I and group II chaperonins. (A and B) Group II chaperonin from *Thermoplasma acidophilum* (the thermosome; Ditzel *et al.*, 1998). (C and D) The group I chaperonin from *Escherichia coli* (GroEL/ES complex; Xu *et al.*, 1997). The chaperonin oligomers are shown in side views (top row) and top views (bottom row). In the side views, the different domains of the individual subunits are colored (red, equatorial domains; green, intermediate domains; yellow, apical domains). In (C and D), the GroES co-chaperonin is colored aqua. To distinguish the monomers from the chaperonin complexes, a single subunit in the upper ring of each side view is colored differently. Figures kindly provided by S. Steinbacher. Reprinted from Cell, 93, Ditzel, L., Löwe, J., Stock, D., Stetter, K.-O., Huber, H., Huber, R., and Steinbacher, S., Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT, 125-138, Copyright 1998, with permission from Elsevier Science.

subunit ribosomal RNA (SSUrRNA) gene sequencing led to the discovery that prokaryotes were in fact comprised of two distinct groups, eubacteria (bacteria) and archaebacteria (archaea). This 'three-domain' view of life, championed by Woese and co-workers (Woese 1987; Woese and Fox 1977), caught on slowly but eventually transformed the fields of molecular and cellular evolution.

While SSUrRNA phylogenies clearly illustrated the distinctness of bacteria, archaea and eukaryotes, they could not determine the relative branching order among the three domains. In the absence of additional evidence, there was no way of deciding which group was ancestral to the other two (i.e., to determine the position of the 'root'). Such evidence eventually came in 1989 in the form of 'reciprocally-rooted' phylogenetic trees constructed from protein sequences. The elongation factors EF-Tu/EF-1 $\alpha$  and EF-G/EF-2 are present in all three domains and are distantly related paralogs: they evolved from one another by gene duplication in the common ancestor of bacteria, archaea and eukaryotes. Because of this, phylogenetic trees can be constructed in which each 'subtree' (e.g., EF-Tu/EF-1 $\alpha$ ) can be used to determine the position of the root to the other (EF-G/EF-2). Iwabe *et al.* (1989) did precisely this, as did Gogarten *et al.* (1989), using an ancient duplication in the proton-pumping ATPase family. The results suggested that archaea and eukaryotes shared a more recent common ancestor than either did with bacteria. Numerous other 'rooting' experiments with larger datasets and additional molecules have supported the notion of an archaea/eukaryote sisterhood (Baldauf, Palmer and Doolittle 1996; Brown and Doolittle 1995; Brown *et al.* 1997; Gribaldo and Cammarano 1998; Lawson, Charlebois and Dillon 1996, but see Brinkmann and Philippe 1999; Philippe and Forterre 1999).

## Eukaryotic evolution

The Archaea immediately became the focus of considerable attention, particularly for molecular evolutionists with an interest in the origin of eukaryotes. Archaea seemed like the best place to search for the molecular antecedents of much of the modern-day complexity of eukaryotic cells. In some ways, this turned out to be true. Many of the features of archaea, such as the transcription and translation apparatus, do appear to be more 'eukaryotic' than 'bacterial' (reviewed extensively by Brown and Doolittle 1997). Nevertheless, the gulf between the prokaryotic and eukaryotic grade of cellular organization is enormous.

Others have focussed on searching for transitional forms within the eukaryotic domain. Cavalier-Smith's 'Archezoa hypothesis' (1983a, 1983b) has been the impetus for the collection of enormous amounts of ultrastructural and molecular data aimed at elucidating some of the early events in eukaryotic evolution. The hypothesis posits that descendants of primitive eukaryotic lineages might still exist on earth, lineages that diverged from other eukaryotes *prior* to the bacterial endosymbioses that gave rise to chloroplasts and mitochondria. The study of such lineages could provide important cellular and molecular clues to the genesis of eukaryote-specific cellular and molecular features.

A sub-kingdom within eukaryotes, the Archezoa, was erected. A variety of amitochondriate protist groups were originally put forth as members of the Archezoa, including the Parabasalia (e.g., the Trichomonad *Trichomonas vaginalis*), the Diplomonadida (e.g., the diplomonad *Giardia lamblia*) and the Microsporidia (e.g., *Nosema locustae*). In addition to their lack of identifiable mitochondria or chloroplasts, these lineages shared as their uniting feature a general simplicity in cellular architecture, which was suggested to be a primitive

feature. Gratifyingly, the first molecular data obtained from these organisms was consistent with this notion. The 5.8S and 28S ribosomal RNA (rRNA) genes of the microsporidian *Variamorpho necatrix* were found to be fused, like those in bacteria but unlike other eukaryotes (Vossbrinck and Woese 1986). Even more exciting was the fact that the first phylogenetic analyses of SSUrRNA sequences from archezoa placed these organisms at the base of eukaryotes (Leipe *et al.* 1993; Sogin 1989; Sogin *et al.* 1989; Vossbrinck *et al.* 1987), consistent with their supposed primitively amitochondriate status. Early protein-based phylogenies also suggested that *Trichomonas*, *Giardia* and *Glugea* (another microsporidian) were among the deepest branches on the eukaryotic tree (Hashimoto *et al.* 1994; Kamaishi *et al.* 1996; Stiller, Duffield and Hall 1998).

The exact membership of the Archezoa has changed over the years, largely in response to the analysis of new molecular sequence data and a growing appreciation of the potential problems and pitfalls of the phylogenetic analysis of divergent molecular sequences. Perhaps the most striking example is that of the Microsporidia. Once considered *bona fide* archezoa, phylogenies constructed from tubulin, Hsp70 and RNA polymerase (Germot, Philippe and Le Guyader 1997; Hirt *et al.* 1999; Keeling and Doolittle 1996; Keeling, Luker and Palmer 2000) strongly suggest that microsporidia are not extant ancestors of an early-diverging lineage but are, in fact, relatives of fungi. Their deep placement in some phylogenetic trees appears to be artifactual, due to an accelerated rate of sequence evolution. With respect to the question of the presence or absence of mitochondria in these organisms, a fungal origin for the microsporidia strongly implies secondary mitochondrial loss, as such a position on the eukaryotic tree means that they are nested within mitochondrion-containing groups.

The discovery of mitochondrial-type molecular chaperones in several archezoan lineages has provided further evidence against the idea that these

organisms are primitively amitochondriate. Most notably, *cpn60* and/or *HSP70* genes of probable mitochondrial origin have been isolated from the nuclear genomes of the parabasalid *Trichomonas vaginalis* (Germot, Philippe and Le Guyader 1996; Roger, Clark and Doolittle 1996), the diplomonad *Giardia lamblia* (Morrison *et al.* 2001; Roger *et al.* 1998), as well as several different microsporidia (Germot, Philippe and Le Guyader 1997; Hirt *et al.* 1997; Peyretailade *et al.* 1998). At present, the data are consistent with the possibility that none of the 'deep-branching' protist groups currently under investigation are living descendents of a primitively amitochondriate phase of eukaryotic evolution.

### **General rationale: the comparative method**

Phylogenetic artifacts and shattered hypotheses aside, the former members of the archezoa are an extraordinarily diverse collection of eukaryotes. As such, they are key to the study of molecular evolution—they are as evolutionarily distant from animals and fungi as any eukaryotic cells presently known.

Much of what we know about the biochemistry and evolutionary history of the chaperonins comes from the study of surprisingly few 'model' organisms. The approach taken in this thesis is to study the evolution of group II chaperonins by comparing protein sequences encoded by orthologous and paralogous genes from a broad diversity of archaea and eukaryotes. Such comparisons allow the identification of universally conserved amino acid residues, poorly conserved residues, and residues that are differentially conserved in particular subsets of chaperonin sequences. The identification of evolutionarily conserved residues is not trivial: these residues are likely to be important for protein structure and function, having been maintained over large time-scales by the process of purifying selection. Amino acid residues that are



unique to (and highly conserved within) a particular subset of chaperonins can elucidate changes in function that have occurred after protein sequences have diverged from one another.

### **Gene duplication and the evolution of 'new' genes and proteins**

The importance of gene duplication as a mechanism for generating evolutionary novelty has long been recognized. Ohno's (1973) classic model for the origin of new genes proposed that after duplication, one copy of a duplicate gene is freed from functional constraint and thus allowed to acquire mutations (both synonymous and non-synonymous) at random. Such mutations would be selectively neutral and would most often result in the degeneration of the duplicate. *Occasionally*, however, a series of mutations might result in the acquisition of a new function. Hughes and Hughes (1993) have shown that the substitution pattern observed between duplicate protein-coding genes in *Xenopus* is inconsistent with Ohno's prediction, and instead indicates that purifying selection usually acts to eliminate amino acid differences between duplicate proteins.

A more realistic model for the evolution of duplicate genes invokes the concept of 'gene sharing'. Simply put, if a single gene encodes a protein that has two (or more) functions, gene duplication can result in a situation where each of the daughter genes assumes one (or more) of these functions—the genes become 'specialized'. Under this model, a period in which one of the duplicates is completely freed of any function is not required (Hughes 1994). The phenomenon of gene sharing has been documented in the evolution of the eye lens crystallins in animals (Piatigorsky and Wistow 1991).

The chaperonins are a unique system with which to study gene duplication, for two reasons. First, duplicate chaperonin subunits often function

in the same oligomeric complex, where they can co-evolve with one another as well as with their substrates. Second, when chaperonin genes from a wide diversity of organisms are considered, a huge range in the degree of sequence divergence between duplicates is observed: some duplications appear to have occurred quite recently, others are extremely ancient. It is thus possible to try to understand the *origin* of duplicate chaperonin genes, and to elucidate the evolutionary forces mediating the probability of their fixation, by comparing and contrasting these two extremes.

This thesis deals primarily with the role of gene duplication in the evolution of group II chaperonins. In the first Chapter, I describe the isolation of chaperonin genes from a variety of 'crenarchaeote' archaea and present a comprehensive analysis of chaperonin protein sequence evolution. I also develop and implement a novel phylogenetic method for scanning DNA and protein sequences for regions of anomalous phylogenetic signal, such as those produced by gene conversion.

The bulk of the work presented in Chapter II describes my attempts to determine when during eukaryotic evolution the gene duplications producing the different CCT (TCP-1) subunit genes occurred, and how the different subunits are related to one another. I also present a site-by-site analysis of the distribution of slowly-evolving amino acid positions in the different CCT subunits, and attempt to shed light on the significance of these observations by mapping the positions of subunit-specific residues onto the crystal structure of an archaeal homolog of CCT. From a theoretical perspective, the work in Chapter II is directly related to that described in Chapter I; as the reader will see, an understanding of the pattern of chaperonin gene duplication observed in archaea is important for understanding the evolution of the eukaryotic chaperonin CCT.

Chapter III presents the results of a phylogenetic analysis of several different molecular chaperones encoded in the 'nucleomorph' genome of the cryptomonad alga, *Guillardia theta*. Cryptomonads are unusual cells. They have acquired photosynthesis secondarily by engulfing and retaining a photosynthetic *eukaryote*, and have kept the cytoplasm and nucleus (the nucleomorph) of the endosymbiont. The nucleomorph genome is highly reduced, and the study of its genes can shed light on the basic set of genes and proteins necessary for maintaining basic eukaryotic cellular processes. Huge differences in the degree of evolutionary conservation were observed among the different classes of nucleomorph-encoded molecular chaperones. As well, many of the co-chaperones and co-factors necessary for protein folding in the 'standard' eukaryotic cytoplasm were absent, suggesting that there may be more flexibility in the exact components of chaperone-assisted protein folding pathways than previously appreciated.

Finally, the analyses presented in Chapter IV expand the taxonomic sampling of eukaryotic CCT genes to include an enigmatic group of mitochondriate protists known as the 'jakobid' flagellates. Jakobids are free-living, heterotrophic flagellated cells found in a variety of marine and fresh-water habitats. These organisms have the most bacterial-like mitochondrial genomes discovered in any eukaryotic cell thus far (Gray *et al.* 1998; Lang *et al.* 1997). Jakobids also possess ultrastructural features that appear to ally them with certain amitochondriate protist groups. If these amitochondriate lineages are at or near the base of the eukaryotic tree, then the jakobids could be living relatives of an early-diverging mitochondriate lineage. At present, there is virtually no data from nuclear genes available to test this hypothesis. To that end, the work presented in the final chapter of the thesis attempts to examine the relationship

between the jakobids and a variety of mitochondriate and amitochondriate protist lineages using the  $\alpha$  subunit of CCT.

## MATERIALS AND METHODS

### Genomic DNA extraction

Genomic DNA (gDNA) was isolated from *Giardia lamblia* cells (strain WB; ATCC#30957) provided by Dr. D. Edgell and Dr. A. Roger (Dalhousie University, Halifax, Nova Scotia) using the CTAB DNA purification protocol described by Clark (1992). *G. lamblia* cells were resuspended in 250 µl of lysis buffer (0.1 M EDTA, pH 8.0, 0.25% SDS ), proteinase K was added to a concentration of 0.1 mg/ml and the solution was incubated for one hour at 55°C. NaCl (0.7 M) and 1% CTAB were then added and the solution was mixed thoroughly and incubated at 65°C for 15 minutes. The mixture was extracted once with an equal volume of chloroform/isoamyl alcohol and once with phenol chloroform (1:1). After centrifugation (14,000 x g), the aqueous phase was removed and DNA was precipitated with 2.5 volumes of ethanol. This solution was centrifuged again, and the resulting pellet was washed in 70% ethanol, air-dried, and resuspended in sterile deionized water.

### Genomic DNAs as gifts

**Archaea.** Dr. H.-P. Klenk (Epidauros Biotechnologie, Bernried, Germany) provided gDNA from *Desulfurococcus mobilis* and Dr. D. Faguy and M. Schenk (Dalhousie University, Halifax, Nova Scotia) provided *Sulfolobus solfataricus* gDNA. Dr. A. Russell (Dalhousie University) provided gDNA from *Sulfolobus acidocaldarius*.

**Parabasalids.** gDNA from *Trichomonas vaginalis* strain NIH-C1 (ATCC#30001) was a gift from Dr. M. Müller (The Rockefeller University, New York).

*Monocercomonas* sp. gDNA was provided by Dr. A. Roger (Dalhousie University). Dr. J. M. Logsdon, Jr. and Dr. N. Fast (Dalhousie University) generously provided a *Trichomonas vaginalis* (ATCC#30001) gDNA library.

**Heterolobosea.** Dr. A. Roger (Dalhousie University) provided *Acrasis rosea* gDNA (strain T-235) and *Naegleria gruberi* gDNA (ATCC#30224) was a gift from Dr. R. J. Redfield (University of British Columbia, Canada).

**Jakobids.** gDNAs from the jakobids examined in this thesis (*Reclinomonas americana* (ATCC#50394 and 50283) and *Malawimonas jakobiformis* (ATCC#50310)) were kindly provided by Dr. B. F. Lang (Université de Montréal, Canada).

**Trypanosoma brucei.** Dr. S. L. Hajduk and M. Oli (University of Alabama, USA) provided gDNA from *Trypanosoma brucei* (strain MiTat 1.2 427/221).

### PCR primer design

A battery of degenerate PCR primers was designed on the basis of chaperonin protein sequence alignments that contained maximal taxonomic diversity. Regions of high conservation were identified as potential primer sites, and primers were ultimately designed to highly conserved stretches of amino acids containing minimal degeneracy in their codons. Primers were between 17 and 21 nucleotides in length and three or four codons at the 3' end were made completely degenerate, except for the last codon of N-terminal (forward) primers, which was truncated to the first two nucleotides to avoid ending the primer in a degenerate position. The 3' end nucleotide was usually C or G. Codons toward the 5' end of the primer were 'fixed', based on codon biases of the target organism(s) (if known) and with the goal of balancing the overall G+C content.

**Universal chaperonin primers.** The ATP-binding and ATP-hydrolysis domains are the most highly conserved regions of chaperonins, and numerous primers were designed to amino acids in these motifs. These primers were used in all possible combinations to amplify chaperonin genes from a variety of genomic DNAs. Two forward primers were targeted against the overlapping motifs I/VGDGTT (primer CCT-1-for) and EI/VGDGT (primer CCT-5-for) and had the following sequences: CCT-1-for: 5'-TACGGTGAYGGNACNAC-3'; CCT-5-for: 5'-GAAATCGGNGAYGGNAC-3'. Reverse primers were designed against the regions P/AGG/AGAP (primer CCT-3-rev) and GG/AGAVE (primer CCT-4-rev). These primers had the following sequences: CCT-3-rev: 5'-TGGAGCTCCNSCNCCNG-3'; CCT-4-rev: 5'-CTCTACAGCNCCNSCNCC-3'. Additional primers were designed to regions conserved among archaeal chaperonins and the different eukaryotic CCT paralogs to varying degrees. These primers were targeted to the following motifs: PVGLDKM (primer CCT-9-for, 5'-CCAGTCGGTCTNGAYAARATG-3'); NDGATIL (primer CCT-2-for, 5'-AACGACGGTGCNACNATHYT-3'); HDA/SL/ICI/VI/V (primer CCT-7-rev, 5'-ACGATGCACATNGHRTCRTG-3'); RIDD/MIK/R (primer CCT-10-rev, 5'-TGATCAGRTCRTCDATNC-3'); ILRIDD/M (primer CCT-11-rev, 5'-AGGTCGTCGATGCKNARDAT-3').

**Archaeal chaperonin primers.** Primers for archaeal chaperonins were designed to amplify specific subunit-encoding genes from a variety of crenarchaeotes. Four different forward primers were targeted against the following amino acid motifs: TTPEGI (primer TF-1-for), EETADG (primer TF-2-for), LLLK/REGT (primer TF-7-for) and K/VT/ILA/LEM (primer TF-8-for). These primers had the following sequences: TF-1-for: 5'-ACAACCTCCNGARGGNAT-3'; TF-2-for: 5'-GAAGAGACNGCNGAYGG-3'; TF-7-for: 5'-CTACTACTTAGRGARGGNAC-3'; TF-8-for: 5'-GTCATACTABYNGARATG-3'. Two reverse primers were directed

against the regions MNAIKAA (TF-5-rev) and IDDL/MIAA (primer TF-9-rev) and had the following sequences: TF-5-rev: 5'-TGCAGCCTTAATNGCRTTCAT-3'; TF-9-rev: 5'-GCAGCTATCARRTCRTC DAT-3'.

**Miscellaneous primers.** A variety of exact-match primers were designed and used to amplify genes from a variety of organisms for which partial fragments had been identified in searches of publicly available incomplete genome sequence data. These primers were sometimes used in combination with degenerate primers. The full-length  $\alpha$  and  $\gamma$  chaperonin subunit genes in *Sulfolobus solfataricus* were amplified with the following primers: Ssol-TF-a-for (5'-TGCATATGGCAGCTCCAGTCTTATTG-3')/ Ssol-TF-a-rev (5'-CTCGAGGTCTCCTAAAGATGGAGTAGATC-3') and Ssol-TF-g-for (5'-CGCATATGGCCTATTTATTAAGAGAAGGAAC-3')/ Ssol-TF-g-rev (5'-CTCGAGACCTAAGTATGGGTTTGGCTGTIG-3'). For *Trypanosoma brucei*, a small fragment of the CCTalpha gene was found by searching genomic data from unfinished microbial genomes. Preliminary sequence data was obtained from The Institute for Genomic Research website at <http://www.tigr.org>. Based on this sequence an exact-match reverse primer (Tbru.CCTa.R-1 [5'-GTTAAGGCGAACACAGTCAT-3']) was designed and used in combination with degenerate forward primers (CCT-2-for, CCT-9-for; above) to amplify most of the CCTalpha coding sequence from *T. brucei* genomic DNA. For *Giardia lamblia*, exact-match primers were designed based on preliminary genome sequence data and, in combination with degenerate primers (above), were used to amplify multiple CCT genes from *G. lamblia* genomic DNA prepared in this study. The primers and sequences used for each gene were as follows. *Giardia Ccta*: GL.alpha.for.1 (5'-GTAGACATGCTTGTCTGCAG-3')/GL.alpha.rev.1 (5'-GTCGTGTATGCTCTAGTAGC-3'), *Giardia Cctb*: GL.beta.for.1 (5'-CCATAGCTGAGTTATAGATG-3')/ GL.beta.rev.2 (5'-



TAATCTTGTCAGAGTCCATG-3'), CCT-1-for (above)/ GL.beta.rev.3 (5'-AGGTGCACAGCTTATTATGC-3'), *Giardia Cctg*: CCT-5-for (above)/GL.gamma.rev.1 (5'-TCCGCAGAACCATACGCCAG-3'), *Giardia Ccte*: GL.eps.for.1 (5'-ATGATTAGTATCTCTCAGTG-3')/GL.eps.rev.1 (5'-GCTGAA CGATCGTTGTCATG-3'), *Giardia Cctq*: GL.theta.for.1 (5'-TTCTTCCATGATGA AGGTGC-3')/GL.theta.rev.1 (5'-GACCACGTACTGCTCTAGAC-3'), *Giardia Cctz*: GL.zeta.for.1 (5'-AGAATTTTCATGTCTGCTATC-3')/GL.zeta.rev.1 (5'-TGCTCAGAACGTGGTATCTG-3'). Preliminary sequence data from the *Giardia* Genome Project was obtained from The Josephine Bay Paul Center WEB site at the Marine Biological Laboratory ([www.bpc.mbl.edu](http://www.bpc.mbl.edu)). Sequencing was supported by the National Institute of Allergy and Infectious Diseases using equipment from LI-COR Biotechnology.

### **DNA amplification and cloning**

PCR reactions were carried out under standard conditions in 25 or 50  $\mu$ l volumes. The reactions contained 10-100 ng of template DNA, 1-2 U of *Taq* DNA polymerase (Gibco BRL or SIGMA), 0.1-0.2 U of *Pfu* (Stratagene) or Vent (NEB) DNA polymerase, 1x PCR buffer (Gibco BRL or SIGMA), 1.5-3 mM MgCl<sub>2</sub>, 10 mM dNTP (each; Gibco BRL) and 1  $\mu$ M of each primer. PCR reactions performed on *Reclinomonas americana*, *Jakoba libera* and *Malawimonas jakobiformis* DNAs contained 5% acetamide (final concentration). Reactions were performed using Ericomp and MJ Research Inc. PTC-100 thermal-cyclers. The cycling parameters and number of cycles employed varied considerably depending on the size of the expected fragment, the degeneracy of the primers and whether the template DNA was prokaryotic or eukaryotic. After an initial denaturation of 3 minutes at 92°C, 35-45 cycles of 92°C for 15-30 seconds, 45-54°C for 30 seconds to 1 minute

and 72°C for 30 seconds to 3 minutes were performed, followed by a final extension of 3 minutes at 72°C. PCR products were visualized by agarose gel electrophoresis (below).

PCR products were cloned by two methods. Initially, bands of the expected size were excised from standard agarose gels and purified using the Prep-a-gene kit from BIO-RAD. These products were cloned into the T-tailed vector pCR2.1 (Invitrogen) with T4 DNA ligase, and the plasmids were transformed into chemically competent *E. coli* cells (INV $\alpha$ F' from Invitrogen) by heat shock as per the manufacturer's instructions. Later, PCR products were cloned directly from low-melt agarose using the TOPO TA Cloning kit from Invitrogen. Briefly, this method utilizes the activity of Topoisomerase I to ligate PCR-generated DNA fragments into a T-tailed vector (pCRII-TOPO) in a single step. Plasmids were transformed into chemically competent TOP10 *E. coli* cells (Invitrogen). For both methods, transformation reactions were plated on LB solid medium containing ampicillin (50  $\mu$ g/ml) and spread with 75  $\mu$ l of 2% X-Gal in dimethylformamide to allow for the blue-white screening of transformants.

## **Electrophoresis**

Standard agarose gels of between 0.8% and 1.5% agarose in 1x TAE buffer were run in submerged electrophoresis chambers containing 1x TAE. 6% polyacrylamide DNA sequencing gels were made and run in 60-cm vertical Sequi-Gen Sequencing Cells from BIO-RAD. The gels contained 1x TBE, 8% acrylamide/bis-acrylamide (38:2) and 50% urea and were run in 1x TBE buffer. Sequencing gels were dried on 3 MM Whatman chromatography paper using a BIO-RAD slabdrier.

## Screening transformants

To screen *E. coli* transformants for the presence of plasmids containing correctly sized inserts, between 7 and 10 white colonies were selected with a sterile toothpick or pipet tip, touched to a 'master' plate, and used to inoculate overnight cultures (LB medium supplemented with ampicillin to 50 µg/ml). Plasmid DNAs were then isolated using several different commercially available alkaline-lysis kits (Eclipse Mini Plasmid Prep Kit, Eclipse Molecular Biologicals; UltraClean Mini Plasmid Prep Kit, MO BIO Laboratories, Inc.; QIAprep Miniprep Kit, QIAGEN). 5 µl of purified plasmid was restricted with *EcoRI* and digests were electrophoresed to identify clones containing inserts.

A faster, PCR-based screening method was also used. This involved choosing 10 white *E. coli* colonies, touching them to a master plate and placing them in a 10-15 µl PCR reaction containing 0.2 U Taq polymerase, 1.5 mM MgCl<sub>2</sub>, 1x PCR buffer, 5 mM dNTPs and 0.5 µM M13 -20 (forward) and reverse primers. The PCR reaction was performed with 30-35 cycles of 94°C for 1 minute, 57°C for 1 minute and 72°C for 1 minute, followed by a final extension of 72°C for 5 minutes. PCR products were visualized by gel electrophoresis and clones that appeared to contain proper inserts were picked from the master plate and used to inoculate cultures to grow overnight for plasmid preparation (as above) and sequencing.

## Southern transfers and hybridizations

To confirm the origin and copy number of PCR-amplified genes, Southern hybridizations were performed. Typically, 0.5-3 µg of archaeal or eukaryotic genomic DNA was digested with selected restriction enzymes and

electrophoresed on 1% TAE gels. Before transfer, gels were treated with 0.25 M HCl to nick and depurinate the DNA and then with 0.4 M NaOH to denature the DNA. DNAs were transferred to a charged nylon membrane (GeneScreenPlus Hybridization Transfer Membrane from DuPont) using a dry blotting procedure. The gel was laid on a piece of plastic wrap and the nylon membrane, which had been soaked in 2x SSC, was placed on top, followed by several layers of Whatman 3 MM filter paper. Multiple layers of paper towel and Kim-Tuffs were then added, and the apparatus was weighted. Transfers were usually performed overnight, and the paper towels and Kim-Tuffs were replaced with fresh ones for an additional hour the following morning. DNA was then fixed to the membrane by UV crosslinking with a Stratalinker (Stratagene). DNA probes were prepared by digesting cloned PCR products with restriction enzymes. Fragments (typically between 400 and 1500 nucleotides in length) were resolved on agarose gels and purified using the Prep-a-gene band isolation kit (BIO-RAD).

For hybridizations, membranes were soaked in 5x SSC for 10-15 minutes to allow hydration, and placed in glass tubes containing hybridization solution (0.5% nonfat dry milk powder, 1% SDS and 4x SSC). Pre-hybridization was then carried out at 65°C for 1-3 hours in a rotary hybridization oven (Hybaid). During this period, 25 ng of purified PCR product (above) was  $\alpha^{32}\text{P}$ -labeled with the Prime-It II random primer labeling kit from Stratagene. Labeled probe solutions were placed in a boiling water bath for 10 minutes, and fresh hybridization solution was added to the tubes, followed by the denatured probes. Hybridization was performed overnight. After hybridization, membranes were treated with wash buffer (0.5% SDS and 2x SSC). A brief room-temperature wash was performed, followed by two or three 15 minute washes at 65°C. After washing, membranes were placed in cassettes and exposed to X-ray film (Kodak)

at  $-70^{\circ}\text{C}$  for varying lengths of time, depending on the intensity of the hybridization signal.

### **Genomic library screening**

PCR products of *Trichomonas vaginalis* CCT genes were used as probes to isolate full-length genes from a *T. vaginalis* genomic library. Library screening was performed essentially as per the manufacturer's instructions. Briefly, XL1-Blue MRF' *E. coli* cells were grown to mid-log phase in LB medium supplemented with 0.2% maltose and 10 mM  $\text{MgSO}_4$ . Cells were pelleted and resuspended in 10 mM  $\text{MgSO}_4$  (to an  $\text{OD}_{600}=0.5$ ) and 600  $\mu\text{l}$  of cells were infected with an appropriate amount of phage stock (determined by titration to yield ~ 15,000 pfu/plate). After a 15-minute incubation at  $37^{\circ}\text{C}$ , cells were mixed with 8 ml of NZY top agar, plated on 150-mm diameter pre-warmed NZY solid medium and incubated overnight at  $37^{\circ}\text{C}$ . 6-8 plates were used for 'primary' library screenings.

Plaque lifts were performed as follows. Hybond N+ nylon membranes (Amersham) were placed on chilled plates for 2 minutes and the orientation of the membrane with respect to the plate was marked by india ink with a syringe tip. Membranes were then placed plaque-side up in a small pool of 0.5 N NaOH for 2 minutes to denature the DNA. Denaturation was repeated in a fresh pool of NaOH, and neutralization was achieved by placing the membrane in a pool of 1 M Tris-HCl for 30 seconds. Membranes were placed on Whatman 3 MM filter paper to dry, and DNA was fixed to the membrane by UV crosslinking. Plaque lifts were usually performed in duplicate.

Membranes were hybridized against the appropriate probe as described above, with the exception that hybridizations were performed in a single

Tupperware container placed in a 65°C water bath/shaker. After washing, membranes were exposed to X-Ray film at -70°C in the presence of intensifier screens for 1-2 days.

To identify putative hybridizing plaques, positive signals on developed autoradiographs were matched to the lifted plates. Appropriate regions were 'cored' with the broad end of a Pasteur pipet and placed in 1 ml of SM buffer and 20 µl of chloroform. Cores were vortexed and left at room temperature for ~5 hours or overnight at 4°C to allow the phage to elute from the core into the buffer. Serial dilutions were made and used to perform 'secondary' screenings by infecting XL1-Blue MRF' cells as described above. Smaller 100-mm NZY plates were used for secondary screenings, and the volume of plating cells and top agar were adjusted accordingly. Single plaques corresponding to hybridization signals were cored (this time with the narrow end of a Pasteur pipet) and placed in 0.5 ml of SM buffer and 20 µl of chloroform to release the phage.

The Lambda Zap Express 'single-clone excision protocol' (Statagene) was used to perform *in vivo* excisions, allowing the isolation of plasmid DNAs containing genomic fragments of interest. Separate overnight cultures of XL1-Blue MRF' cells (supplemented with 0.2% maltose and 10 mM MgSO<sub>4</sub>) and XLOLR cells were grown in NZY broth and resuspended in 10 mM MgSO<sub>4</sub> to an OD<sub>600</sub> of 1.0. Next, 200 µl of XL1-Blue MRF' cells were mixed with 250 µl of phage stock and 1 µl of ExAssist helper phage. This mixture was incubated at 37°C for 15 minutes, added to 3 ml of NZY broth and incubated at 37°C for 2-3 hours to overnight with shaking. Cultures were incubated at 65°C for 20 minutes and centrifuged for 15 minutes (1000 x g). The resulting supernatant contained excised pBK-CMV phagemid vector packaged as filamentous phage particles. 100 µl of supernatant was used to infect 200 µl of XLOLR cells, and these cells were incubated at 37°C for 15 minutes. 300 µl of NZY medium was then added,

cultures were incubated at 37°C for an additional 45 minutes, and 200 µl of the cell mixture was incubated overnight (37°C) on LB solid medium containing kanamycin (50 µg/ml). Individual colonies were grown and PBK-CMV plasmid was isolated (as above) for analysis.

To identify positive genomic library clones, single and double digests were performed with *KpnI* and *SacI* on 10 µl of purified plasmid. Digests were electrophoresed, transferred to nylon membranes, and hybridized with the appropriate probe as described previously. Clones exhibiting positive hybridization signals were subjected to DNA sequencing.

### **DNA sequencing**

Manual sequencing was used to screen cloned PCR products. Standard dideoxy sequencing reactions were performed with  $\alpha$ -<sup>35</sup>S-dATP on plasmid DNAs using T7 DNA polymerase from the Pharmacia <sup>17</sup>Sequencing Kit. Reactions were performed in 96-well microtiter plates on top of a heating block. A combination of ABI and LiCor automated sequencing was used to 'polish' PCR-generated and genomic library-derived clones. Automated sequencing was carried out by the Joint Lab Sequencing Service (NRC, Halifax, Nova Scotia) and the Roger/Doolittle/Gray Automated Sequencing Facility (Dalhousie University). Most clones were completely sequenced on both strands, and multiple independent clones of PCR products were sequenced to rule out PCR-generated errors.

## DNA sequence analysis and assembly

Nucleotide sequences obtained from sequencing gels were entered manually into the sequence analysis program EditSeq (LaserGene). These were compared to known gene sequences in the public databases using the BLAST suite of programs (Altschul *et al.* 1990; Altschul and Koonin 1998; Altschul *et al.* 1997). DNA sequences were assembled into contigs using Sequencher 3.1 (Gene Codes Corp., Ann Arbor, MI). DNASTrider (version 3.1) was used to analyze the open reading frames and restriction patterns of newly sequenced genes and, if introns were present, to examine exon and intron boundaries. EditSeq and DNASTrider were also used to obtain the molecular weights and isoelectric points of inferred protein sequences.

## Phylogenetic analysis

ClustalW (Thompson, Higgins and Gibson 1994) was used to align protein sequences inferred from newly sequenced genes with homologs obtained from public databases. Alignments were adjusted manually in the text editor of PAUP\* 4.0 (Swofford 1998), taking into account published alignments (if available) and, in some cases, protein crystal structures. For the chaperonins, two different 'master' alignments were used, one containing representative archaeal and eukaryotic (i.e., group II) chaperonin sequences and diverse bacterial/organellar homologs (group I chaperonins) and another containing all known archaeal and eukaryotic cytosolic chaperonin sequences. An alignment of animal, fungal and plant heat shock transcription factors (HSFs) was constructed from sequences collected from the databases by using the *Guillardia theta* nucleomorph HSF sequence as a probe. For Hsp70 and Hsp90, 'starter' alignments were kindly



provided by A. Roger (Dalhousie University) and additional sequences obtained from the databases were added to these alignments manually based on globally conserved regions. Smaller alignments containing various subsets of sequences were created from these datasets. Ambiguous regions of alignments were removed prior to phylogenetic analysis.

Phylogenetic trees were inferred using maximum parsimony (MP), distance-based, and maximum-likelihood (ML) methods of tree reconstruction. PAUP\* 4.0 (Swofford 1998) was used for all parsimony analyses. The heuristic search option was used to find the maximum parsimony tree and 10-100 random sequence addition replicates were performed with tree-bisection-reconnection (TBR) branch swapping. Bootstrap analyses were performed on 100 or 1000 resampled datasets, depending on the number of taxa in the alignment. A 'simple sequence addition' heuristic search was used to identify the maximum parsimony tree for each bootstrap replicate.

Distance analyses were performed using programs in Felsenstein's PHYLIP 3.57 package (Felsenstein 1995). PROTDIST was used to calculate amino acid distance matrices (using the PAM001 matrix as the model of amino acid substitution), and trees were inferred from these distance matrices using the neighbor-joining and Fitch-Margoliash methods (NEIGHBOR and FITCH programs in PHYLIP 3.57, respectively). Support for distance trees was obtained by bootstrapping with 100 or 1000 resampling replicates, generated with the SEQBOOT program. The majority-rule consensus tree was constructed using CONSENSE.

ML analyses were performed with protML in the MOLPHY package (Adachi and Hasegawa 1996). The Jones, Taylor and Thornton amino acid substitution matrix was used and adjusted to account for the amino acid frequencies observed in the dataset (option JTT-F). Due to the computationally

intensive nature of these analyses, ML trees were most often inferred using the heuristic (quick-add OTU) search option in protML (option -q). Occasionally, exhaustive ML searches (option -e) were performed using partially constrained input trees determined by consideration of other methods and analyses (in general, nodes with bootstrap support greater than 90% in parsimony and distance analyses were constrained). REL values (re-sampling estimated log likelihoods) were used as measures of support for ML trees and were obtained by quick-add searches of 100, 1000 or 2000 trees in protML. REL values were calculated with the programs mol2con.pl (Dr. A. Stoltzfus, unpublished) and CONSENSE (Felsenstein 1995).

ML trees were also inferred using the quartet puzzling method of Strimmer and von Haeseler (1996) in PUZZLE 4.0 and 4.02 (Strimmer and von Haeseler 1997). The JTT-F model of amino acid substitution was used and among-site rate variation (ASRV) was taken into account by using an eight rate category discrete approximation to the  $\Gamma$  distribution, plus an additional 'invariable site' category. Quartet puzzling values were used to quantify branch support for quartet puzzling trees. For ML-distance trees, PUZZLE 4.02 was used to calculate ML distance matrices (using an ASRV model, as above). Trees were constructed using FITCH and NEIGHBOR in PHYLIP 3.57 (Felsenstein 1995). Support values for these trees were obtained by bootstrapping (500 replicates) with PUZZLEBOOT 1.02 (A. Roger and M. Holder; <http://members.tripod.de/korbi/puzzle/>). PUZZLE was also used to determine the proportion of constant amino acid positions in alignments, to perform Chi-square tests for the detection of amino acid composition biases, and to statistically assess the significance of different tree topologies using the Kishino-Hasagawa test (Kishino and Hasegawa 1989). Alternative input topologies were constructed using the program TreeViewPPC 1.5.3 (Roderic D.M. Page, Glasgow, Scotland),

and differences in log likelihood between topologies of greater than 1.96 standard errors were considered significantly worse at the 5% level. To estimate site-by-site evolutionary rates across protein sequence alignments, discrete  $\Gamma$  distributions with eight variable site rate categories, or one invariable plus eight variable site rates categories, were calculated over neighbor-joining or Fitch-Margoliash trees using the JTT-F model of amino acid substitution in PUZZLE. Molecular-clock likelihood ratio tests were also performed to statistically assess differences in substitution rates among the different eukaryotic chaperonin subunits.

Phylogenetic analyses were also performed with nucleotide sequences. A DNA sequence alignment of archaeal chaperonin genes was constructed using `aa-dna-align.pl` (unpublished, O. Feeley, Dalhousie University). Given an alignment of protein sequences and their respective coding sequences, this program creates a nucleotide alignment that is gapped with respect to the amino acid sequence alignment. Additional DNA sequences were subsequently added manually based on highly conserved regions, ensuring that the proper reading frame was maintained.

Using an alignment of 39 archaeal chaperonin DNA sequences and 1335 unambiguously aligned sites, the data were initially tested for the optimal fit of various models of nucleotide sequence evolution using MODELTEST (version 3.0b3; Posada and Crandall 1998). The base frequencies and the proportion of invariable sites ( $P_{INV}$ ) were also estimated from the data, and a discrete  $\Gamma$  distribution was approximated by four rate categories. By comparing the log-likelihood ( $-\ln L$ ) scores obtained with the various models, a general time-reversible (GTR) model (Rodriguez *et al.* 1990) that incorporated a  $\Gamma$  distribution (with four rate categories) and the proportion of invariable sites (GTR+ $\Gamma$ + $P_{INV}$ ) was found to be significantly better than any of the other models tested.

PAUP\* version 4.0b5 (Swofford 1998) was used for all DNA-level phylogenetic analyses. Maximum likelihood (ML) and ML-distance trees were inferred with the GTR+ $\Gamma$ + $P_{\text{INV}}$  model using the heuristic search option. Starting trees for TBR branch swapping were obtained by the neighbor-joining method, or by stepwise addition with the starting sequence chosen by 10 random additions. Branch support for ML and ML-distance trees was obtained by bootstrapping with 100 or 1000 resampling replicates, often with 10 random additions performed per bootstrap replicate. To control for the effects of saturation at highly degenerate codon positions, phylogenetic trees constructed from alignments containing first and second positions only were compared to those inferred from all three codon positions.

### **Tests for gene conversion**

GENECONV (Sawyer 1989; <http://lado.wustl.edu/~sawyer/geneconv/index.html>) was used to test for instances of gene conversion in the archaeal chaperonin dataset. Briefly, GENECONV detects region(s) between pairs of sequences in an alignment that share more consecutive identical silent polymorphisms than would otherwise be expected by chance. For protein coding genes, it is possible that a given pair of sequences in an alignment exhibit significant similarity due to functional (selective) constraints on the protein sequence, rather than because of gene conversion. To control for this, GENECONV provides the option of focussing on 'silent polymorphic sites'—degenerate sites in an alignment whose codons all specify the same amino acid (Drouin *et al.* 1999). For the archaeal chaperonin dataset, DNA sequence alignments (above) containing varying numbers of taxa were used as input to GENECONV. All polymorphic sites (default), as well as silent-site-only

(synonymous) polymorphisms (*-seqtype=silent*), were tested for evidence of gene conversion using mismatch penalties of 0 (default), 1, 2, and 3 (*gscale=0-3*). N=10,000 (default) random permutations of the polymorphic sites were performed in each analysis to assess the statistical significance of putative gene conversion tracts.

Two additional approaches were used to detect anomalously evolving regions within archaeal chaperonin genes, both utilizing a 'sliding window' analysis of the data. The first was to assess the % nucleotide sequence identity shared between duplicate genes within a given genome in 50- or 100-nucleotide windows across their entire sequence, advancing the window in one-nucleotide increments. Discrete regions of particularly high nucleotide sequence identity between two otherwise divergent duplicates would suggest areas of recent gene conversion between the two genes. Protein sequences were also examined in this fashion using 20- or 50-amino acid sliding windows and one-amino acid increments.

The second approach involved 'scanning' the alignment for regions of contradictory phylogenetic signal using a maximum likelihood approach. This was done by employing a novel method in which the topology of the maximum likelihood tree obtained from the complete alignment (i.e., the whole molecule) was compared to those inferred from subsets of the data. This was done using all positions in the alignment as well as for first and second codon positions only. Using 'character set definitions' in PAUP\* (Swofford 1998), a 100-nucleotide window was systematically advanced across the alignment in 10-nucleotide increments. For each window, the log likelihood of the best tree from a heuristic maximum likelihood search was obtained, as was the likelihood of the data present in the window given the topology obtained from the analysis of the complete alignment. The difference between these two likelihoods reflects the

degree of congruence between the phylogenetic signal present in a given 100-nucleotide subset of the molecule and that in the molecule as a whole. These differences can be plotted to visualize trends. If a significantly worse likelihood is obtained when the topology from the complete alignment is forced upon the data in a 100-nucleotide window, an anomalous phylogenetic history likely exists for this region of the gene. In these analyses, only full-length sequences were considered, and crenarchaeotes and euryarchaeotes were analyzed separately (12 sequences for crenarchaeotes, 21 for euryarchaeotes). The ML sliding window profiles for smaller datasets containing subsets of these sequences were also examined. For both types of sliding window analyses, the data (% sequence identity or differences in log likelihoods ( $\Delta \ln Ls$ )) were analyzed and plotted using Microsoft Excel.

### **Chaperonin crystal structure visualization**

The crystal structure of the *Thermoplasma acidophilum* chaperonin determined by Ditzel *et al.* (Ditzel *et al.* 1998) (1A6D) was visualized using Insight II (MSI, San Diego, CA) and Cn3D version 3.0 (<http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml>). Individual residues were highlighted on the structure with the use of a script written by C. Blouin (Dalhousie University).

## CHAPTER I

### Archaeal Chaperonin Evolution: Recurrent Paralogy

This chapter includes work published in Archibald, J. M., J. M. Logsdon, Jr., and Doolittle, W. F. 1999. Recurrent paralogy in the evolution of archaeal chaperonins. *Curr. Biol.* **9**: 1053-1056, and Archibald, J. M., C. Blouin, and W. F. Doolittle. 2001. Gene duplication and the evolution of group II chaperonins: implications for structure and function. *J. Struct. Biol.* (in press).

New sequences have been deposited in Genbank under accession numbers AF149920-AF149925 and AF181261.

## INTRODUCTION

The chaperonins are a ubiquitous class of molecular chaperone involved in the folding of non-native proteins (reviewed in Bukau and Horwich 1998; Ranson, White and Saibil 1998; Sigler *et al.* 1998). Individual chaperonin monomers assemble to form multi-subunit complexes that harbor substrates within a central cavity and facilitate protein folding through the hydrolysis of ATP. On the basis of protein sequence similarity, two distantly related types of chaperonins are apparent, and these have distinctive phylogenetic distributions. Group I chaperonins are found in Bacteria and their eukaryotic organellar derivatives, while group II chaperonins are present in Archaea (Archaeobacteria) and the eukaryotic cytosol (Willison and Horwich 1996; Willison and Kubota 1994). Comparison of the atomic structures of group I and group II chaperonins reveals significant evolutionary conservation (Braig *et al.* 1994; Ditzel *et al.* 1998). Despite sharing only 20-25% amino acid sequence identity, their basic

architecture is very similar, and both assemble to form multisubunit oligomers with double-ring quaternary structures. Both chaperonin types possess equatorial and apical domains linked by a flexible intermediate domain. The equatorial domain is involved in the binding and hydrolysis of ATP and provides most of the intra- and inter-ring contacts, while the apical domain is involved primarily in the binding of substrate (Sigler *et al.* 1998).

Group I and group II chaperonins also differ from one another in several important respects. In bacteria, chaperonin monomers assemble to form seven-membered rings (Braig *et al.* 1994), while archaeal and eukaryotic cytosolic chaperonin complexes are composed of eight- or nine-membered rings (reviewed in Gutsche, Essen and Baumeister 1999; Klumpp and Baumeister 1998; Willison and Horwich 1996). As well, no direct counterpart to the bacterial co-chaperonin GroES/cpn10 has been identified in the group II chaperonin system. Instead, group II chaperonins possess a built-in 'lid'—an extension of the apical domain called the helical protrusion—that is absent in group I chaperonins. The helical protrusions are thought to seal off the central cavity of the chaperonin oligomer in a fashion analogous to GroES/cpn10, and have also been suggested to play a role in substrate recognition (Horwich and Saibil 1998; Klumpp and Baumeister 1998; Klumpp, Baumeister and Essen 1997). Furthermore, group II chaperonins are known to interact with novel co-factors that have no apparent homologs in the bacterial system (Siegert *et al.* 2000; Gebauer, Melke and Gehring 1998; Geissler, Siegers and Schiebel 1998; Siegers *et al.* 1999; Vainberg *et al.* 1998). The precise roles of these co-factors in the protein folding process, and how they interact with the group II chaperonins, are as yet unclear.

The most striking feature of the group II chaperonins is their hetero-oligomeric composition. While bacterial chaperonins (e.g., GroEL in *E. coli*) are generally homo-oligomeric (constructed from seven identical subunits), the CCT



complex in eukaryotes is completely hetero-oligomeric, possessing eight distinct (but homologous) subunits (CCT $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$ , and  $\theta$ ) whose genes evolved from a common ancestor by gene duplication (Kubota *et al.* 1994; Willison and Kubota 1994). Each subunit species is thought to occupy a unique position in each of the eight-membered CCT rings (Liou and Willison 1997). Recent biochemical and electron microscopic studies have revealed that the binding of actin and tubulin (two well-defined CCT substrates) within the central cavity of the chaperonin particle is both subunit-specific and geometry-dependent (i.e., a unique arrangement of CCT subunits is required; Llorca *et al.* 2000; Llorca *et al.* 1999a). The transition from homo-oligomeric chaperonin rings, like those in *E. coli*, to completely hetero-oligomeric ones like CCT was therefore an important step in chaperonin evolution.

In archaea, chaperonin complexes exhibit an intermediate degree of complexity. Archaeal chaperonins were first described, as heat shock proteins, in two hyperthermophilic organisms, *Sulfolobus shibatae* and *Pyrodictium occultum* (Phipps *et al.* 1991; Phipps *et al.* 1993; Trent *et al.* 1991). The *Sulfolobus* protein showed remarkable amino acid sequence similarity with *t*-complex polypeptide-1 (TCP-1; Willison, Dudley and Potter 1986), now known to be the  $\alpha$  subunit of CCT (Frydman *et al.* 1992; Lewis *et al.* 1992). The chaperonin complex in the euryarchaeote *Thermoplasma acidophilum* (called the thermosome) has been crystallized and is composed of two homologous subunits,  $\alpha$  and  $\beta$ , that alternate in each of its eight-membered rings (Ditzel *et al.* 1998). While chaperonin complexes in the crenarchaeote *Pyrodictium* also appear to have this arrangement (Phipps *et al.* 1991; Phipps *et al.* 1993), the organization of subunits in the nine-membered chaperonin rings observed in *Sulfolobus* species remains unclear. The *Sulfolobus* chaperonins were first described as homo-oligomeric (Trent *et al.* 1991)

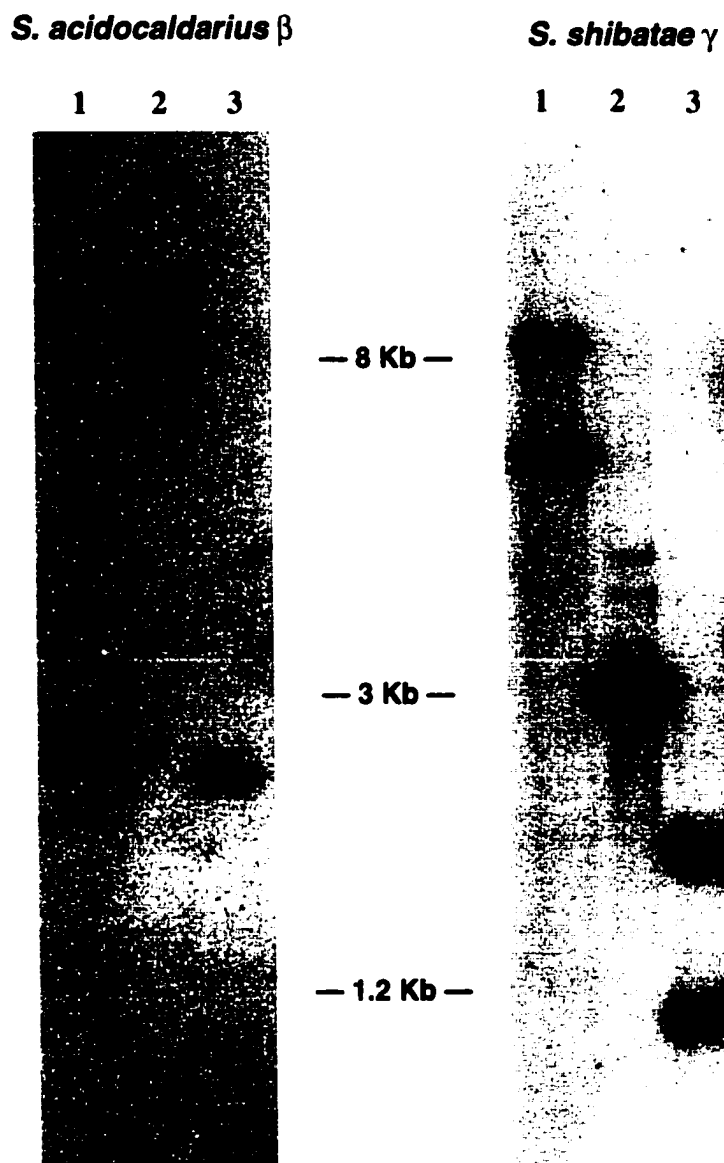
but were later found to possess two different subunit species (Kagawa *et al.* 1995; Knapp *et al.* 1994; Nakamura *et al.* 1997).

In an attempt to elucidate the evolutionary history of archaeal chaperonins, I isolated and sequenced chaperonin genes from a variety of crenarchaeotes, and performed phylogenetic analyses on all available archaeal chaperonin protein sequences. In this chapter I demonstrate that archaea exhibit a remarkably complex pattern of chaperonin subunit evolution involving gene duplication, gene conversion and gene loss. Hetero-oligomeric chaperonin complexes appear to have arisen—and been lost—multiple times independently in the evolutionary history of this group. I propose a ‘neutral’ evolutionary model that attempts to explain this pattern, in which co-evolution between duplicate subunits can lead to obligatory hetero-oligomerism.

## RESULTS

### Archaeal chaperonin sequences

Degenerate PCR was used to isolate chaperonin subunit-encoding genes from *Sulfolobus* and *Desulfurococcus* genomic DNAs (gDNAs). Near full-length orthologs of the previously described  $\alpha$  and  $\beta$  subunit genes were cloned and sequenced from *S. acidocaldarius* with the primer sets TF-8-for/TF-9-rev ( $\alpha$ ) and TF-1-for/TF-5-rev ( $\beta$ ) while the primer pair TF-2-for/TF-9-rev was used to amplify the  $\beta$  subunit gene from *S. solfataricus*. A  $\beta$ -like gene was also isolated from *Desulfurococcus mobilis* using the primers TF-7-for and TF-9-rev. The sources of these genes were confirmed by hybridizing isolated PCR products to their respective gDNAs. Representative Southern hybridizations are shown in Figure 1.1.



**Figure 1.1** Southern hybridizations of labeled chaperonin gene fragments to restricted *Sulfolobus* genomic DNAs. *S. acidocaldarius*  $\beta$ : lane 1- *Eco*R1; lane 2- *Hind*III; lane 3- *Eco*R1/*Hind*III. *S. shibatae*  $\gamma$ : lane 1- *Hind*III; lane 2- *Xho*I; ; lane 3- *Hind*III/*Xho*I. Both genes appear to be present in single copy (the *S. shibatae*  $\gamma$  gene possesses a *Hind*III restriction site).

The *S. solfataricus*  $\alpha$  subunit gene was obtained by searching incomplete genome sequence data (Sensen *et al.* 1998), as was the complete sequence of the partial  $\beta$  subunit gene amplified by PCR (above). A third, previously undescribed, chaperonin subunit gene was also discovered, and named  $\gamma$ . The presence and sequence of the  $\alpha$  and  $\gamma$  genes in *S. solfataricus* was subsequently confirmed by PCR amplification with the following primer pairs:  $\alpha$ ; Ssol-TF-a-for/Ssol-TF-a-rev,  $\gamma$ ; CCT-5-for/CCT-3-rev and Ssol-TF-g-for/Ssol-TF-g-rev. A  $\gamma$  ortholog was also isolated from *S. shibatae* by PCR (using the primer pair CCT-5-for/CCT-3-rev) and successfully hybridized to *S. shibatae* genomic DNA (Figure 1.1). To confirm that the *S. shibatae* gDNA used to amplify the  $\gamma$  subunit gene was the same as that used by other authors, a fragment of the  $\beta$  gene was amplified using the CCT-5-for/TF-5-rev primer pair. A portion of this fragment was sequenced and found to be identical over 280 nt to the *S. shibatae*  $\beta$  sequence described previously (Kagawa *et al.* 1995; Trent *et al.* 1991). Attempts to amplify a  $\gamma$  subunit gene from *S. acidocaldarius* as well as  $\alpha$  and  $\gamma$  genes from *D. mobilis* using all appropriate pairs of degenerate primers failed.

Using the procedures described in the Materials and Methods, amino acid sequences inferred from the newly sequenced genes were aligned with homologs obtained from public databases. Chaperonin sequences from the crenarchaeote *Pyrobaculum aerophilum* were obtained (with permission from S. Fitz-Gibbon, Los Angeles CA) from the *Pyrobaculum* project homepage (<http://www.doe-mpi.ucla.edu/PA/>). Figure 1.2 shows an alignment of all available archaeal chaperonin protein sequences (as of 03/2001) and illustrates the high degree of amino acid identity shared among chaperonins from diverse archaeal lineages.

**Figure 1.2** Alignment of archaeal chaperonin protein sequences. All available euryarchaeal and crenarchaeal sequences are present (40 in total, 605 aligned amino acid positions). Genes sequenced in this study are indicated by asterisks. Amino acid residues present in at least 60% of the sequences are shaded black and chemically similar amino acids (if present in  $\geq 60\%$  of the sequences) are shaded gray. Taxon abbreviations: *Taci*, *Thermoplasma acidophilum*; *Tvol*, *T. volcanium*; *Mthe*, *Methanobacterium thermoautotrophicum*; *Mkandl*, *Methanopyrus kandleri*; *Mjanna*, *Methanococcus jannaschii*; *Mtherm*, *M. thermolithotrophicus*; *Hvol*, *Haloferax volcanii*; *Fialo*, *Halobacterium* sp. NRC-1; *Aful*, *Archaeoglobus fulgidus*; *ThK1*, *Thermococcus* strain KS-1; *ThK8*, *Thermococcus* strain KS-8; *Phorik*, *Pyrococcus horikoshii*; *Pfurio*, *P. furiosus*; *Pabyss*, *P. abyssi*; *Susp*, *Sulfolobus* sp. S7; *Saci*, *S. acidocaldarius*; *Sshi*, *S. shibatae*; *Ssol*, *S. solfataricus*; *Pocc*, *Pyrodictium occultum*; *Aper*, *Aeropyrum pernix*; *Paer*, *Pyrobaculum aerophilum*; *Dmob*, *Desulfurococcus mobilis*.



Figure 1.2 Alignment of archaeal chaperonin protein sequences

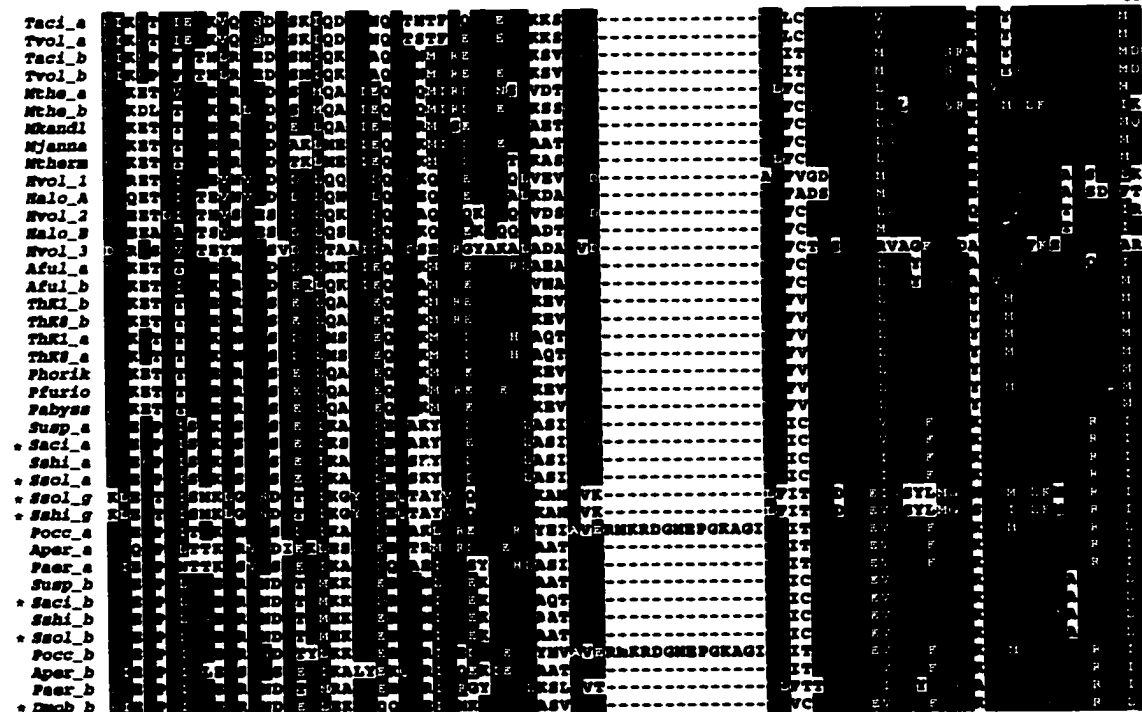


Figure 1.2 Alignment of archaeal chaperonin protein sequences

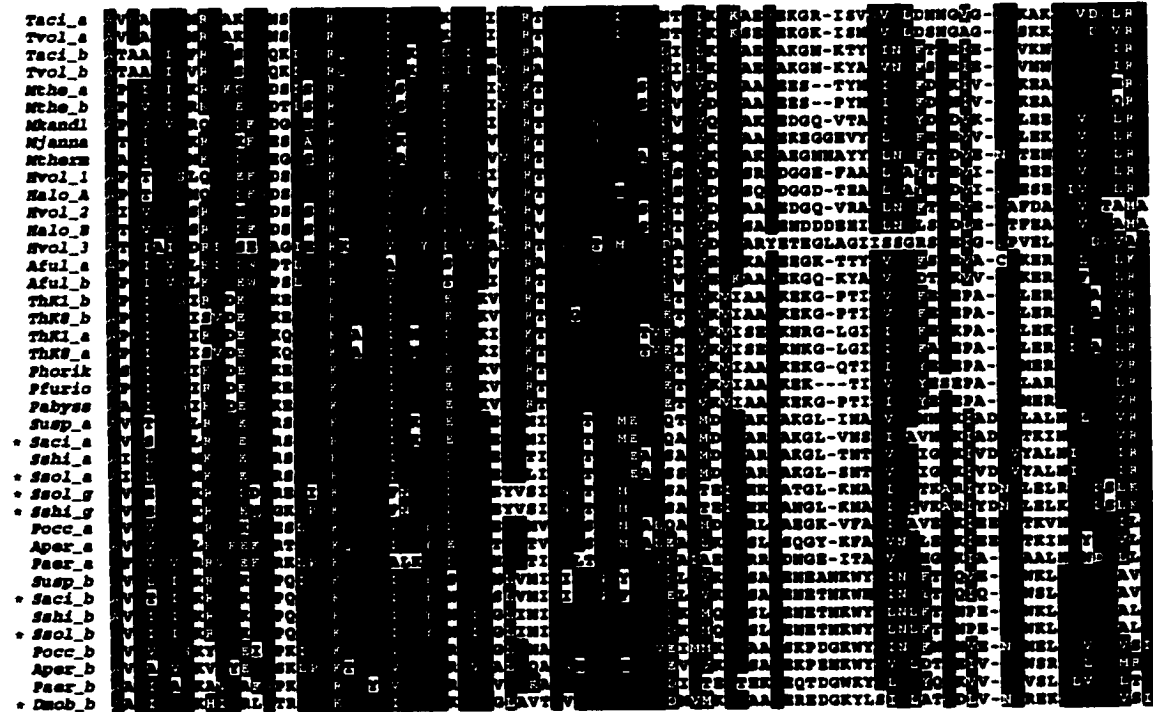


Figure 1.2 Alignment of archaeal chaperonin protein sequences



605

```

Taci_a  MKSTPPSGQQGGQQGMPGGMP EY
Tvol_a  MKSTPPSNQPGQGAGAPGGMP EY
Taci_b  MKSSSSSNPPKSSSSSSSD
Tvol_b  MKSSSSSNPPKSPSSSSSD
Mtha_a  MSQSSSEKHEKGGGKGGGPFK
Mtha_b  RGFVSEKDEEDMEKGGGKGGGPFK
Mkandl  MELKKEEHEEHEGGSSEF
Mjanna  MKVKGDERKGGEGGDMGGDEP
Mtharn  MKLSQGSQGGDMGGGKGGGKGGM
Nvol_1  GDLGQQTGSDDDDDGGAFGGKGGGKGGGKGGGKGGM
Nalo_A  GDLGQQTGSDDDDDGGAFGGKGGGKGGGKGGGKGGM
Nvol_2  GDLSTGGDDDEGGAFGGKGGGKGGGKGGGKGGM
Nalo_B  GDLSTDKDDDDGAGGKGGGKGGGKGGM
Nvol_3  K
Aful_a  KGLEKKEKQPEKESGGEEDSEK
Aful_b  KGLEKKEKGGGEGGMP EHP EY
Thki_b  SKLEKDKGGKGGTDFGSDLD
Thki_a  SKLEKDKGGKGGTDFGSDLD
Thki_s  KATKPEKGGQGGMPGGGKGGDMGM
Phorik  KVSKPEKGGKGGKGGGSEKSGSDLD
Pfurio  K
Pabyss  KGLEKKEKGGKGGGSEKSGSDLD
Susp_a  APLKSGEKKGGKGGSEKSGSDLD
Saci_a  APLKSEKGGGKGGKGGSEKSGGAGTFSLGD
Sshi_a  APLKSEKGGGKGGKGGSEKSGGAGTFSLGD
Ssol_g  AFAKQQFPQPPFFLG
Sshi_g  AFAKQQFPQPPFFLG
Pocc_a  APTKSEKGGKGGKGGSEK
Aper_a  APTKSEKGGKGGKGGSEKGGKGGSEK
Pacr_a  GAPKSEKGGKGGKGGSEKGGKGGSEK
Susp_b  GKSEKGGKGGKGGSEKGGKGGSEK
Saci_b  K
Sshi_b  GKSGSEKGGKGGKGGSEKGGKGGSEK
Ssol_b  ARKSEKGGKGGKGGSEKGGKGGSEK
Aper_b  ARKSEKGGKGGKGGSEKGGKGGSEK
Pacr_b  SKLEKKEKGGKGGKGGSEKGGKGGSEK
Dacb_b  LAK

```

Figure 1.2 Alignment of archaeal chaperonin protein sequences

### Archaeal chaperonin phylogeny: a pattern of recurrent paralogy

The  $\alpha$  and  $\beta$  subunits of *Thermoplasma acidophilum* share ~60% amino acid sequence identity (Waldmann *et al.* 1995a; Waldmann *et al.* 1995b), and evolved from a common ancestral chaperonin by gene duplication. When compared to chaperonin protein sequences from other archaea, the two subunits are more similar to each other than to other archaeal chaperonin sequences, suggesting that the gene duplication that produced them occurred after this lineage diverged from other archaea. In fact, 'lineage-specific' gene duplication (paralogy) is a recurring theme in the evolution of archaeal chaperonins.

The Archaea are generally considered to be a monophyletic group containing two evolutionarily distinct lineages, euryarchaeotes (e.g., *Thermoplasma*) and crenarchaeotes (e.g., *Sulfolobus*; Woese, Kandler and Wheelis 1990, but see Brown *et al.* 1994). Figure 1.3 shows the results of a phylogenetic analysis performed with an alignment containing all known archaeal chaperonin protein sequences and 452 unambiguously aligned amino acid positions. The deepest branching in the phylogeny of archaeal chaperonins separates euryarchaeotes and crenarchaeotes, consistent with the notion that a single chaperonin subunit gene in their common ancestor gave rise to all modern archaeal genes. Nevertheless, within both euryarchaeotes and crenarchaeotes, paralogy is rampant: a minimum of eight events of chaperonin gene duplication can be inferred.

Within euryarchaeotes, lineage-specific duplications of chaperonin genes have occurred in *Methanobacterium thermoautotrophicum*, *Archaeoglobus fulgidus*, and the *Thermococcus/Pyrococcus* clade, in addition to the duplication in the *Thermoplasma* lineage. Recently, the genome of *Thermoplasma volcanium*, an organism closely related to *T. acidophilum*, was completely sequenced (Kawashima *et al.* 2000). *T. volcanium* also possesses genes encoding two subunits

**Figure 1.3** Phylogenetic analysis of archaeal chaperonins. The tree shown is a maximum likelihood (ML) tree ( $\ln L = -15,614.72$ ) inferred from a heuristic search of 1000 trees in protML (Adachi and Hasegawa, 1996). A protein sequence alignment containing 40 sequences and 452 unambiguously aligned amino acid positions was used. Sequences determined in this study are in bold, and asterisks appear next to sequences from organisms whose genomes have been completely sequenced. The two recognized kingdoms within Archaea (euryarchaeotes and crenarchaeotes) are labeled, and inferred gene duplications and gene losses are indicated (see text). Within euryarchaeotes, regions of the tree in which lineage-specific gene duplications have occurred are shaded. For crenarchaeotes, the three different gene/subunit families ( $\alpha$ ,  $\beta$  and  $\gamma$ ) are indicated. Statistical support for significant nodes appear above and below the branches (above, ML REL bootstrap values inferred from a heuristic search of 1000 trees in protML (Adachi and Hasegawa, 1996); below, quartet puzzling support values). The position of the root of the archaeal chaperonin tree (labeled *ROOT*) is highlighted with an arrow and was determined by the analysis of a larger dataset that included 16 representative eukaryotic CCT sequences (see Figure A.1, Appendix A; 56 taxa in total, 348 unambiguously aligned sites). Support values for this node are provided as above. The scale bar represents the estimated number of amino acid substitutions per site.

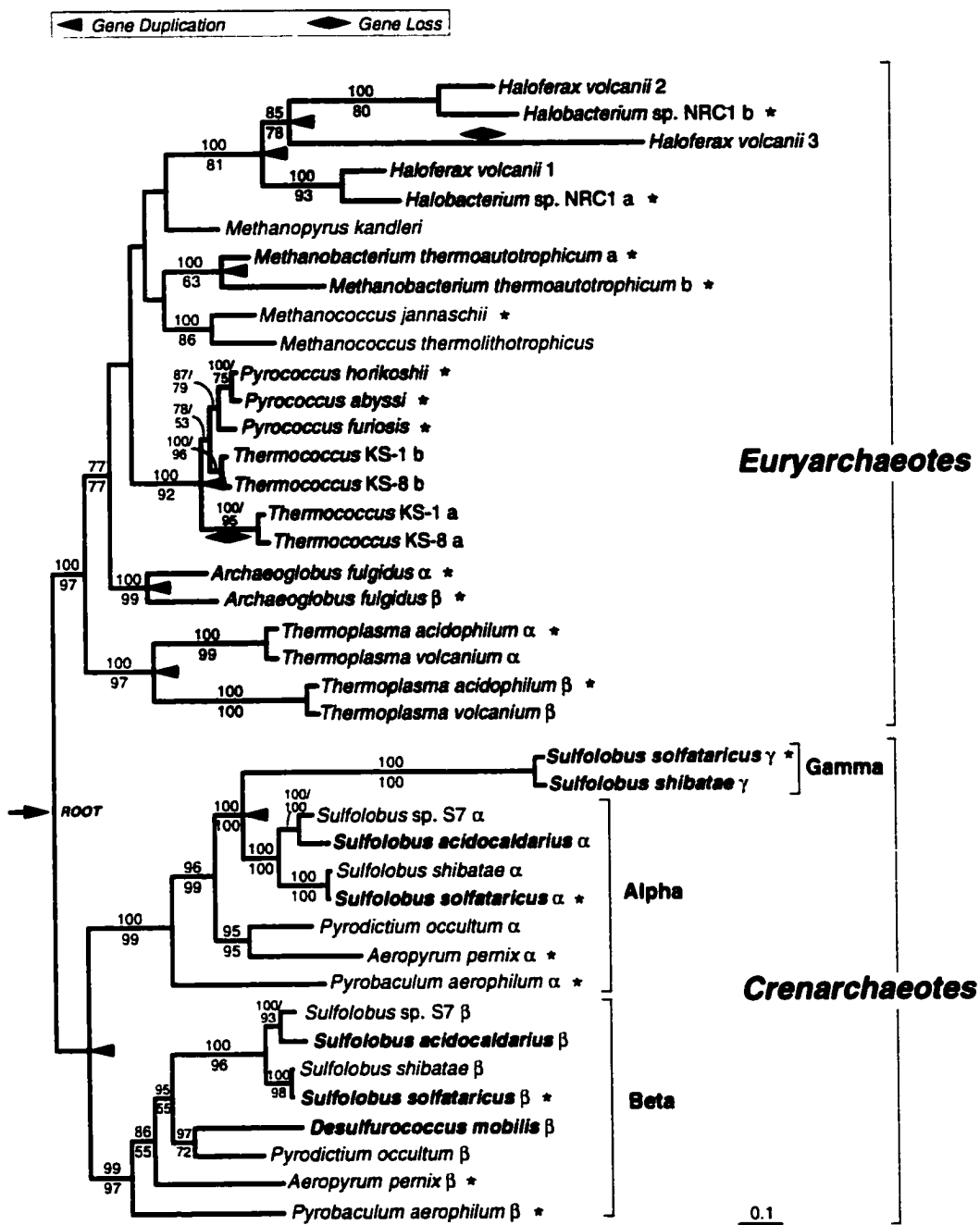


Figure 1.3 Phylogenetic analysis of archaeal chaperonins

very similar to those in *T. acidophilum*; this indicates that the duplications producing the subunits occurred in their common ancestor, but independent of duplications in other archaea (Figure 1.3). Two separate gene duplications have also occurred within the halophilic archaea (e.g., *Haloferax*). Amino acid identities between euryarchaeal paralogs range from 46.1% (*H. volcanii* 2 and 3) to 80.6% (*Thermococcus* KS-1  $\alpha$  and KS-1  $\beta$ ), suggesting that some duplications occurred more recently than others. It should be noted that, for the most part, the designation of duplicate subunits as  $\alpha$  and  $\beta$  (or a and b) in euryarchaeotes is arbitrary. For example, there is no sense in which the  $\alpha$  subunit in *T. acidophilum* is more similar to the  $\alpha$  subunit of *A. fulgidus* than it is to the *T. acidophilum*  $\beta$  subunit.

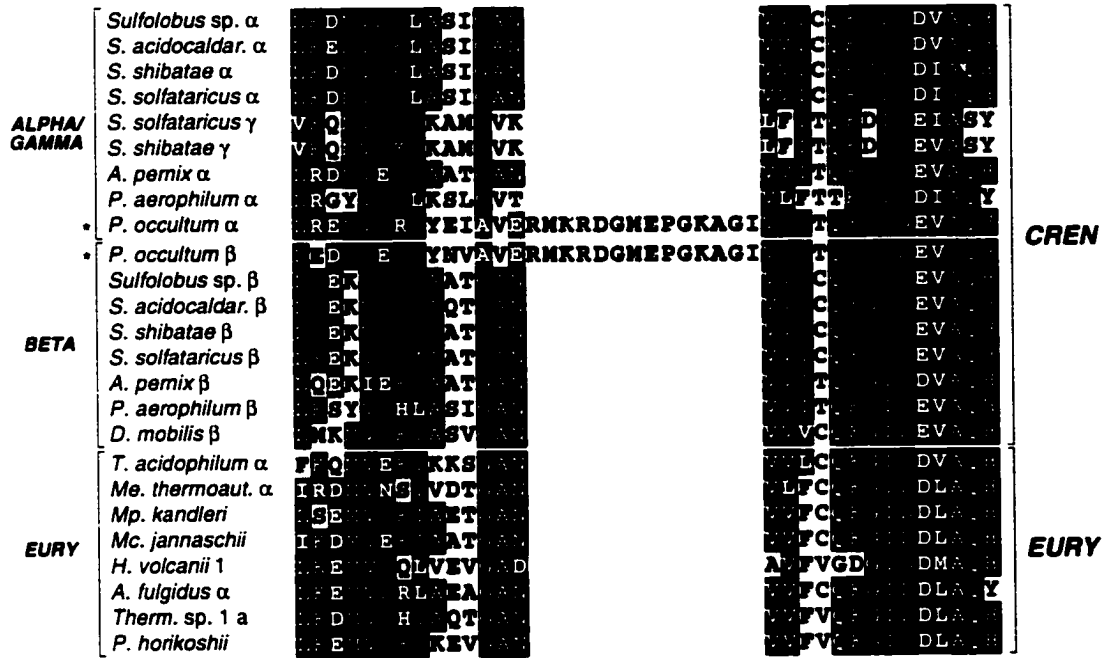
Chaperonin sequences are now available from a wide diversity of crenarchaeotes, as the complete sequences of three crenarchaeal genomes have recently been released (*Aeropyrum pernix* (Kawarabayasi *et al.* 1999), *Pyrobaculum aerophilum* (<http://www.doe-mbi.ucla.edu/PA/>), and *Sulfolobus solfataricus* (<http://www.cbr.nrc.ca/sulphome/>)). The tree shown in Figure 1.3 reveals that the gene duplication producing the  $\alpha$  and  $\beta$  subunits occurred in the common ancestor of the *Sulfolobus*, *Pyrodictium*, *Desulfurococcus*, *Aeropyrum* and *Pyrobaculum* lineages. This can be inferred because the  $\beta$  sequences from these organisms cluster together with high statistical support, to the exclusion of the  $\alpha$  sequences. The gene duplication producing the  $\gamma$  subunit appears to have occurred more recently, as the *S. solfataricus* and *S. shibatae*  $\gamma$  sequences cluster strongly with the  $\alpha$  sequences from four different *Sulfolobus* species, to the exclusion of those in *Pyrodictium occultum*, *Aeropyrum pernix* and *Pyrobaculum aerophilum*.

Where complete genome sequences are available, several instances of chaperonin gene loss can also be inferred. For example, the *Thermococcus* strains

KS-1 and KS-8 each have two subunits, 'a' and 'b', which are 80.6% identical: genes encoding these subunits duplicated and diverged prior to the splitting of these two lineages (Figure 1.3). Significantly, three completely sequenced *Pyrococcus* genomes (*P. furiosus*, *P. abyssi* and *P. horikoshii*) reveal the presence of a single chaperonin subunit gene. When compared to the *Thermococcus* proteins, the three *Pyrococcus* sequences appear most similar to the 'b' subunit from the *Thermococcus* strains (Figure 1.3). This suggests that the duplication and divergence of the 'a' and 'b' chaperonin subunit genes in these lineages occurred in their common ancestor, and that subunit 'a' was subsequently lost in the *Pyrococcus* lineage. Chaperonin gene loss can also be inferred for the halophilic euryarchaeote *Halobacterium* sp. NRC-1. The genome of this organism has recently been completely sequenced (Ng *et al.* 2000) and possesses genes encoding two subunits. However, a closely related halophile, *Haloferax volcanii*, has three chaperonin subunits, and the two *Halobacterium* subunits cluster strongly with two of the three *H. volcanii* sequences in phylogenetic analyses (Figure 1.3). Again, one can infer that the ancestor of these two organisms had three subunits and that the third subunit was lost in *Halobacterium* sp. NRC-1.

### **Gene conversion in archaeal chaperonins**

Close examination of the  $\alpha$  and  $\beta$  genes in the crenarchaeote *Pyrodicticum occultum* revealed that both subunits contain an identical 14 amino-acid insertion in their respective apical domains that is absent from all other archaeal sequences (Figure 1.4). The distribution of this insertion is completely at odds with the inferred phylogeny, as the duplication producing the  $\alpha$  and  $\beta$  subunits clearly predates the radiation of all crenarchaeotes examined thus far (Figure 1.3). A DNA sequence alignment of this region is shown in Figure 1.5 and reveals that the unexpectedly high degree of sequence identity shared between the two genes



**Figure 1.4** An homologous insertion in the  $\alpha$  and  $\beta$  chaperonin subunits of *Pyrodictium occultum*. A protein sequence alignment of the insertion and flanking region for all known crenarchaeal chaperonins (CREN) is shown, and the different subunit families ( $\alpha$ ,  $\beta$  and  $\gamma$ ) are indicated. Representative euryarchaeal sequences (EURY) are also included. Residues present in at least 60% of the sequences are shaded black and conservative substitutions are shaded gray. The *P. occultum* sequences are highlighted with an asterisk (\*).

**Figure 1.5** DNA sequence alignment of a putative gene conversion tract between the *P. occultum*  $\alpha$  and  $\beta$  subunit genes. The alignment shows a portion of coding sequence surrounding the 14 amino acid (42 nucleotide) insertion in the apical domain. All crenarchaeal and representative euryarchaeal sequences are present. The *P. occultum* sequences are highlighted with an asterisk, and the stretch of sequence showing particularly high sequence identity between the two *P. occultum* genes is shaded gray. The boundaries of this region were determined by the gene conversion detection program GENECONV (Sawyer 1989; see text). Dots (.) indicate an identical residue to that in the reference sequence (*P. occultum*  $\alpha$ ) and dashes (-) indicate gaps. The reference coordinates correspond to the *P. occultum*  $\alpha$  subunit gene.



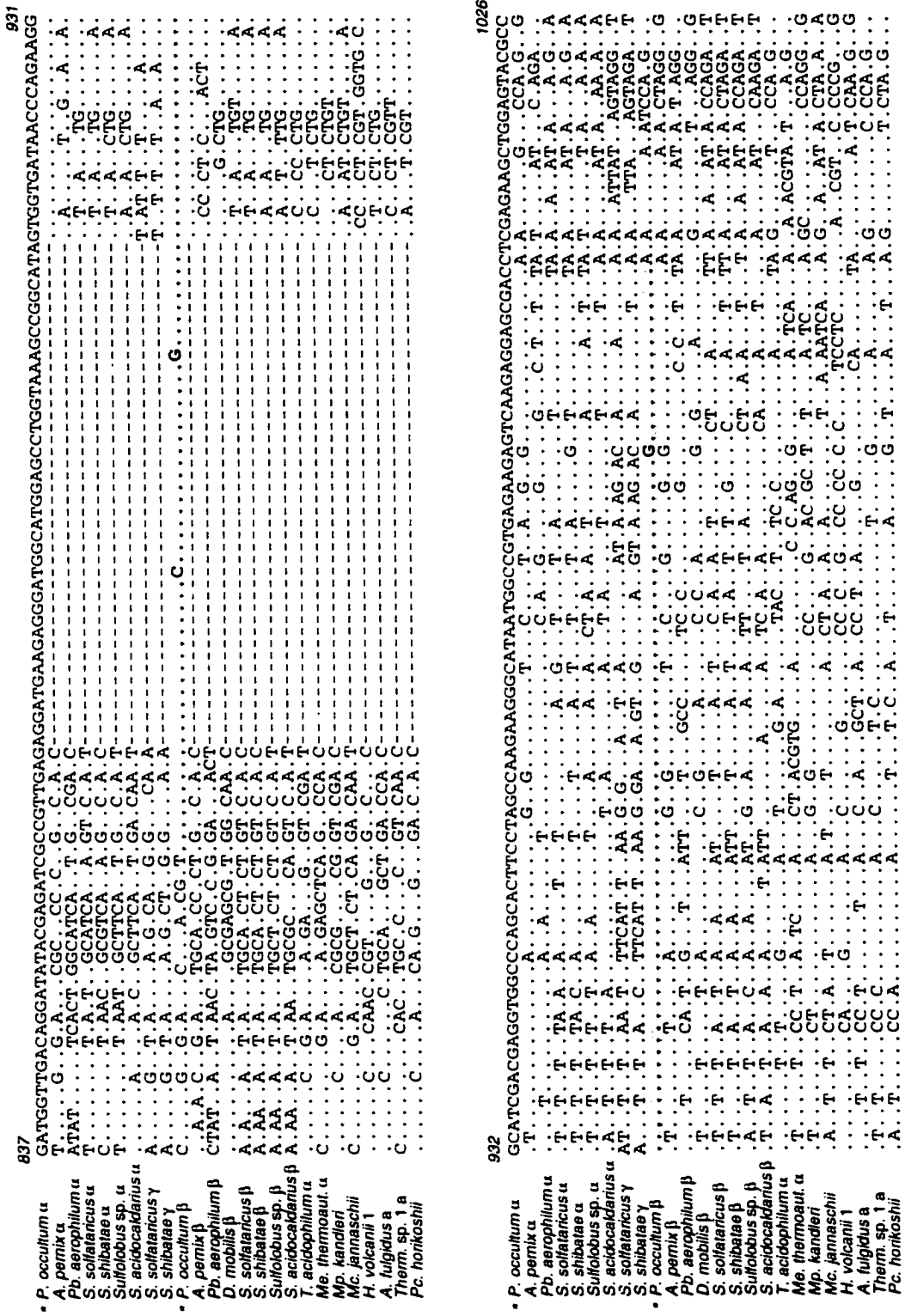
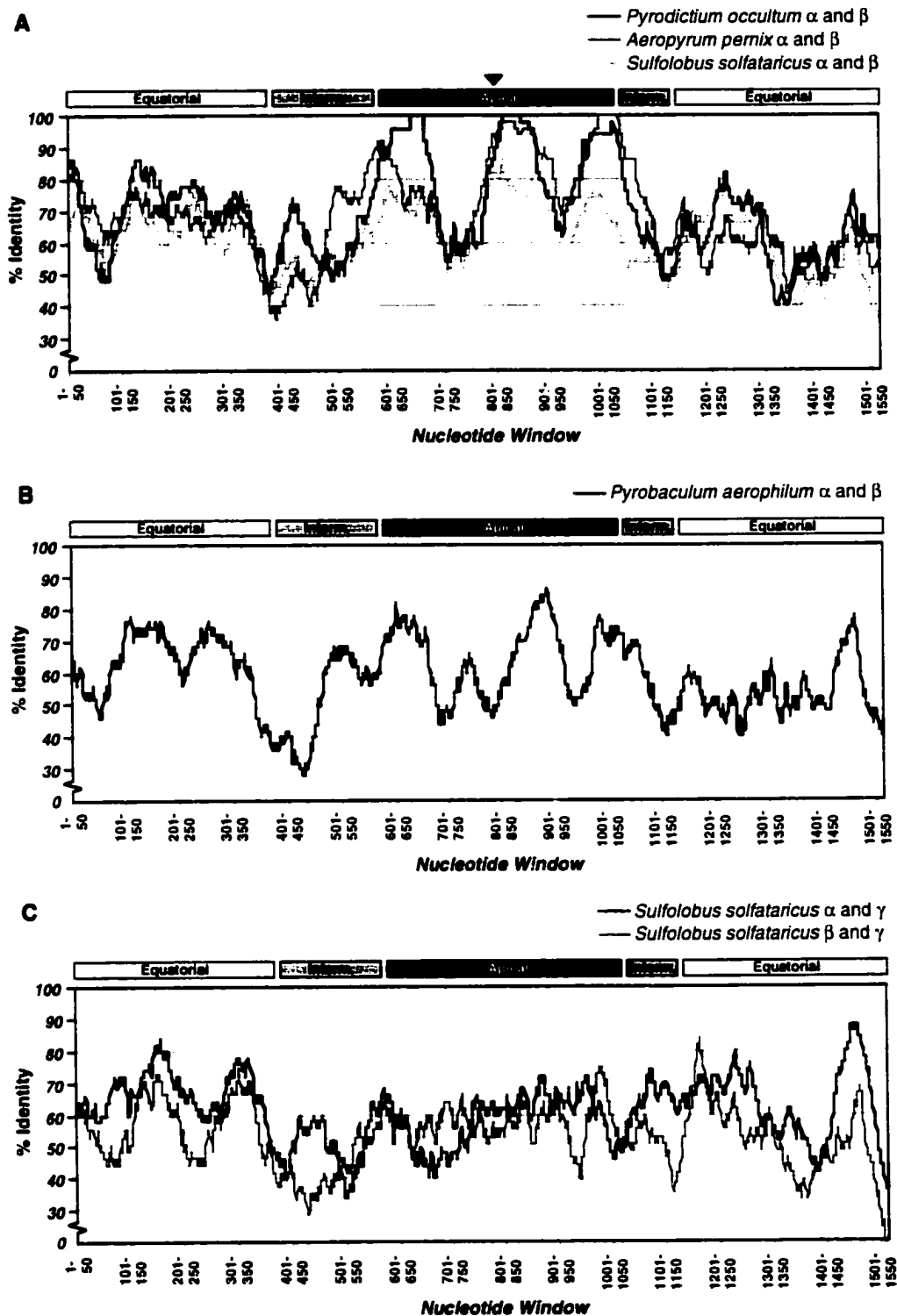


Figure 1.5 DNA sequence alignment of a putative gene conversion tract between the P. occultum alpha and beta subunit genes

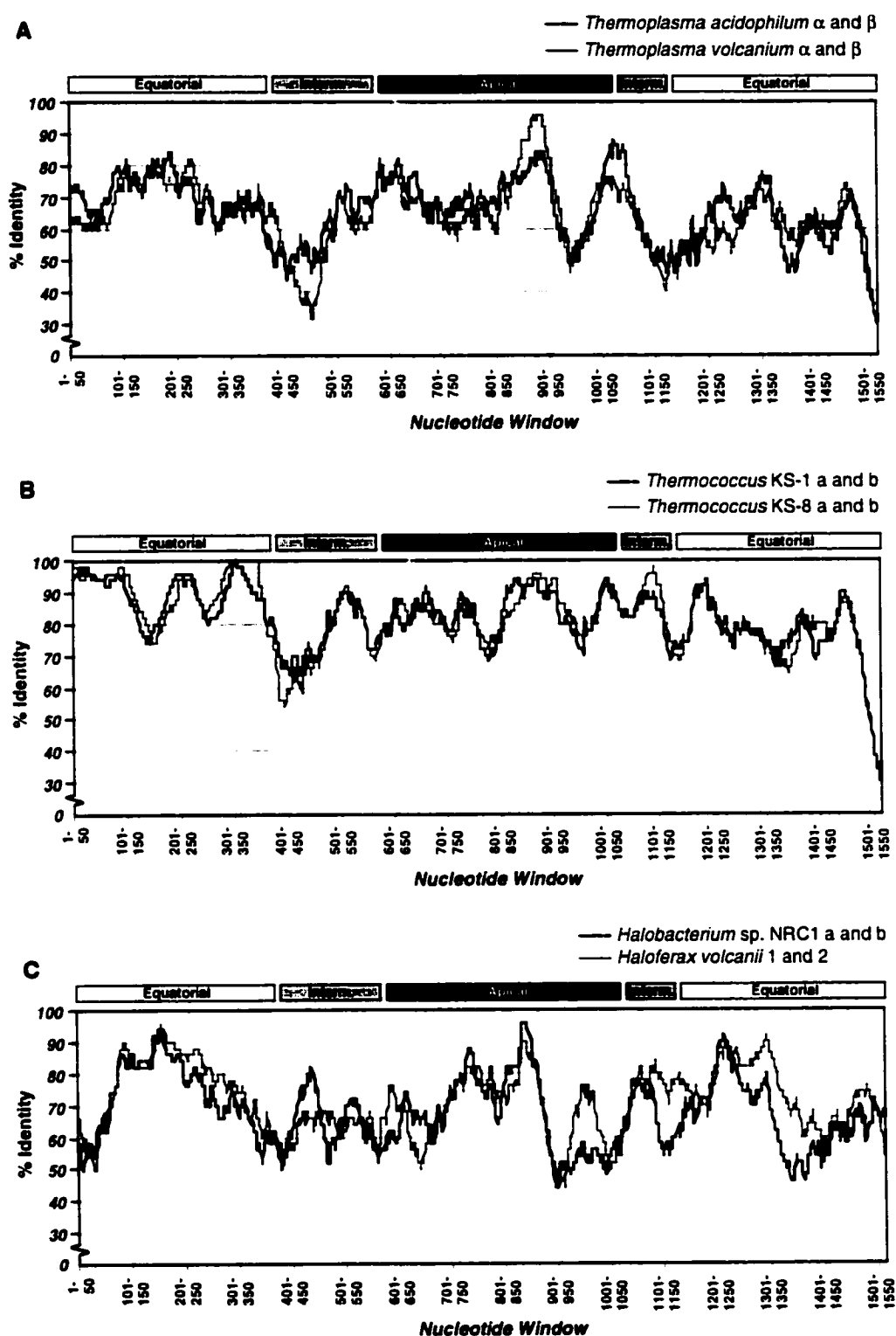
in the insertion also extends into the flanking area, mostly in the 3' direction. The most parsimonious explanation for the presence of this insertion in both *P. occultum* genes is that it arose in one of the paralogs sometime after the divergence of *P. occultum* from other crenarchaeal lineages and was subsequently 'passed'—by gene conversion—to the other. To investigate the possibility that gene conversion has played a significant role in the evolution of duplicate chaperonin genes in archaea, the data were analyzed using several different methods aimed at detecting anomalously evolving regions within the genes.

Figure 1.6 shows the results of % identity sliding window analyses performed on duplicate crenarchaeal genes. The % identity shared between pairs of sequences was determined in 50-nucleotide windows across the alignment, advancing the window in one-nucleotide increments. To ensure that the % identity spectra for all pairs of sequences were directly comparable, all pairwise alignments used in the calculations contained the same positions (1557 sites) and were devoid of gaps. When the *P. occultum*  $\alpha$  and  $\beta$  sequences were compared in this manner (Figure 1.6A), three stretches of extremely high sequence identity were readily apparent in the apical domain region (the second of which flanked the 42 nucleotide insertion discussed above; see Figure 1.5). Surprisingly, another crenarchaeote, *Aeropyrum pernix*, possesses 'peaks' of high sequence identity in two of these same areas (between nucleotide windows 801-850 and 901-950 and around window 1001-1050). The  $\alpha/\beta$  identity spectra for *Sulfolobus solfataricus* (Figure 1.6A) and *Pyrobaculum aerophilum* (Figure 1.6B) also show peaks in these regions, although to a much lesser degree. In contrast, the  $\alpha/\gamma$  and  $\beta/\gamma$  comparisons in *S. solfataricus* (Figure 1.6C) reveal a much lower level of sequence identity shared between these genes in the apical domain.

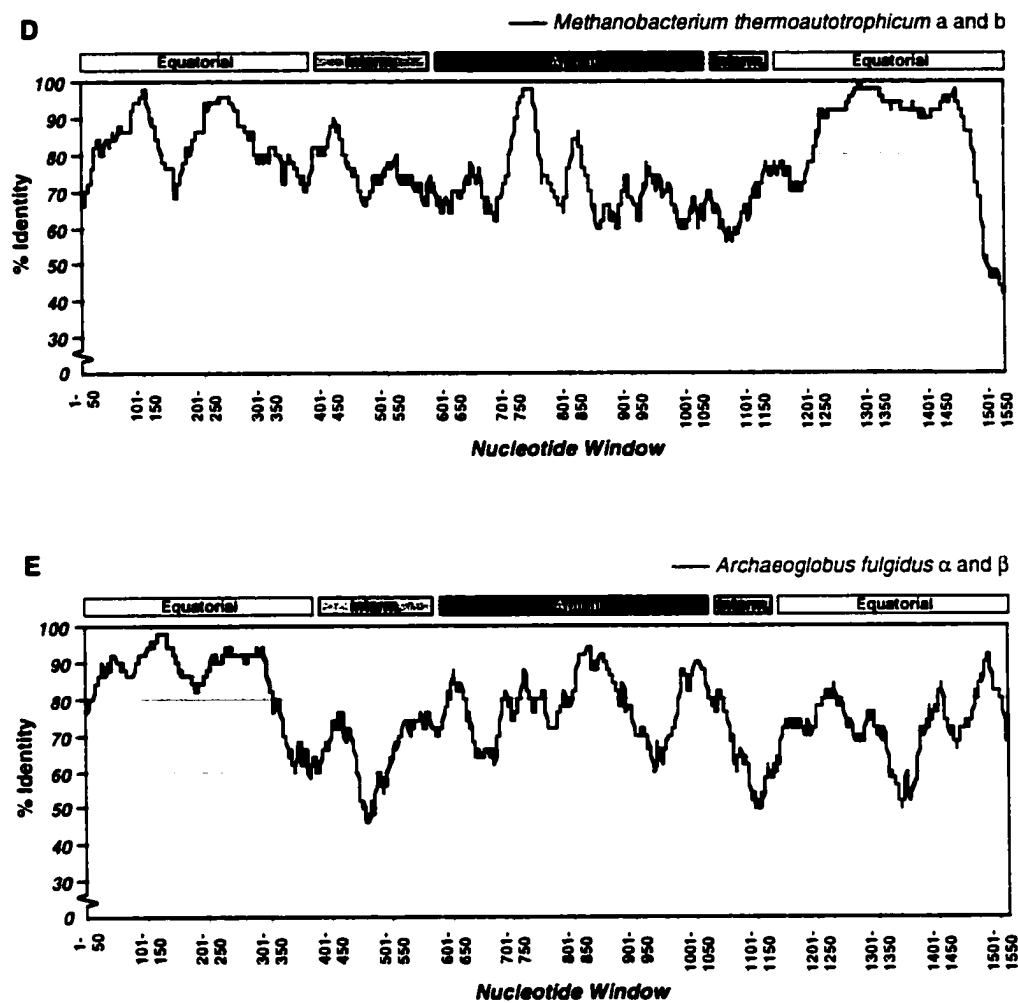
The % identity plots for euryarchaeotes are shown in Figure 1.7, and emphasize the wide range in the degree of sequence identity shared between



**Figure 1.6** Sliding window analyses of percent (%) identity shared between duplicate crenarchaeal chaperonin genes. (A-C) % identity plots for various pairs of sequences calculated with a 50-nucleotide window moved across the complete alignment in 1-nucleotide increments. For reference, the regions of the molecule corresponding to the equatorial, intermediate and apical domains of the chaperonin are indicated. In (A), the area of the 14-amino-acid insertion in the *P. occultum*  $\alpha$  and  $\beta$  subunits (see text) is indicated by a triangle and the putative gene conversion tracts identified by GENECONV (see text) are shaded gray. An alignment containing 1557 nucleotide positions was used for all analyses.



**Figure 1.7** Sliding window analyses of percent (%) identity shared between duplicate euryarchaeal chaperonin genes. (A-E) % identity plots for the indicated pairs of sequences inferred with a 50-nucleotide sliding window moved across the alignment in 1-nucleotide increments. Putative gene conversion tracts identified by GENECONV (see text) are shaded gray. The regions corresponding to the equatorial, intermediate and apical domains of the chaperonin are indicated. An alignment containing 1557 nucleotide positions was used for all analyses.



**Figure 1.7** Sliding window analyses of percent (%) identity shared between duplicate euryarchaeal chaperonin genes

duplicates in the different lineages (e.g., compare Figure 1.7A and B). Given that phylogenetic analyses indicate that the duplications and divergences in euryarchaeal chaperonins occurred independently of one another and of those in crenarchaeotes (Figure 1.3), it is interesting that the profiles are quite similar. For example, most pairwise comparisons reveal a reduced level of sequence identity at the boundary between the amino-terminal equatorial and intermediate domains (around nucleotide window 401-450), a pattern also seen in the crenarchaeal sequences. Several of the profiles also possess a 'spike' in sequence identity between nucleotide windows 101-150 and 201-250, as well as in several regions of the apical domain (e.g., compare Figure 1.6A to Figure 1.7A and E). There were several exceptions, however, as the *Methanobacterium thermoautotrophicum* a and b genes share a long (and unique) stretch of high sequence identity toward their 3' ends (Figure 1.7D). The  $\alpha$  and  $\beta$  genes in *Archaeoglobus fulgidus* share a region of high identity at the 5' end (Figure 1.7E).

### **Statistical tests for detecting gene conversion**

Attempts to determine the statistical significance of these observations using Sawyer's gene conversion detection program GENECONV (Sawyer 1989; see Materials and Methods) produced ambiguous results. Initially, a DNA sequence alignment containing the full diversity of euryarchaeal and crenarchaeal sequences (39 taxa, 1815 nucleotide positions) was used as GENECONV input. When all sites in the alignment were considered, numerous pairs of sequences in both crenarchaeotes (*P. occultum* and *A. pernix*) and euryarchaeotes (*T. volcanium*, *Thermococcus* KS-8, *M. thermoautotrophicum* and *A. fulgidus*) showed possible evidence for partial gene conversion, with significance values ranging from  $p=2.58 \times 10^{-3}$  to  $7.75 \times 10^{-10}$ . As expected, the putative gene conversion tracts corresponded to spikes of high sequence identity in their

respective % identity spectra, and these regions are highlighted gray in Figures 1.6 and 1.7 for comparison. As suggested by the % sequence identity plot, the two candidate gene conversions in the *A. pernix*  $\alpha$  and  $\beta$  genes were in the same region as two of the three tracts in *P. occultum* (between nucleotide windows 801-850 and 901-950 and around window 1001-1050; Figure 1.6A).

To control for the possibility that a given stretch of identity shared between two sequences is due to functional constraints at the protein level and not gene conversion, Sawyer (1989) suggests focussing on silent polymorphic sites (degenerate positions in the alignment whose codons all specify the same amino acid; see Materials and Methods). Unfortunately, this criterion makes GENECONV poorly suited for detecting conversions between distantly related protein-coding genes, as the presence of divergent sequences in the alignment drastically reduces the number of silent polymorphic sites considered by the program (Drouin *et al.* 1999). Under this more strict criterion, none of the putative gene conversion tracts identified previously were deemed significant (including the prominent one in *P. occultum*), despite the fact that many of them were fairly obvious by visual inspection. In an attempt to minimize this effect, closely related subsets of the data were analyzed separately. However, most of the possible gene conversion tracts were still not identified as being statistically significant under the 'silent site only' criterion. The notable exceptions were those in *P. occultum* and *A. pernix*. When certain subsets of crenarchaeal sequences were analyzed, the entire apical domain in both the *P. occultum* and *A. pernix*  $\alpha$  and  $\beta$  sequences was identified as a possible gene conversion tract, although with modest statistical support ( $p=7.45 \times 10^{-2}$  and  $p=1.08 \times 10^{-1}$ , respectively). From these experiments, it is thus not clear whether a single gene conversion event homogenized the entire apical domain coding region in these

lineages or whether the three distinct peaks of high identity shared between the  $\alpha$  and  $\beta$  genes were caused by independent conversions.

### **A phylogenetic approach to gene conversion**

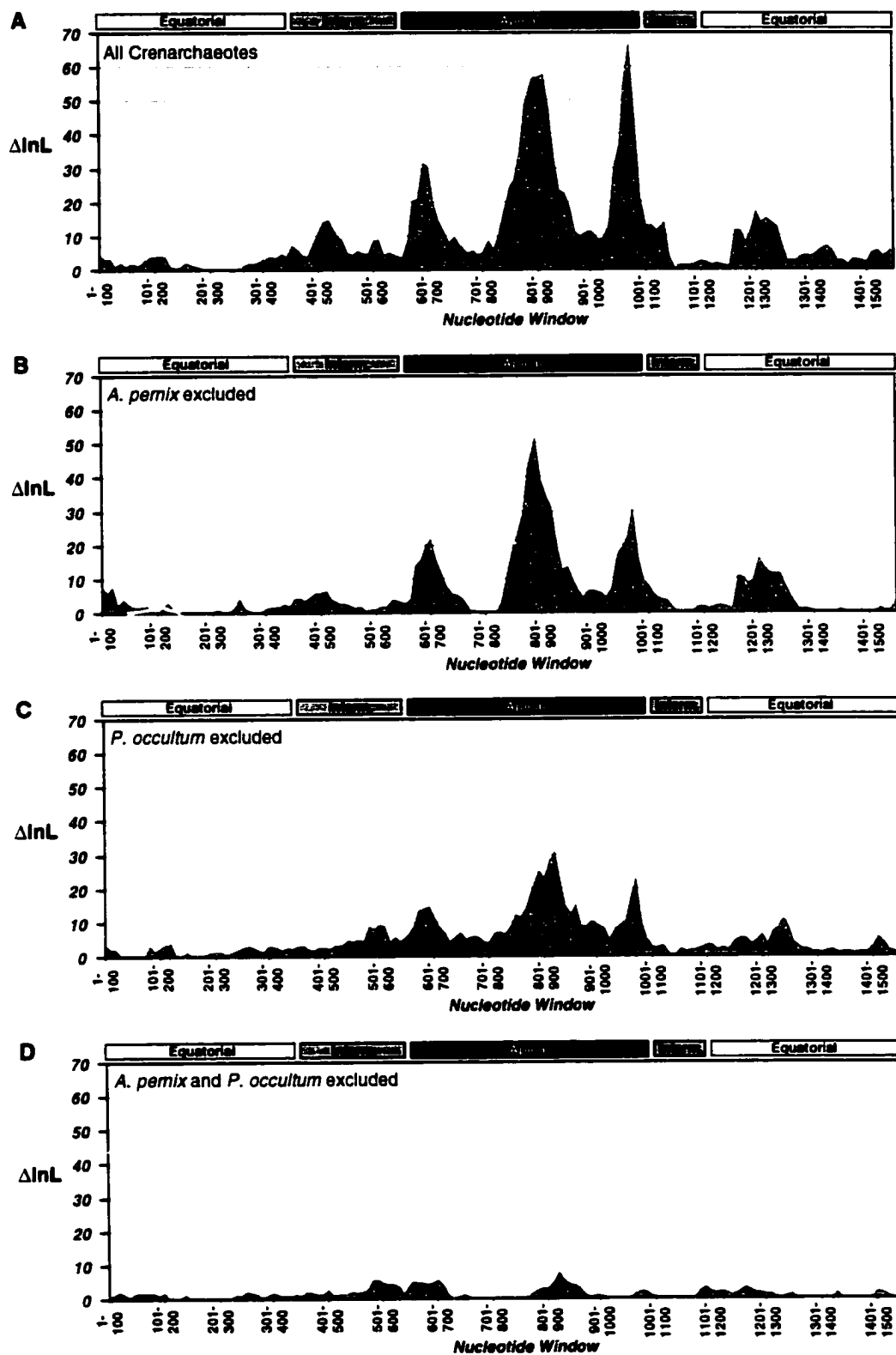
The paralogy in the  $\alpha$  and  $\beta$  subunits of crenarchaeotes presents a unique opportunity to study gene conversion from a phylogenetic perspective. Since this paralogy predates organismal divergence, partial conversions between  $\alpha$  and  $\beta$  subunit genes that occur *after* lineage splitting should be identifiable as regions of the gene with a phylogenetic history contradictory to the molecule as a whole. In these regions,  $\alpha$  and  $\beta$  paralogs would show an organism-specific branching pattern instead of a gene-specific one.

With this in mind, alignments were 'scanned' for regions of incongruent phylogenetic signal using a novel maximum likelihood (ML) approach. With PAUP\* (Swofford 1998), ML trees were inferred from 100-nucleotide windows of the alignment, systematically advancing the window across the alignment in 10-nucleotide increments (see Materials and Methods). For each window, the log likelihood (lnL) of the best tree was obtained, as was the lnL of the data present in the window given the topology obtained from a ML analysis of the complete alignment. The difference between these two likelihoods ( $\Delta$ lnL) reflects the degree of congruence between the phylogenetic history of a given 100-nucleotide subset of the data and that of the whole molecule.

The results of the ML sliding window analyses for all full-length crenarchaeal  $\alpha$  and  $\beta$  subunit genes are presented in Figure 1.8. The data are presented as a plot of  $\Delta$ lnL values for the various 100-nucleotide windows spanning the gene. Three large peaks in the apical domain are evident. Again, these peaks represent areas of incongruence between the phylogenetic signal in the whole gene. Interestingly, when the *P. occultum* and *A. pernix*  $\alpha$  and  $\beta$



**Figure 1.8** Maximum likelihood (ML) sliding window analyses of crenarchaeal chaperonin DNA sequences. For each dataset the best ML tree was inferred from 100-nucleotide windows across the alignment in 10-nucleotide increments, as was the log likelihood ( $\ln L$ ) of the data in each window given the topology inferred from the whole molecule. The difference between the two ( $\Delta \ln L$ ) was plotted for each of the 146 windows spanning the whole alignment. (A)  $\Delta \ln L$  profile obtained with all full-length  $\alpha$  and  $\beta$  crenarchaeal sequences (*Pyrodictium occultum*, *Aeropyrum pernix*, *Pyrobaculum aerophilum*, *Sulfolobus* sp. S7, *S. solfataricus*, *S. shibatae*). (B-D)  $\Delta \ln L$  profiles obtained with the *A. pernix* and/or *P. occultum* sequences excluded. For reference, the regions of the molecule corresponding to the equatorial, intermediate and apical domains of the chaperonin are indicated.



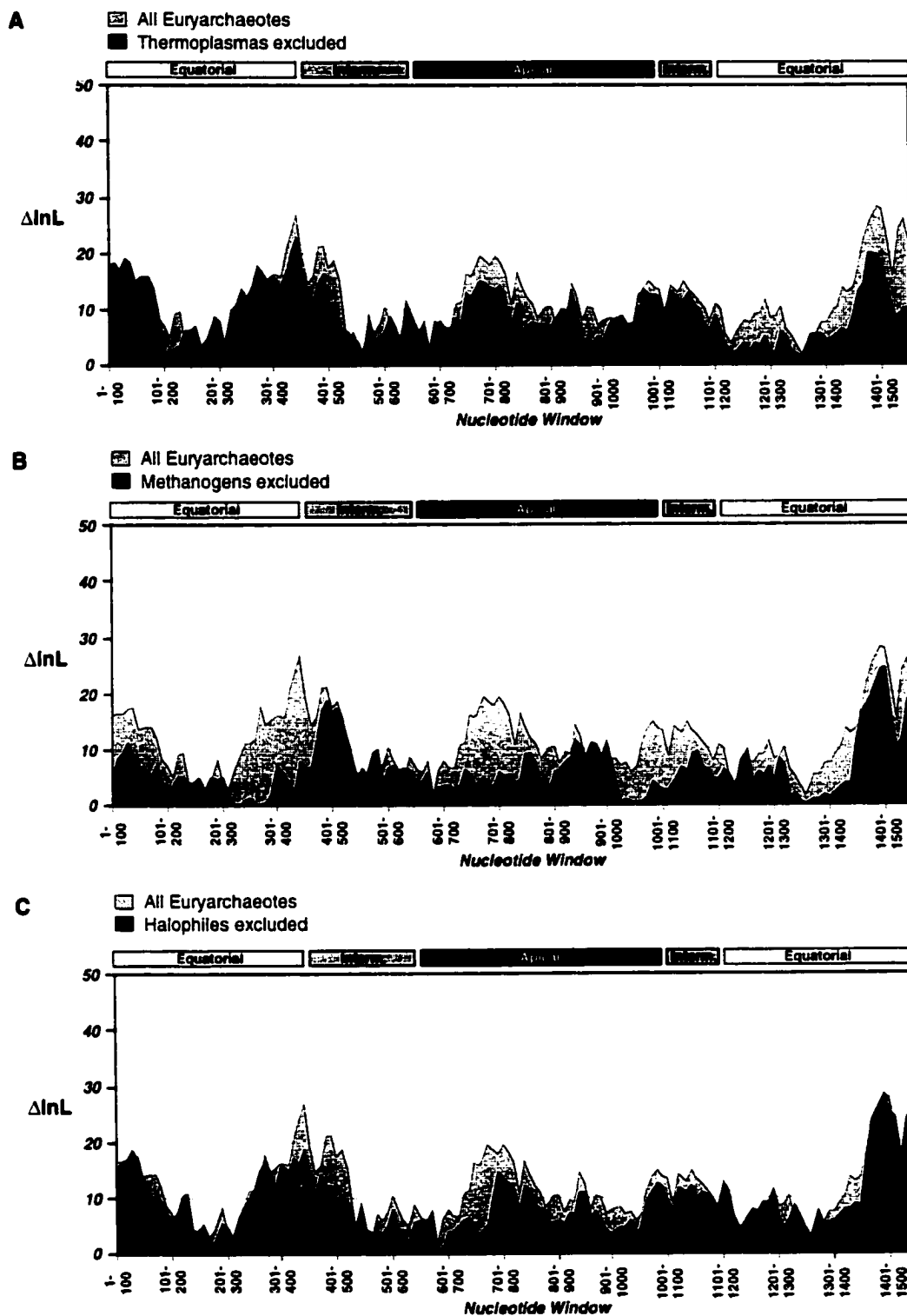
**Figure 1.8** Maximum likelihood (ML) sliding window analyses of crenarchaeal chaperonin DNA sequences

sequences were systematically removed and the analyses repeated (Figure 1.8B-D), the bulk of the phylogenetic incongruence present in the data disappeared.

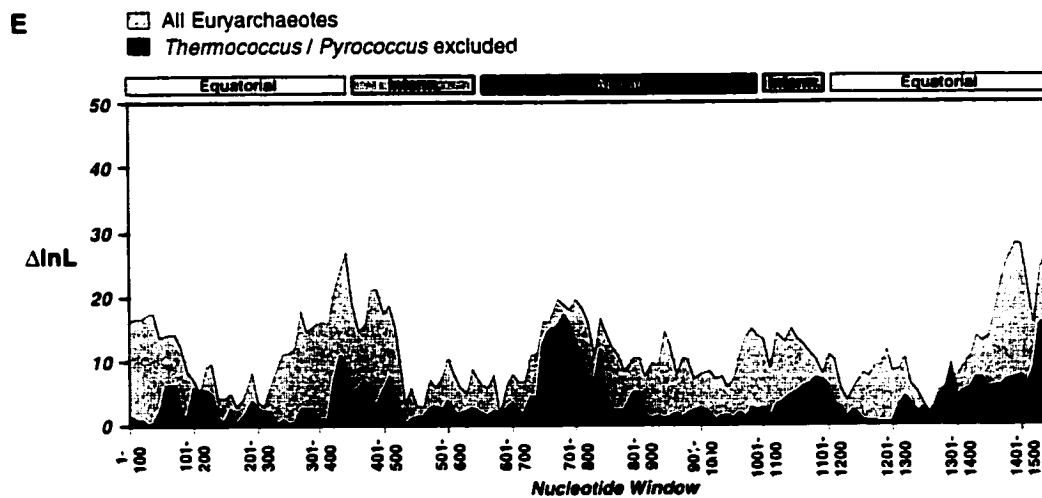
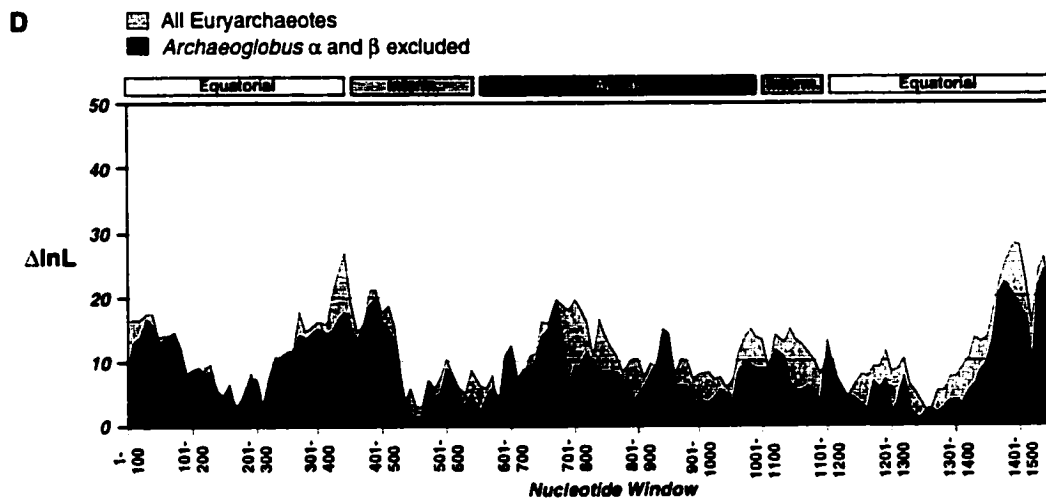
The euryarchaeal chaperonin dataset was also examined in this fashion. The fact that the euryarchaeal chaperonin tree is characterized by lineage-specific gene duplications means that partial gene conversions may go undetected by the ML sliding window method. This is because paralogs in a given genome are *already* more similar to one another than to other genes in the alignment and thus branch together to the exclusion of other sequences in the phylogeny, regardless of whether or not they have experienced gene conversion. Indeed, the  $\Delta\ln L$  spectrum obtained with the complete euryarchaeal dataset (Figure 1.9) was not as striking as that observed for crenarchaeotes, although a slightly higher background  $\Delta\ln L$  was observed. This may be (at least in part) due to the presence of more sequences in the analysis (21 compared to 12 for crenarchaeotes), and thus more possible alternative tree topologies for a given window of sequence. Several discrete regions of the molecule did appear to exhibit a somewhat increased  $\Delta\ln L$ , however, the most prominent of which included a region roughly between nucleotide windows 201-300 and 401-450 and a region surrounding window 1401-1500.

To determine which sequences in the alignment gave rise to the  $\Delta\ln L$  'peaks' (i.e., which sequences contained the phylogenetic incongruence), the major euryarchaeal clades (see Figure 1.3) were systematically removed from the alignment and the analyses were repeated. The results are shown in Figure 1.9, plotted against the spectrum obtained with all euryarchaeal sequences to allow for direct comparison. The removal of the Thermoplasmatales (Figure 1.9A), the halophiles (Figure 1.9C) and the *Archaeoglobus* sequences (Figure 1.9D) appeared to have little effect on the  $\Delta\ln L$  profile. In contrast, removing the methanogens

**Figure 1.9** Maximum likelihood (ML) sliding window analyses of euryarchaeal chaperonin DNA sequences. For each dataset the best ML tree was inferred from 100-nucleotide windows across the alignment in 10-nucleotide increments, as was the log likelihood ( $\ln L$ ) of the data in each window given the topology inferred from the molecule as a whole. The difference between the two ( $\Delta \ln L$ ) was plotted for each of the 146 windows spanning the complete alignment. (A-E) The  $\Delta \ln L$  profile obtained with all euryarchaeal sequences (21 in total) plotted against the  $\Delta \ln L$  profiles obtained when various euryarchaeal clades were removed. The regions of the molecule corresponding to the equatorial, intermediate and apical domains of the chaperonin are highlighted.



**Figure 1.9** Maximum likelihood (ML) sliding window analyses of euryarchaeal chaperonin DNA sequences



**Figure 1.9** Maximum likelihood (ML) sliding window analyses of euryarchaeal chaperonin DNA sequences

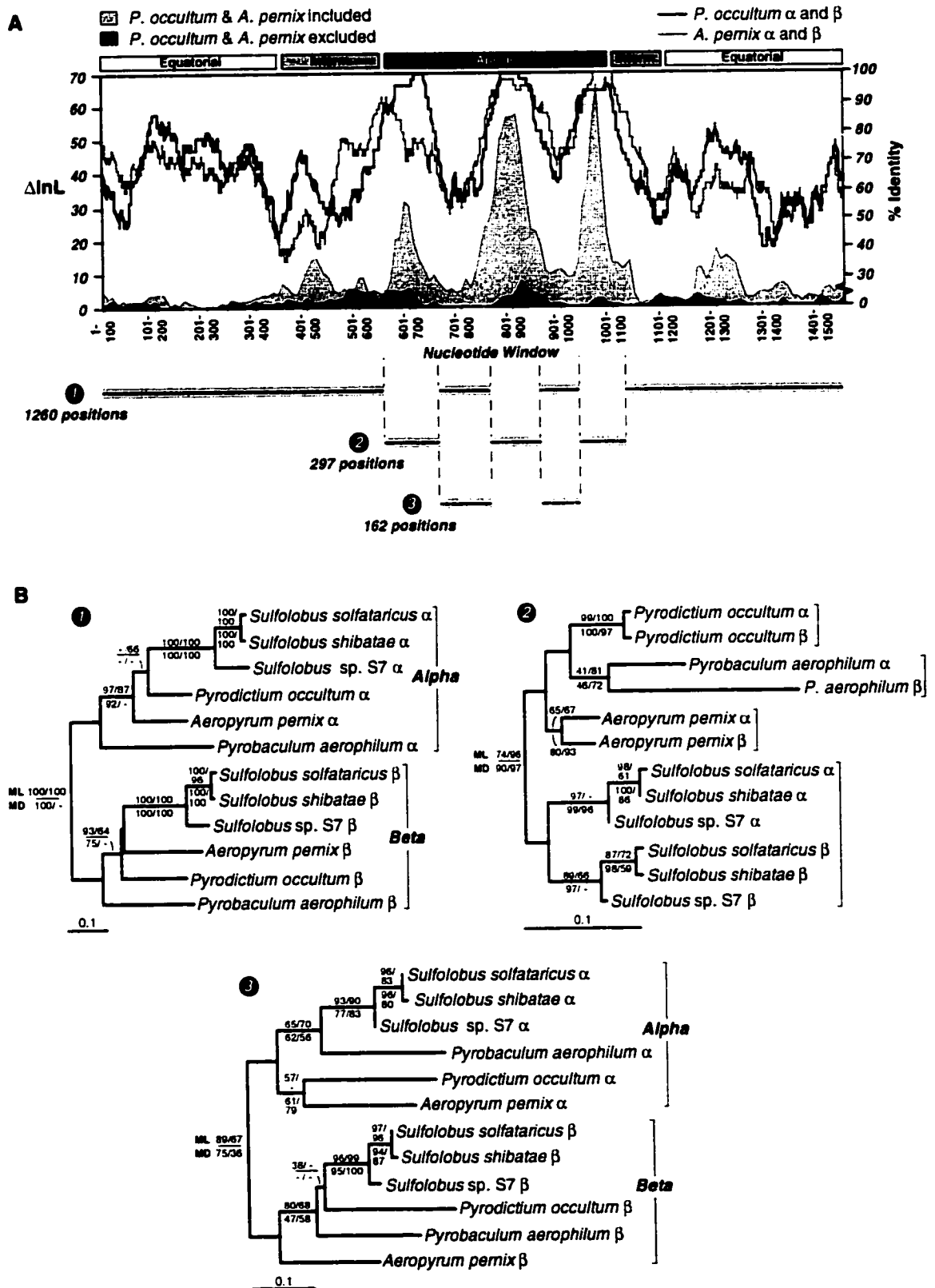
(*Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii* and *M. thermolithotrophicus*, *Methanopyrus kandleri*) resulted in a reduction in a peak in the apical domain around nucleotide window 701-800, as well as in the broad peak around window 301-400 (Figure 1.9B). The removal of the *Pyrococcus* / *Thermococcus* clade appeared to have the greatest effect (Figure 1.9E), substantially reducing the peaks in the area between windows 201-300 and 401-500 and around 1401-1500 (Figure 1.9E). Interestingly, the peak between windows 201-300 and 401-500 roughly corresponds to a region suggested (but not confirmed) to be a region of gene conversion between the *Thermococcus* KS-8 a and b genes (Figure 1.7B).

In an attempt to more accurately interpret the significance of the % sequence identity and ML sliding window results, phylogenetic analyses were performed on three discrete portions of the alignment. ML and ML-distance trees were inferred from (1) the complete alignment, minus the putative gene conversion tracts, (2) the regions corresponding to the putative gene conversion tracts in the apical domain of *P. occultum* and (3) the two regions between the gene conversion tracts. The results are shown in Figure 1.10. To illustrate the correspondence between the peaks in  $\Delta\ln L$  and regions of high sequence identity shared between the  $\alpha$  and  $\beta$  paralogs in *P. occultum* and in *A. pernix*, the results of the % identity and ML sliding window analyses were plotted against one another in Figure 1.10A. For reference, the  $\Delta\ln L$  spectrum obtained when these sequences were removed (Figure 1.8D) is also included.

As expected, phylogenies inferred from the complete alignment minus the putative gene conversion tracts (dataset 1) were identical to those obtained from protein sequence analyses of the complete molecule in separating the crenarchaeotes into two highly supported  $\alpha$  and  $\beta$  clades, irrespective of organism (Figure 1.10B, tree 1). In stark contrast, phylogenetic trees inferred from

**Figure 1.10** Incongruent phylogenetic signal in the apical domain of crenarchaeal chaperonins. (A) Results of the % identity and maximum likelihood (ML) sliding window analyses shown in Figures 1.6A and 1.8A and D, plotted against one another for comparative purposes. The coordinates of three subsets of the data selected for phylogenetic analysis are shown. Dataset 1 contained 1260 nucleotide positions, while dataset 2 contained 297 sites and corresponded to the putative gene conversion tracts identified in the  $\alpha$  and  $\beta$  subunit genes of *P. occultum* (see text). Dataset 3 contained 162 positions and corresponded to the regions between the putative gene conversion tracts in dataset 2. (B) ML trees inferred using the 1st and 2nd codon positions of the three datasets. Analyses were also performed using all 3 codon positions, and bootstrap support is given for both (1st and 2nd positions and all 3 positions, respectively). Bootstrap support values for ML analyses are given above the branch and ML-distance bootstrap values (for analyses with 1st and 2nd, and all 3 codon positions) are given below. The scale bar indicates the estimated number of substitutions per site.

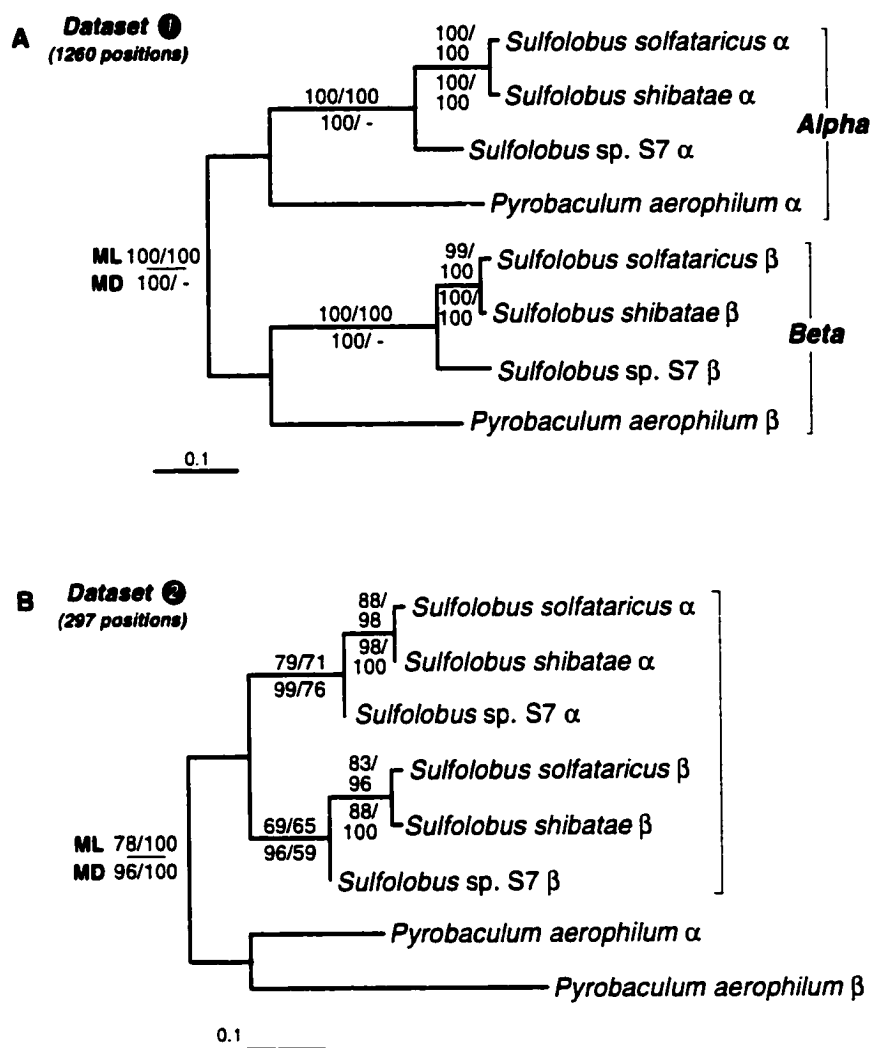




**Figure 1.10** Incongruent phylogenetic signal in the apical domain of crenarchaeal chaperonins

the gene conversion tracts themselves (dataset 2) showed the opposite pattern, with the  $\alpha$  and  $\beta$  paralogs branching in an organism-specific fashion (Figure 1.10B, tree 2). This result was expected for *P. occultum* and *A. pernix*, as their % identity spectra indicated that the  $\alpha$  and  $\beta$  genes in these organisms were extremely similar to one another in these discrete regions. Most unexpected was the observation that the *Sulfolobus*  $\alpha$  and  $\beta$  sequences also branched in a lineage-specific manner (considering the three closely related *Sulfolobus* species as a single group), as did those in *Pyrobaculum aerophilum*. This suggests that in addition to gene conversions in *P. occultum* and *A. pernix*, similar events have occurred between the  $\alpha$  and  $\beta$  subunits within *Pyrobaculum* and *Sulfolobus* evolution, although clearly not as recently. This is supported by the observation that within the *Sulfolobus* clade, the *S. solfataricus*, *S. shibatae* and *Sulfolobus* sp. S7  $\alpha$  sequences branch robustly with each other to the exclusion of the  $\beta$  sequences (Figure 1.10B, tree 2), suggesting that gene conversions have not occurred between the  $\alpha$  and  $\beta$  genes since the divergence of these three lineages.

To rule out the unlikely possibility that the anomalously evolving *P. occultum* and *A. pernix* sequences were causing the *Sulfolobus*  $\alpha$  and  $\beta$  and *Pyrobaculum*  $\alpha$  and  $\beta$  sequences to branch together artifactually in 'tree 2' from Figure 1.10B, datasets 1 and 2 were re-analyzed with these sequences excluded. The results show that with dataset 1 (the complete molecule, minus putative conversion tracts), the *Pyrobaculum aerophilum* and *Sulfolobus*  $\alpha$  sequences branch together with high bootstrap support, as do the *P. aerophilum* and *Sulfolobus*  $\beta$  genes (Figure 1.11A). However, with dataset 2, the inferred phylogeny again switches to one in which the genes branch in a lineage-specific fashion (Figure 1.11B). It is significant that in the phylogenies shown in Figures 1.10 and 1.11, the same topologies were almost always obtained when the trees were inferred from all positions in the alignment and when they were inferred from the first and



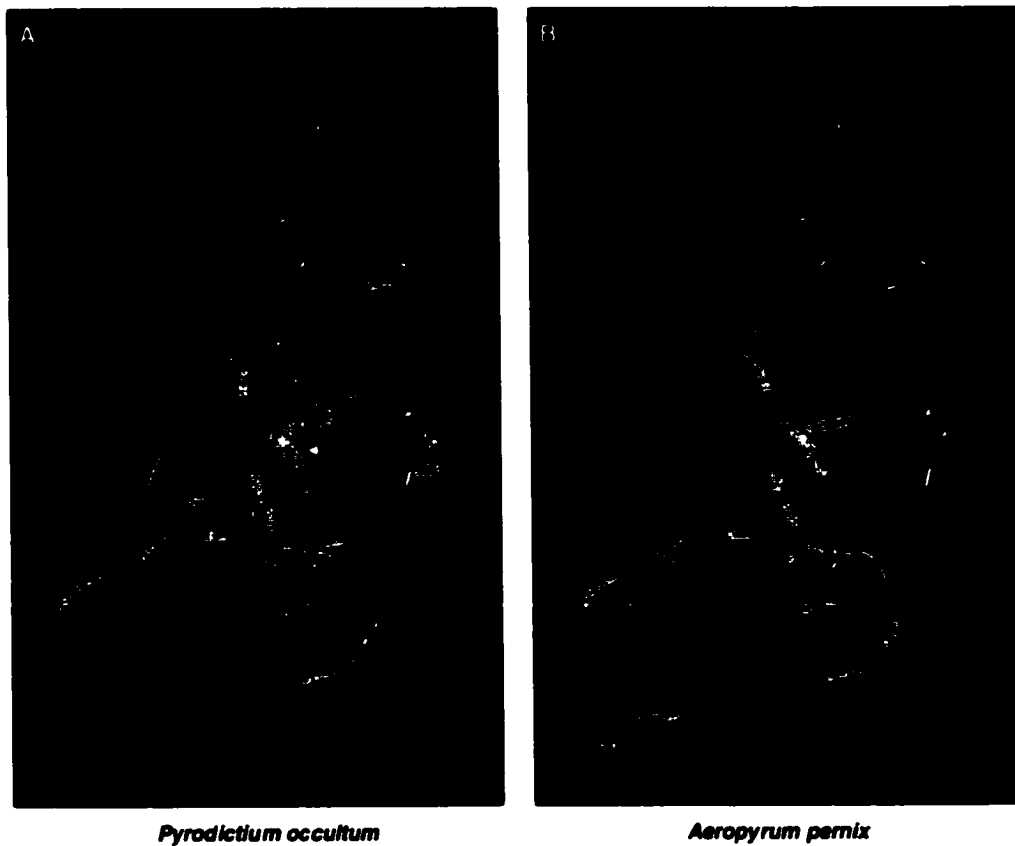
**Figure 1.11** Testing the removal of the *P. occultum* and *A. pernix* sequences on the incongruent phylogenetic signal in crenarchaeal chaperonins. (A) ML tree inferred from the 1260-nucleotide dataset used in Fig. 1.10. (B) ML tree inferred from the 297-nucleotide dataset containing the putative gene conversion tracts in *P. occultum* and *A. pernix*. Both trees were inferred using the 1st and 2nd codon positions. Analyses were also performed using all 3 codon positions. Bootstrap support is given for 1st and 2nd positions and all 3 positions, respectively. ML bootstrap values are given above the branch and ML-distance bootstrap values are given below. The scale bar indicates the inferred number of substitutions per site.

second codon positions only. This rules out the possibility that in the putative gene conversion regions, the  $\alpha$  and  $\beta$  paralogs branch together in an organism-specific fashion due to lineage-specific or genome-specific codon bias effects.

Another important conclusion can be drawn from the analysis of the small regions between the three putative gene conversion tracts. Phylogenetic trees inferred from the 162 nucleotides in this region resemble those obtained from whole molecule (minus the gene conversion tracts) and *not* those inferred from the gene conversion tracts themselves (Figure 1.10B, tree 3). This observation is consistent with a scenario whereby the discrete regions of high sequence identity shared between  $\alpha$  and  $\beta$  paralogs in the apical domain region are the result of multiple independent gene conversion events occurring *independently* in each crenarchaeal lineage, as opposed to a single event in each lineage.

### **Correlations with structure**

The apical domain of the chaperonin monomer has been shown to play an important role in the recognition and binding of substrate (for review, see Gutsche, Essen and Baumeister 1999; Leroux and Hartl 2000; Willison and Grantham 2001). Gene conversion events between paralogs in this region of the molecule could therefore have important implications for subunit-specific roles in protein folding in hetero-oligomeric chaperonin complexes. To address this issue, the putative gene conversion tracts between  $\alpha$  and  $\beta$  subunits in *P. occultum* and *A. pernix* were mapped onto the known crystal structure of the euryarchaeal chaperonin (the thermosome) in *Thermoplasma acidophilum* (Ditzel *et al.* 1998). The results are shown in Figure 1.12. Remarkably, these tracts map predominantly to the middle portion of the apical domain below the helical protrusion (helix H10), precisely in the region that, by analogy to the eukaryotic chaperonin system (Llorca *et al.* 2000; Llorca *et al.* 1999a), appears to mediate



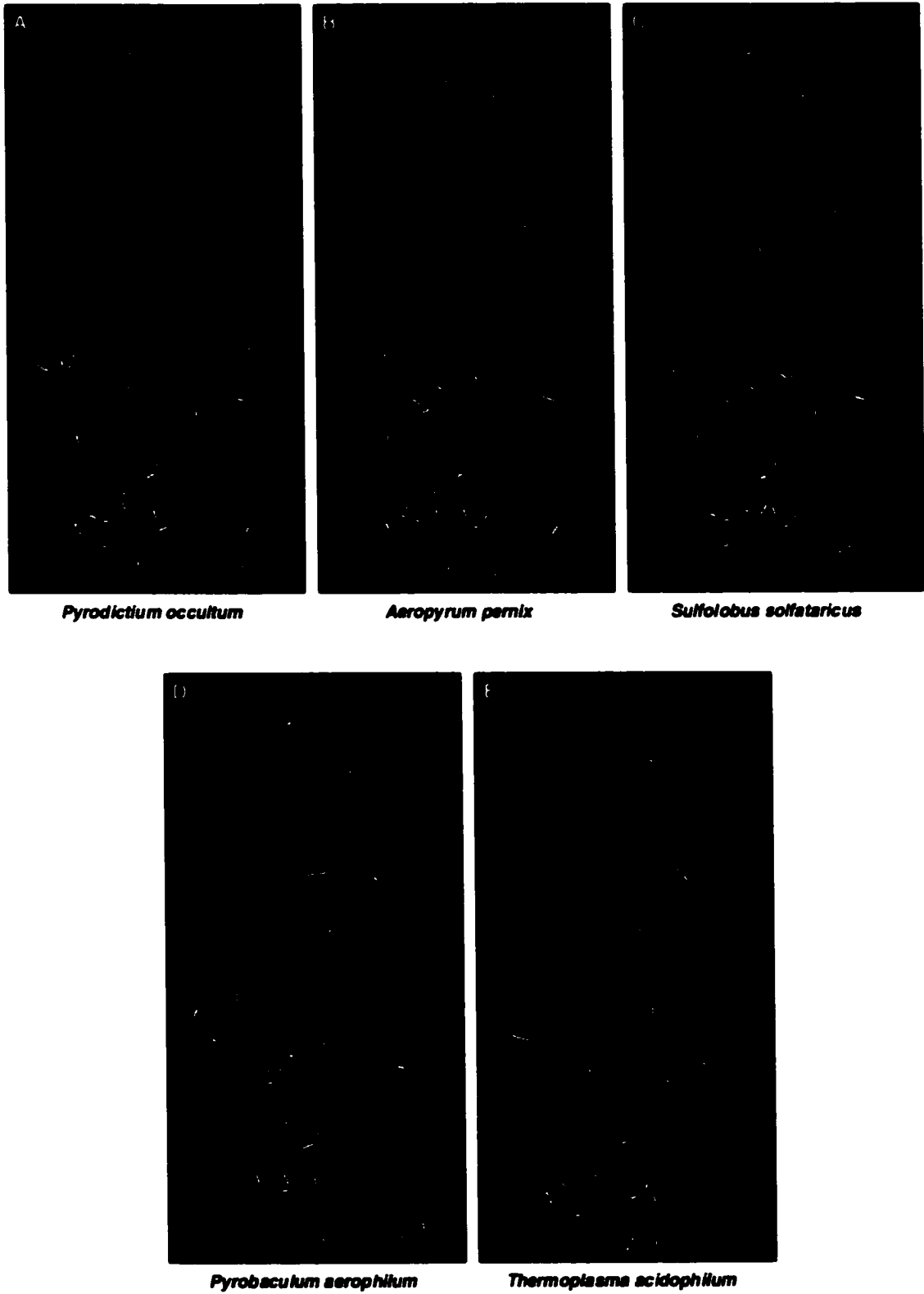
**Figure 1.12** Putative gene conversion tracts in *P. occultum* and *A. pernix* map to the apical domain. Regions of the *P. occultum* (A) and *A. pernix* (B) apical domains that have undergone gene conversion between the  $\alpha$  and  $\beta$  subunit genes were mapped onto the crystal structure of the  $\alpha$  subunit of the *Thermoplasma acidophilum* thermosome (Ditzel *et al.* 1998). The thermosome consists of two homologous subunits,  $\alpha$  and  $\beta$ ; since the two subunits are essentially identical in structure (Ditzel *et al.* 1998), the  $\alpha$  subunit was arbitrarily used. The apical domains are viewed from within the central cavity of the thermosome. The apical domain is colored orange and the regions corresponding to putative gene conversion tracts are highlighted yellow. In both *P. occultum* and *Aeropyrum pernix*, these tracts extend a short distance into the equatorial domain (highlighted blue). The area of the 14 amino acid insertion in the *P. occultum*  $\alpha$  and  $\beta$  genes is indicated by an arrow. Secondary structural elements are labeled according to Ditzel *et al.* (1998) (H=helix).

substrate binding. It is also significant that the 14-amino-acid insertion present in the *P. occultum*  $\alpha$  and  $\beta$  subunits maps to this area (Figure 1.12A)—both subunits likely form a unique (and identical) structure in this portion of the apical domain.

Another way to assess the structural/functional significance of the apical domain gene conversion data is to map onto the thermosome structure the positions of all the amino acid substitutions that have occurred between duplicate subunits in the different lineages. This was done for the  $\alpha$  and  $\beta$  subunits in the crenarchaeotes *P. occultum*, *A. pernix*, *Pyrobaculum aerophilum* and *S. solfataricus*, and for the euryarchaeote *Thermoplasma acidophilum*. The results are shown in Figure 1.13. For crenarchaeotes, the substitution patterns in the different species are correlated to a certain extent, since the  $\alpha/\beta$  gene duplication predates organismal divergence. Nevertheless, the regions of the molecule that have been 'homogenized' by gene conversion are quite prominent, especially for *P. occultum* and *A. pernix* (Figure 1.13A and B). The  $\alpha$  and  $\beta$  sequences in these organisms are absolutely identical to one another over much of the apical domain, but differ substantially in the helical protrusion region. The  $\alpha$  and  $\beta$  pairs for *Sulfolobus solfataricus* and *Pyrobaculum aerophilum* show a similar pattern (Figure 1.13C and D), with relatively few substitutions in the central portion of the apical domain but a large number of substitutions in the helical protrusion. Interestingly, the equatorial and intermediate domains, which provide the bulk of the subunit-subunit contact points, appear to possess the greatest number of substitutions between duplicate pairs (Figure 1.13A to D).

The substitution pattern for the  $\alpha$  and  $\beta$  subunits of *T. acidophilum* (Figure 1.13E) serves as a control for the crenarchaeote data, since the two proteins are the product of an independent duplication and divergence. Notably, the observed pattern was extremely similar to that in crenarchaeotes—the two subunits share a high degree of similarity in the mid portion of their apical domains (Figure

**Figure 1.13** Distribution of amino acid substitutions between the  $\alpha$  and  $\beta$  paralogs in crenarchaeotes and *T. acidophilum*. (A-D) For each crenarchaeote, the positions of all amino acid substitutions between the  $\alpha/\beta$  subunit pairs were mapped onto the crystal structure of the  $\alpha$  subunit of the *T. acidophilum* thermosome (Ditzel *et al.* 1998). (E) The substitution pattern between the  $\alpha$  and  $\beta$  subunits of *T. acidophilum*. The apical, intermediate and equatorial domains are colored orange, blue and pink, respectively. Secondary structural elements are labeled according to Ditzel *et al.* (1998) (H=helix, S=sheet).



**Figure 1.13** Distribution of amino acid substitutions between the  $\alpha$  and  $\beta$  paralogs in crenarchaeotes and *T. acidophilum*



1.13E), yet have experienced numerous substitutions in the helical protrusion region and in the equatorial and intermediate domains. Bosch, Baumeister and Essen (2000) recently crystallized the apical domain of the  $\alpha$  subunit in *T. acidophilum* and showed that its helical protrusion adopts a different conformation than that found in the crystal structure of the  $\beta$  subunit apical domain.

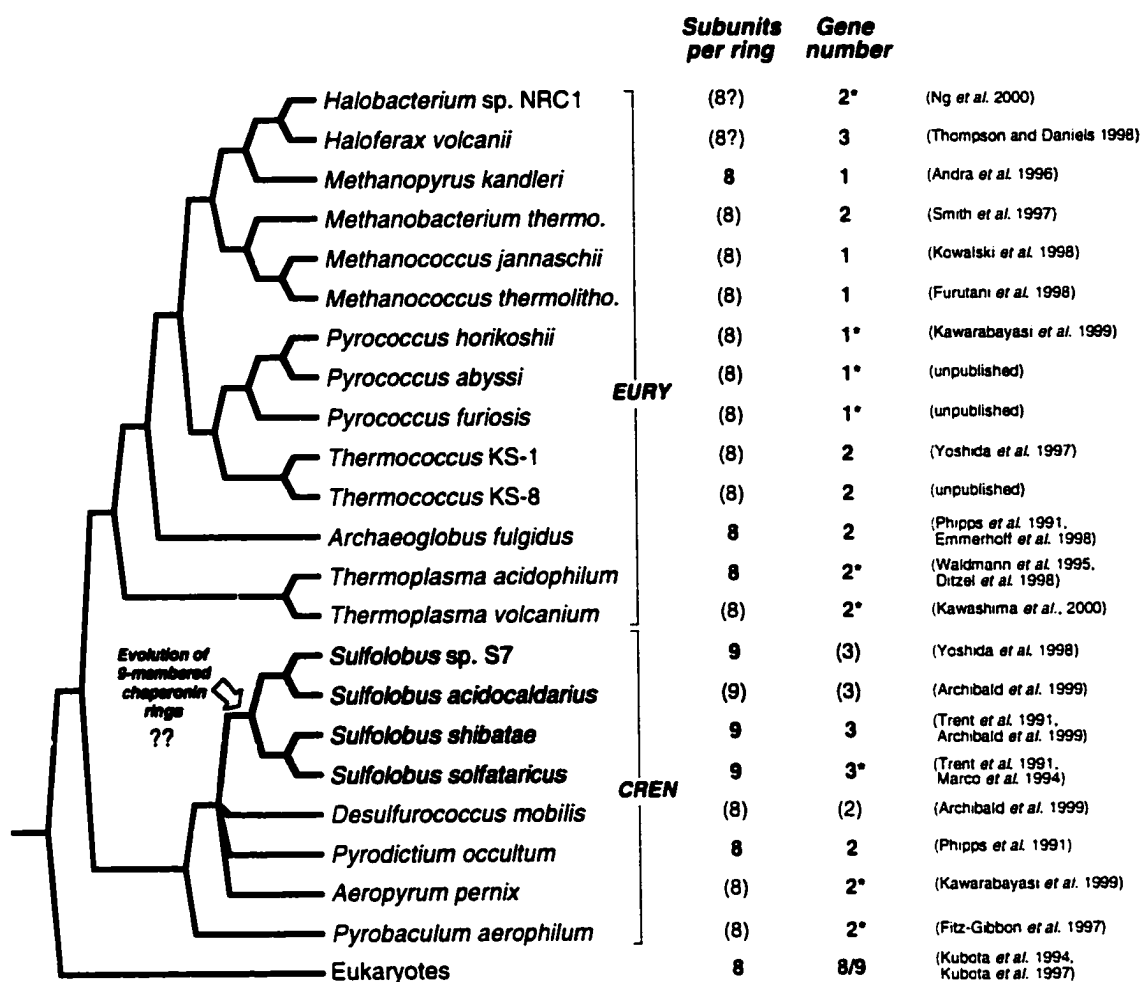
## DISCUSSION

Taken as a whole, archaeal chaperonins have had an extraordinarily complex evolutionary history. In euryarchaeotes, multiple lineage-specific gene duplications and divergences have occurred, while in crenarchaeotes, a duplication (into  $\alpha$  and  $\beta$  subunits) occurred in their common ancestor. The significance of these results in terms of chaperonin structure and function is that archaea appear to have made the transition from homo-oligomeric chaperonin complexes to hetero-oligomeric ones *at least* six times independently, and duplicate subunits have been lost at least twice (Figure 1.3). Gene conversion also appears to have played a role in archaeal chaperonin evolution, particularly in crenarchaeotes where it is clear that the  $\alpha$  and  $\beta$  subunit apical domains have been 'homogenized' several times independently. However, if it were not for the demonstrably homologous insertion shared between the  $\alpha$  and  $\beta$  paralogs in *P. occultum*, the gene conversions could easily have been overlooked. These results have important implications for understanding the function of duplicate subunits in hetero-oligomeric chaperonin complexes, and the origin and evolution of hetero-oligomerism itself.

## Evolution of archaeal chaperonin complex architecture

The discovery that at least some *Sulfolobus* species possess an additional chaperonin subunit gene ( $\gamma$ ) is particularly intriguing in light of observed differences in archaeal chaperonin complex architectures. As discussed previously, the chaperonin complex in the euryarchaeote *Thermoplasma acidophilum* is composed of  $\alpha$  and  $\beta$  subunits that alternate in each of its eight-membered chaperonin rings (Ditzel *et al.* 1998). While this arrangement also appears to exist in the crenarchaeote *Pyrodictium* (i.e. eight-membered rings with alternating subunits; Phipps *et al.* 1991, 1993), the organization of subunits in the nine-membered chaperonin rings observed in *Sulfolobus* species remains unclear. Figure 1.14 summarizes what is currently known from electron-microscopic and biochemical studies about the structure of chaperonin complexes in the archaeal lineages analyzed in this study. Many of the chaperonins in these organisms have, unfortunately, not been investigated with these techniques. Nevertheless, the widespread distribution of eight-membered chaperonin rings outside of the Sulfolobales—in all eukaryotes and euryarchaeotes examined thus far, as well as in *Pyrodictium*—suggests that this is an ancestral state, and that a transition from eight-membered to nine-membered rings occurred in a ‘recent’ ancestor of *Sulfolobus* (Figure 1.14).

Chaperonin complexes in *Sulfolobus* were initially thought to be homo-oligomeric (Trent *et al.* 1991) but were later found to possess two different subunit species, named  $\alpha$  and  $\beta$  (previously TF56 and TF55; Kagawa *et al.* 1995; Knapp *et al.* 1994; Nakamura *et al.* 1997). Since two subunits could not alternate equally in a nine-membered ring, Kagawa *et al.* (1995) proposed that *Sulfolobus* chaperonins consisted of two homo-oligomeric rings, one  $\alpha$ , and the other  $\beta$ . More recently, Ellis *et al.* (1998) examined two-dimensional crystals prepared from *S. solfataricus* and proposed that each ring is three-fold symmetric, with an



**Figure 1.14** Evolution of archaeal chaperonin complex architecture. A cladogram of archaeal relationships based on Figure 1.3 is shown on the left (CREN; crenarchaeotes, EURY; euryarchaeotes). The number of subunits per chaperonin ring and the number of known or inferred chaperonin genes present in each organism are shown on the right. *Subunits per ring*: bold values indicate known subunit stoichiometry from electron-microscopic studies; values in parentheses are predicted. *Gene number*: bold values indicate known gene number from sequence data; asterisks (\*) indicate that the total number of chaperonin genes is confirmed by complete genome sequence; values in parentheses are predicted. In mammals, nine chaperonin subunit genes are known; the ninth subunit shows testes-specific expression (Kubota *et al.* 1997).

$(\alpha_2\beta)_3$  arrangement. The discovery of a third chaperonin subunit-encoding gene raises the interesting possibility that the nine-membered rings of *Sulfolobus* chaperonins in fact have an  $(\alpha\beta\gamma)_3$  arrangement. The *S. solfataricus*  $\alpha$  and  $\gamma$  subunits described here should have nearly identical gross physical properties: the  $\gamma$  gene encodes a predicted protein of 535 amino acids [MW=58.771 kDa, pI=5.25] that shares 54.6 and 42.8% identity with *S. solfataricus*  $\alpha$  [MW=59.674 kDa, pI=5.29] and  $\beta$  [MW=60.366 kDa, pI=5.55], respectively. It is thus possible that the structural studies of Ellis *et al.* are of insufficient resolution to distinguish  $(\alpha_2\beta)_3$  from an  $(\alpha\beta\gamma)_3$  arrangement, and that previous descriptions of a 2:1  $\alpha$  to  $\beta$  subunit ratio in *Sulfolobus* (Knapp *et al.* 1994) in fact represent a 1:1:1 ratio of  $\alpha$ ,  $\beta$  and  $\gamma$  subunits.

While the expression of the  $\gamma$  subunit gene has yet to be confirmed, it appears fully functional. The upstream region of the gene in *S. solfataricus* contains promoter elements like those described for the  $\alpha$  and  $\beta$  subunits in *S. shibatae* (Kagawa *et al.* 1995) and two poly-T transcription terminators occur downstream of the stop codon (data not shown). More compelling evidence comes from a comparison of the  $\gamma$  gene sequences in *S. solfataricus* and *S. shibatae*. A significant bias towards synonymous (silent) substitutions was observed between the two genes ( $K_A/K_S \cong 0.1$ ), as would be expected if they are expressed and are evolving under selection. It is curious that the  $\gamma$  subunit appears to be quite divergent in sequence relative to the  $\alpha$  and  $\beta$  subunits, particularly in its apical domain (Figure 1.6C; see also Figure 1.3). This finding could have implications for the function of this subunit in *Sulfolobus* chaperonin complexes.

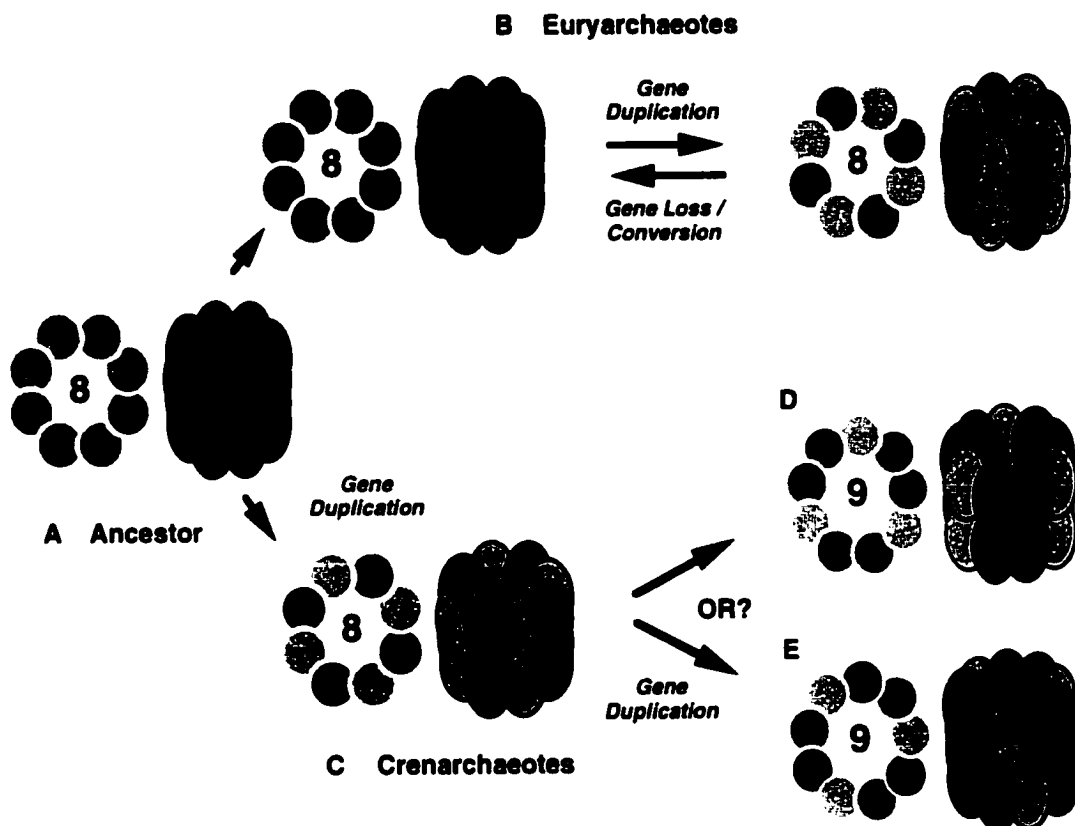
Interestingly, the euryarchaeote *Haloferax volcanii* also possesses a third chaperonin gene (Figures 1.3 and 1.14; Thompson and Daniels 1998). Unlike the current situation for *Sulfolobus*, the presence of three distinct protein subunits in this organism has been confirmed by 2-D gel electrophoresis (P. Lund, personal

communication), although electron-microscopic/crystallographic studies have not been performed on the chaperonin complexes themselves. If nine-membered chaperonin rings were found, this would indicate that a transition from eight to nine-membered rings has occurred at least twice independently during archaeal evolution. The elucidation of the structure of chaperonin complexes in the closely related halophile, *Halobacterium* sp. NRC-1, would also be particularly relevant, as the gene orthologous to the third subunit gene in *H. volcanii* appears to have been lost in this lineage (see above). Figure 1.15 provides a schematic overview of the salient features of archaeal chaperonin evolution from a structural perspective.

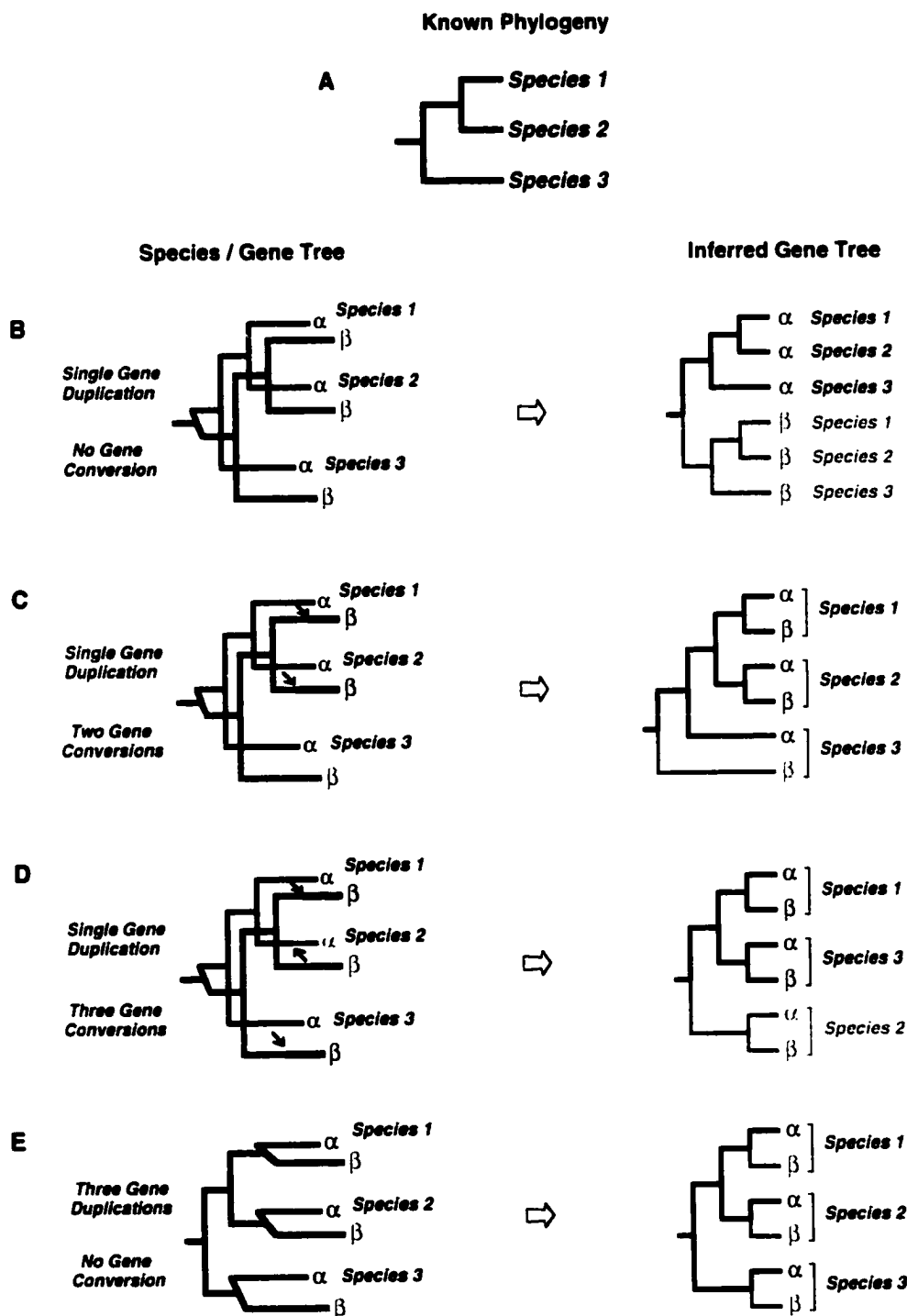
### **The problem with paralogy**

Paralogy is well known as a source of confusion in the interpretation of molecular phylogenies. Figure 1.16 serves to illustrate this point by way of a hypothetical three-species tree. The evolutionary relationship among the three groups is known (species 3 (species 1, species 2)), and various numbers of gene duplications and gene conversions have occurred. In Figure 1.16B, a scenario is depicted in which a single gene duplication occurred in the common ancestor of the three species. In Figure 1.16E, gene duplications have occurred in each of the three lineages after they diverged from one another. With respect to archaeal chaperonin phylogeny, scenario B most resembles the pattern observed in crenarchaeal genes (i.e., paralogy predating organismal divergence), while scenario E appears similar to the euryarchaeal pattern ('lineage-specific' paralogies).

The evolutionary scenarios shown in Figures 1.16C and D are identical to that in B (i.e., a single gene duplication in the common ancestor of the three species) except that various gene conversion events have occurred between



**Figure 1.15** Archaeal chaperonin evolution by recurrent paralogy. Schematic representation of chaperonin complexes: multimeric chaperonin rings are composed of individual subunits that interact asymmetrically (side-to-side and top-to-bottom). (A) Hypothetical chaperonin complex ancestral to euryarchaeotes, crenarchaeotes, and, probably, eukaryotes: eight-membered homo-oligomeric rings (see text). (B) Gene duplications have occurred independently in at least five different euryarchaeal lineages (distinct subunits encoded by duplicate genes are depicted as light and dark shaded subunits). At least two gene losses have occurred. (C) A gene duplication took place early in crenarchaeal evolution and a more recent duplication also took place in a *Sulfolobus* ancestor. (D,E) Two possible nine-membered chaperonin ring structures. (D) The  $(\alpha_2\beta)_3$  arrangement of Ellis *et al.* (1998), inferred from two-dimensional crystalization of chaperonin particles. (E) My prediction of alternating  $\alpha$ ,  $\beta$  and  $\gamma$  subunits in each *Sulfolobus* chaperonin ring.



**Figure 1.16** Hypothetical gene duplication and gene conversion scenarios and their effects on the inference of a 'known' phylogeny. (A) Known phylogenetic relationship of the three species (species 3 is an outgroup to species 1 and 2). (B-E) Three-dimensional species/gene trees with various combinations of gene duplication and complete gene conversion events and their inferred gene trees. In three-dimensional species/gene trees, arrows indicate the direction of gene conversion events.

paralogs *after* the three lineages have diverged. Significantly, the gene trees inferred in these two scenarios differ—in scenario C, the genes from species 1 and 2 are most closely related, while in scenario D, genes from species 1 and 3 appear as sister groups. Most interesting is that the gene trees inferred in scenarios D and E differ in their species affiliations but show the *same* lineage-specific gene duplication pattern, despite the fact that a single duplication occurred in D while three duplications occurred in E.

Of course, the chaperonin tree of euryarchaeotes differs from the hypothetical tree in that the phylogeny is not known—there is as yet no general consensus on the relationships among the major euryarchaeal clades. It is therefore impossible to know for sure which has been the predominant evolutionary force, gene duplication or gene conversion. Some support for the possibility that hidden paralogies and differential gene losses or conversions have in fact occurred in the evolutionary history of euryarchaeotes (analogous to scenario D, above) comes from a comparison of the topology of the euryarchaeal chaperonin tree to those constructed from other molecules. The chaperonin tree is, for instance, most unlike that inferred from small subunit ribosomal RNA (SSUrRNA) in placing *Methanopyrus kandleri* in a highly derived position, next to the halophiles (Figure 1.3; see for example Takai and Horikoshi 1999). As well, the Thermoplasmatales are moderately supported as the deepest branch in the chaperonin tree, a position not occupied by this group in the SSUrRNA tree (Takai and Horikoshi 1999; Takai and Sako 1999). These incongruencies could be the result of paralogy in the chaperonin dataset.

Strictly speaking then, the phylogenetic pattern observed in euryarchaeotes (Figure 1.3) is compatible with a pattern of somewhat fewer gene duplications combined with periodic complete gene conversion (as illustrated in Figure 1.16D). Complete gene conversions would not be detected by any of the



methods used here (e.g., GENECONV), and would be indistinguishable from the loss of one gene followed by duplication of the other. However, the highest degree of amino acid identity shared between subunits encoded in the same archaeal genome is only 80.6% (*Thermococcus* KS-1 a and b), suggesting that if complete gene conversions between duplicate genes have occurred, they have not happened recently.

### **Chaperonin evolution by 'recurrent paralogy'**

In any event, it is the persistence of paralogy in so many separate lineages that begs for an explanation. The fact that some chaperonin complexes are homo-oligomers and others hetero-oligomers suggests that, for archaea as a whole, the presence of two (or more) subunits is not a strict requirement for chaperonin function. Nevertheless, it is possible that, in each of these instances, paralogy is maintained because the hetero-oligomeric chaperonin has acquired novel functions that its homo-oligomeric ancestor lacked. The lineages examined here comprise non-thermophilic halophiles (e.g., *Haloferax*) and both autotrophic (e.g., *Methanococcus*) and heterotrophic (e.g., *Archaeoglobus*) thermophiles, and homo- and hetero-oligomeric chaperonin complexes appear randomly distributed among them with respect to environment and/or lifestyle (Figures 1.3 and 1.14). From this perspective at least, there is little reason to suspect that homo- and hetero-oligomeric chaperonins function differently.

Co-evolved interdependence between the subunits of a hetero-oligomeric chaperonin complex seems a more appealing possibility. Ancestrally, chaperonin complexes would be constructed from homo-oligomeric rings, the subunits of which are products of a single gene. Gene duplication into 'a' and 'b' paralogs would be followed by sequence divergence, through the fixation of mutations in both paralogs that are neutral or only slightly deleterious in their consequences

for chaperonin assembly or function. Duplicate chaperonin genes would thus encode functionally identical subunits that assemble into chaperonin rings in random proportions determined by their cellular abundance. At this stage, one or the other duplicate could be lost as inconsequentially as it was gained, and partial or complete gene conversion events between recent duplicates might periodically reset the 'divergence clock'. With time, mutations in the regions of intra-ring contact of one subunit followed by (or coincident with) compensatory changes in a newly-evolved duplicate subunit could produce a tendency toward the assembly of an ordered arrangement of subunits within a chaperonin ring (like that observed in the thermosome). Such a process would make the formation of homo-oligomers from either duplicate subunit increasingly unlikely and make the loss of either gene increasingly disadvantageous. Obligatory hetero-oligomerism could thus evolve in the absence of 'specialized' roles for the duplicate subunits in the protein folding process. The completely hetero-oligomeric CCT complex found in the cytosol of eukaryotes may in fact be an example of such a process gone to completion. Individual CCT subunits share approximately 30% identity, and seem to occupy specific positions relative to the others, in each 8-membered ring (Lin *et al.* 1997; Lin and Sherman 1997; Liou and Willison 1997; Willison and Kubota 1994). The origin and evolution of CCT is discussed more fully in Chapter II.

This 'neutral' explanation for the persistence of hetero-oligomerism is consistent with the observation that eight-membered rings of euryarchaeal chaperonins are in some species made from single protein subunits (e.g., *Methanopyrus kandleri*) and in other species from two (e.g., *Thermoplasma acidophilum*; Figure 1.14). As well, heterologous expression studies suggest that proper formation and function of homo-oligomeric chaperonins *in vivo* is likely a function of the degree of sequence divergence between paralogous subunits.

*Escherichia coli*-expressed  $\alpha$ -only and  $\beta$ -only ('homo-oligomeric') chaperonins from *Thermococcus* strain KS-1, whose  $\alpha$  and  $\beta$  subunits are 80.6 % identical, showed ATPase activity and protein-folding ability (Yoshida *et al.* 1997). In contrast, heterologously expressed homo-oligomeric chaperonin complexes from *Sulfolobus sp.* S7 ( $\alpha$  and  $\beta$  share only 55.5 % identity) were unstable, prone to dissociation into monomers and showed no ATPase activity (Yoshida *et al.* 1998). In *P. occultum*, recombinant  $\alpha$ -only and  $\beta$ -only (61.8% identical) thermosomes were microscopically indistinguishable from their native counterparts, yet exhibited reduced thermal stability and were deemed only partly functional (Minuth *et al.* 1998).

As alluded to previously, an important corollary of the fact that hetero-oligomerism evolved multiple times independently in archaea is that subunit-specific roles in protein folding, if they exist, must also have evolved multiple times. Unfortunately, little data is currently available with which to test this hypothesis: next to nothing is known about the substrates of archaeal chaperonins, and, apart from *Thermoplasma acidophilum*, very little biochemical and structural information is available on chaperonin complexes from diverse archaeal lineages (Gutsche, Essen and Baumeister 1999). The observation that in crenarchaeotes, gene conversion appears to have repeatedly homogenized large portions of the substrate-binding domains of duplicate subunits seems to argue against the idea that positive selection for subunit 'specialization' has been the evolutionary force leading to obligatory hetero-oligomerism. It is interesting that duplicate pairs of subunits appear to be most differentiated in their equatorial and intermediate domains—the areas involved in most of the inter-subunit contact points (Ditzel *et al.* 1998). This is more consistent with a model for the origin of hetero-oligomerism that emphasizes co-evolution between duplicate subunits, as opposed to subunit and substrate.

It is possible that gene conversions have occurred multiple times independently in the apical domain region of the gene simply because this region of the protein is most highly conserved between duplicates, and is therefore the most similar region at the DNA level. However, the % identity spectra obtained for duplicate pairs of sequences, particularly those in euryarchaeotes (Figures 1.6 and 1.7), do not support this notion. No particular part of the gene stands out as being consistently more highly conserved. It may be that repeated gene conversions (homogenizations) in the apical domain have instead been driven by selection.

The data presented here suggest that if subunit-specific roles in substrate binding do exist for duplicate subunits in archaea, the only place in which such specificity is likely to reside is within the helical protrusion. Unlike the core region of the apical domain, duplicate subunits contain numerous substitutions in this area (Figure 1.13). A possible role for the protrusion in substrate recognition has in fact been suggested based on the presence of several clusters of surface-exposed hydrophobic residues in this area (Klumpp and Baumeister 1998; Klumpp, Baumeister and Essen 1997). By analogy to the well-studied GroEL system in bacteria, hydrophobic-hydrophobic interactions could provide the basis by which non-native protein substrates are recognized by archaeal chaperonins.

While the evolutionary processes that gave rise to the completely hetero-oligomeric chaperonin complex in eukaryotes are obscure, a better understanding of the biochemistry and structural biology of the moderately hetero-oligomeric chaperonins found in diverse archaeal lineages would undoubtedly help. There are many issues to be addressed. Are duplicate subunits always present in the same ring (as in *T. acidophilum* and the eukaryotic chaperonin CCT), or have multiple distinct homo-oligomeric chaperonin

complexes evolved in some lineages? What are the *in vivo* substrates of archaeal chaperonins? Do they interact with a wide range of substrates as does the bacterial chaperonin GroEL, or are they more selective, as appears to be the case for CCT? Answers to these questions should bring us closer to a more complete understanding of the origin and evolution of the CCT complex in the eukaryotic cytosol and its role in protein folding.

### ADDENDUM

The expression of the *Sulfolobus shibatae*  $\gamma$  subunit gene characterized in this chapter has recently been confirmed (Kagawa, Yaoi, and Trent, in preparation). The exact role of the  $\gamma$  subunit in *Sulfolobus* chaperonin complexes is under investigation.

## CHAPTER II

### The Origin and Evolution of the Eukaryotic Chaperonin CCT

This chapter includes work published in Archibald, J. M., J. M. Logsdon, Jr., and Doolittle, W. F. 2000. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol. Biol. Evol.* 17(10): 1456-1466, and Archibald, J. M., C. Blouin, and W. F. Doolittle. 2001. Gene duplication and the evolution of group II chaperonins: implications for structure and function. *J. Struct. Biol.* (in press).

New sequences have been deposited in Genbank under accession numbers AF226714-AF226726.

## INTRODUCTION

Chaperonin-mediated protein folding is a universal cellular process. Eukaryotic cells possess two distantly related (but clearly homologous) chaperonin classes with different evolutionary histories. Bacterial-type (group I) chaperonins, called cpn60 or hsp60, reside in eukaryotic organelles, while archaeal-type chaperonins (group II; called CCT or TRiC) are present in the eukaryotic cytosol (Frydman *et al.* 1992; Kubota, Hynes and Willison 1995a; Trent *et al.* 1991; Willison and Kubota 1994).

Crystal structure comparisons of group I and group II chaperonins reveal remarkable structural conservation (Ditzel *et al.* 1998). However, there are significant differences between the two chaperonin types. While group I chaperonins utilize the co-chaperonin GroES/cpn10 in the protein folding process, no such homolog functions in the group II chaperonin complex. Instead,

an extended 'apical domain', present in the group II chaperonins but absent in the group I's, is thought to cap the central cavity in a manner analogous to GroES/cpn10 (Horwich and Saibil 1998; Klumpp, Baumeister and Essen 1997; Llorca *et al.* 1999b). Recent experiments suggest that novel co-chaperonins, unrelated to GroES/cpn10, interact with CCT to assist protein folding (Gebauer, Melki and Gehring 1998; Geissler, Siegers and Schiebel 1998; Siegers *et al.* 1999; Vainberg *et al.* 1998). Group I and group II chaperonins also differ in the number of subunits present in each chaperonin ring. *Escherichia coli* GroEL, the archetypal bacterial chaperonin, has a double-ring structure with seven subunits per ring (Braig *et al.* 1994), while archaeal and eukaryotic cytosolic chaperonin complexes are composed of eight- or nine-membered rings (reviewed in Gutsche, Essen and Baumeister 1999; Klumpp and Baumeister 1998; Willison and Horwich 1996).

Group I and group II chaperonins may also differ in their degree of substrate specificity. While the *E. coli* chaperonin GroEL is known to mediate the folding of a wide range of proteins (Houry *et al.* 1999), the diversity of substrates that interact with the eukaryotic chaperonin CCT seems more limited. The abundant cytoskeletal proteins actin and tubulin seem to be the predominant substrates of CCT (Willison and Horwich 1996). However, other substrates do continue to be identified (e.g., Farr *et al.* 1997; Feldman *et al.* 1999; Melki *et al.* 1997; Srikakulam and Winkelmann 1999; Won *et al.* 1998), and the full range of *in vivo* CCT activity is in fact unknown (Leroux and Hartl 2000; Thulasiraman, Yang and Frydman 1999; Willison 1999).

The most unusual feature of both archaeal and eukaryotic chaperonins is their hetero-oligomeric nature. Unlike the homo-oligomer GroEL, archaeal chaperonins are often composed of several different (but homologous) subunits. In the previous chapter, I concluded that in the chaperonin complexes of archaea, hetero-oligomerism likely evolved multiple times independently. In archaeal

genomes, duplicate chaperonin genes (paralogs) are often more similar to each other than to those in other archaea, suggesting recent (lineage-specific) duplication. Compared to archaeal chaperonins, the eukaryotic CCT is even more hetero-oligomeric. This was first suggested on biochemical grounds (Frydman *et al.* 1992; Lewis *et al.* 1992; Rommelaere *et al.* 1993), and subsequent sequence comparisons of CCT genes in mouse confirmed the existence of eight distinct subunit species ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\eta$ ,  $\theta$ , and  $\zeta$ ), each thought to occupy a unique position in the eight-membered CCT rings (Kubota *et al.* 1994; Kubota, Hynes and Willison 1995a; Liou and Willison 1997). Orthologs of each of the mouse CCT subunit genes have also been described in yeast (Kim, Willison and Horwich 1994; Kubota *et al.* 1994; Stoldt *et al.* 1996).

Is the presence of eight CCT subunits a universal feature of eukaryotic cells? An early eukaryotic lineage that diverged from other eukaryotes prior to multiple CCT gene duplications might be expected to possess a smaller and/or different complement of CCT genes. The taxonomic diversity needed to address this question is, however, currently lacking. To this end, I sought to (1) increase the phylogenetic diversity of known CCT genes, (2) perform phylogenetic analyses of archaeal and eukaryotic chaperonins to determine when during eukaryotic evolution (and in what order) the gene duplications that gave rise to the CCT subunits occurred, and (3) address specific hypotheses regarding the origin and evolution of CCT from an archaeal-like homo- or moderately hetero-oligomeric chaperonin complex ancestor. I focussed on isolating CCT genes from two amitochondriate protists, *Trichomonas vaginalis* (a parabasalid) and *Giardia lamblia* (a diplomonad). The parabasalids and diplomonads consistently branch at or near the base of the eukaryotic tree in molecular phylogenies (Roger 1999), and have been suggested to be early diverging eukaryotic lineages.



Previous comparative sequence analyses (Kubota *et al.* 1994) have indicated that a completely hetero-oligomeric CCT was present in the common ancestor of animals and fungi. The results presented in this chapter push back the origin of the CCT gene duplications to the common ancestor of animals, fungi, plants, parabasalids and diplomonads, and likely to the common ancestor of all extant eukaryotes. While the exact position of the archaeal root to the eukaryotic CCTs is ambiguous, no close phylogenetic relationship between the archaeal chaperonins and specific eukaryotic CCT subunits was observed, suggesting that the eukaryotic CCT complex became hetero-oligomeric independent of the archaeal chaperonins. The gene duplications producing the CCT $\delta$  and  $\epsilon$  subunits, as well as those in the CCT $\alpha$ / $\beta$ / $\eta$  clade, appear to represent the most recent duplications of the CCT gene family.

I also present data describing the distribution of slowly evolving subunit-specific 'signatures' and variable regions among the different CCT subunits. These signatures are found in the equatorial, intermediate and apical domains of the protein. Within the apical domain, a distinct cluster of subunit-specific signatures is located in the putative substrate-binding region of several of the CCTs: these signatures likely mediate the subunit-specific functions described for the binding of actin (Llorca *et al.* 1999a), tubulin (Llorca *et al.* 2000) and perhaps other CCT substrates.

## RESULTS

### **Cloning and sequencing of *Trichomonas* and *Giardia* CCT genes**

A battery of degenerate PCR primers was designed to the universally conserved regions of group II chaperonins (see Materials and Methods) and used to amplify CCT subunit-encoding genes from *Trichomonas vaginalis* and *Giardia*

*lamblia* genomic DNAs. The *Trichomonas* CCT genes were obtained using the following primer combinations: *Trichomonas Ccta* (CCT $\alpha$ ): CCT-1-for/CCT-4-rev, CCT-5-for/CCT-4-rev, CCT-9-for/CCT-7-rev; *Trichomonas Cctd* (CCT $\delta$ ): p80-4B (5'-CTGCCATTYGTGGCNATG-3')/P80-5 (5-AGCGATGAACTTNARDAT-3'); *Trichomonas Cctg* (CCT $\gamma$ ): CCT-1-for/CCT-4-rev; *Trichomonas Cctz* (CCT $\zeta$ ): CCT-5-for/CCT-3-rev, CCT-5-for/CCT-4-rev, CCT-5-for/CCT-7-rev.

In an attempt to obtain full-length sequences of these genes, PCR-generated CCT gene fragments were isolated (BIORAD, Prep-a-gene), labeled with  $\alpha^{32}\text{P}$  (Prime-It II random primer labeling kit, Stratagene), and used as probes to screen a *Trichomonas vaginalis* genomic DNA library (constructed previously by N. Fast and J. Logsdon, Dalhousie University). Genomic clones containing two different *Ccta* genes and two different *Cctd* genes were obtained. While both *Cctd* genes code for identical proteins (but have many synonymous substitutions), the two *Ccta* genes encode slightly different proteins. The full-length CCT $\alpha$ -1 has an unusually short and divergent 3' end, while the 5'-truncated CCT $\alpha$ -2 clone encodes a protein with a carboxyl terminus typical of other CCTs. Southern hybridization of a *Ccta* PCR product to *Trichomonas* genomic DNA produced multiple hybridizing bands, confirming the presence of at least two genomic copies of this gene (data not shown). Genomic library clones containing a full-length *Cctz* gene, as well as a 5'-truncated clone containing *Cctg*, were also isolated. A *Trichomonas* cDNA library clone encoding a protein with significant similarity to CCT $\eta$  was a gift from R. Hirt and M. Embley (Natural History Museum, London).

For *Giardia*, a portion of the *Cctd* (CCT $\delta$ ) gene was obtained with degenerate PCR primers (as above) with the CCT-1-for/CCT-4-rev primer pair. A recent sequence survey of the *Giardia lamblia* genome (Smith *et al.* 1998) revealed the presence of coding regions with similarity to several additional CCT

subunits. Exact-match PCR primers were designed based on preliminary genome sequence data from *Giardia* and, in combination with various degenerate primers, were used to amplify additional CCT genes (*Giardia Ccta* (CCT $\alpha$ ): GL.alpha.for.1/GL.alpha.rev.1; *Giardia Cctb* (CCT $\beta$ ): GL.beta.for.1/GL.beta.rev.2, CCT-1-for/GL.beta.rev.3; *Giardia Cctg* (CCT $\gamma$ ): CCT-5-for/GL.gamma.rev.1; *Giardia Ccte* (CCT $\epsilon$ ): GL.eps.for.1/GL.eps.rev.1; *Giardia Cctq* (CCT $\theta$ ): GL.theta.for.1/GL.theta.rev.1; *Giardia Cctz* (CCT $\zeta$ ): GL.zeta.for.1/Gl.zeta.rev.1). No spliceosomal introns were found in any of the *Giardia* or *Trichomonas* genes presented here, consistent with their complete absence from the protein-coding genes described in these organisms thus far.

To further increase the taxonomic sampling of CCT sequences for comparative study and phylogenetic analysis, I searched the public sequence databases by BLAST (Altschul *et al.* 1997) using the *Trichomonas* and *Giardia* CCT sequences as queries. Complete sets of CCT protein sequences (eight or nine) were obtained for human, mouse (Kubota *et al.* 1994; Kubota, Hynes and Willison 1995b), yeast and *Caenorhabditis elegans*, as well as single or multiple CCT sequences for *Plasmodium falciparum*, *Leishmania major*, and a variety of animals, plants, fungi and ciliates. Several of the *C. elegans* sequences (CCT $\gamma$ ,  $\eta$  and  $\theta$ , obtained from the Sanger center (<http://www.sanger.ac.uk/>)) contained unique insertions and/or deletions that were most likely the result of incorrect intron/exon boundary predictions; in each case, I identified alternate splice sites that removed the apparent insertions or added the missing exons. As of 03/2001, the complete group II chaperonin dataset contained 40 archaeal sequences and 124 eukaryotic CCTs.

An alignment of the inferred *Trichomonas* and *Giardia* protein sequences with representative archaeal chaperonins and diverse CCTs is shown in Figure 2.1. The new sequences possess putative ATP-binding/ATP-hydrolysis sequence

**Figure 2.1** Alignment of archaeal chaperonins and eukaryotic CCT protein sequences. The alignment contains 11 representative archaeal sequences and between 6 and 10 sequences from each of the 8 CCT subunit families (75 sequences in total, 716 aligned amino acid positions). The *Trichomonas* and *Giardia* genes sequenced in this study are indicated by asterisks. Amino acid residues present in at least 50% of the sequences are shaded black and chemically similar amino acids (if present in  $\geq 50\%$  of the sequences) are shaded gray. Taxon abbreviations: *Taci\_a*, *Thermoplasma acidophilum*  $\alpha$  subunit; *Mthe\_a*, *Methanobacterium thermoautotrophicum*  $\alpha$  subunit; *Mkandl*, *Methanopyrus kandleri*; *Mjanna*, *Methanococcus jannaschii*; *Hvol\_1*, *Haloferax volcanii* subunit 1; *Aful\_a*, *Archaeoglobus fulgidus*  $\alpha$  subunit; *Phorik*, *Pyrococcus horikoshii*; *Ssol\_a*, *S. solfataricus*  $\alpha$  subunit; *Aper\_a*, *Aeropyrum pernix*  $\alpha$  subunit; *Mmus*, *Mus musculus*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Ddis*, *Dictyostelium discoideum*; *Tpyr*, *Tetrahymena pyriformis*; *Pfal*, *Plasmodium falciparum*; *Tvag*, *Trichomonas vaginalis*; *Glam*, *Giardia lamblia*; *Atri*, *Aedes triseriatus*; *Gmax*, *Glycine max*; *Osat*, *Oryza sativa*; *Asat*, *Avena sativa*; *Csat*, *Cucumis sativus*; *Onov*, *Oxytricha nova*; *Ehis*, *Entamoeba histolytica*. CCT subunit abbreviations: *alp*, CCT $\alpha$  (alpha); *bet*, CCT $\beta$  (beta); *del*, CCT $\delta$  (delta); *eps*, CCT $\epsilon$  (epsilon); *eta*, CCT $\eta$  (eta); *gam*, CCT $\gamma$  (gamma); *the*, CCT $\theta$  (theta); *zet*, CCT $\zeta$  (zeta).

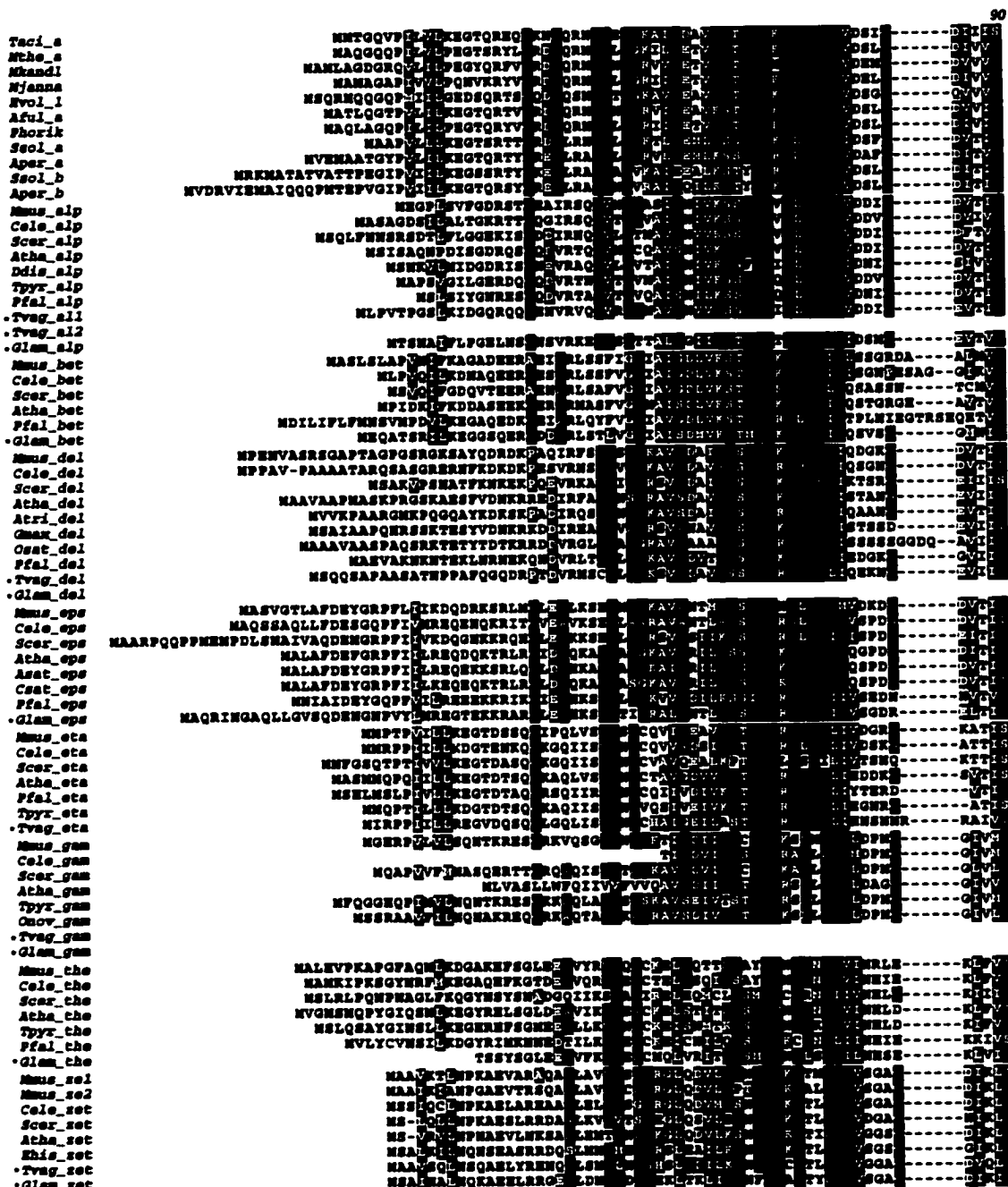


Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences

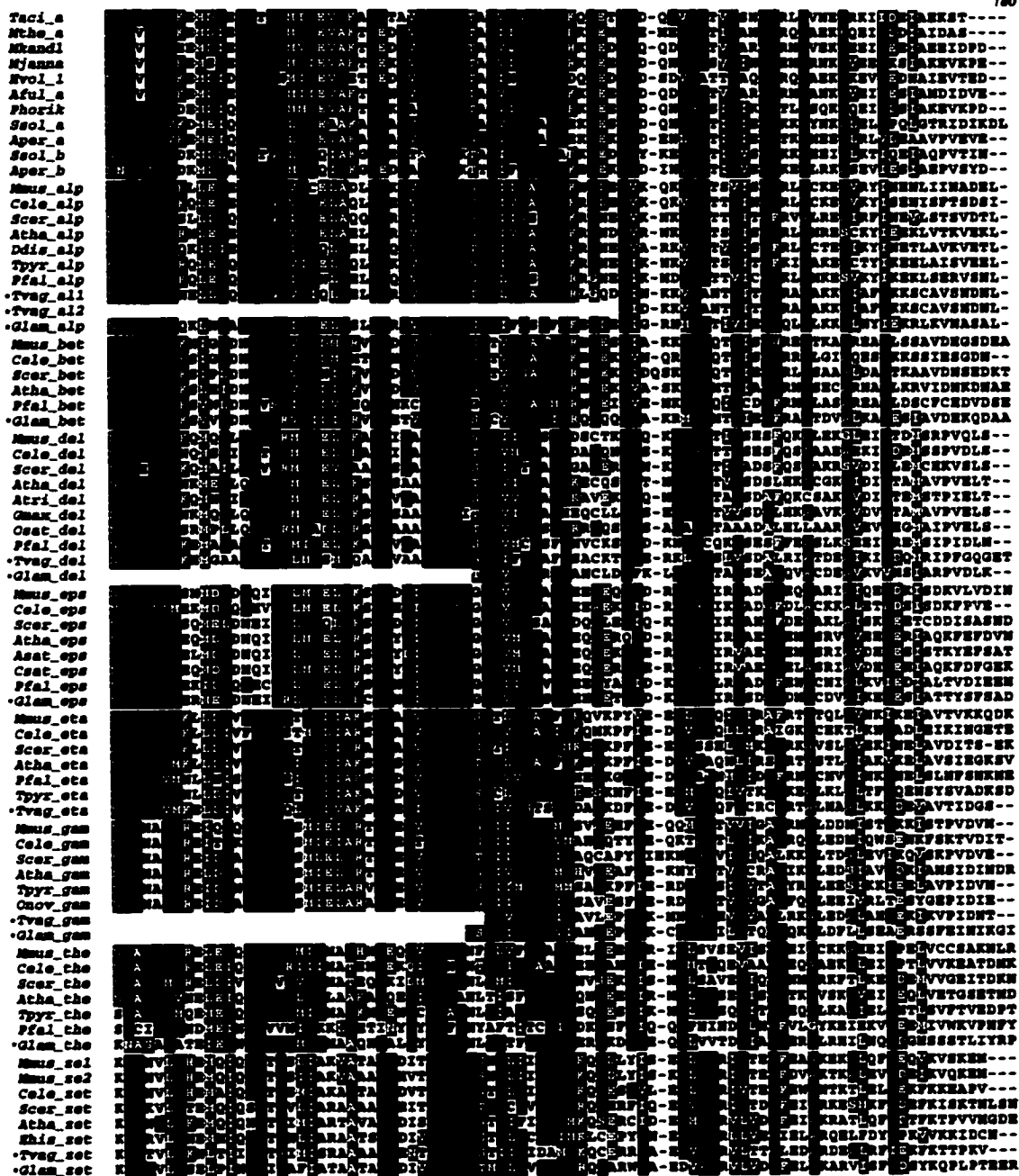


Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences

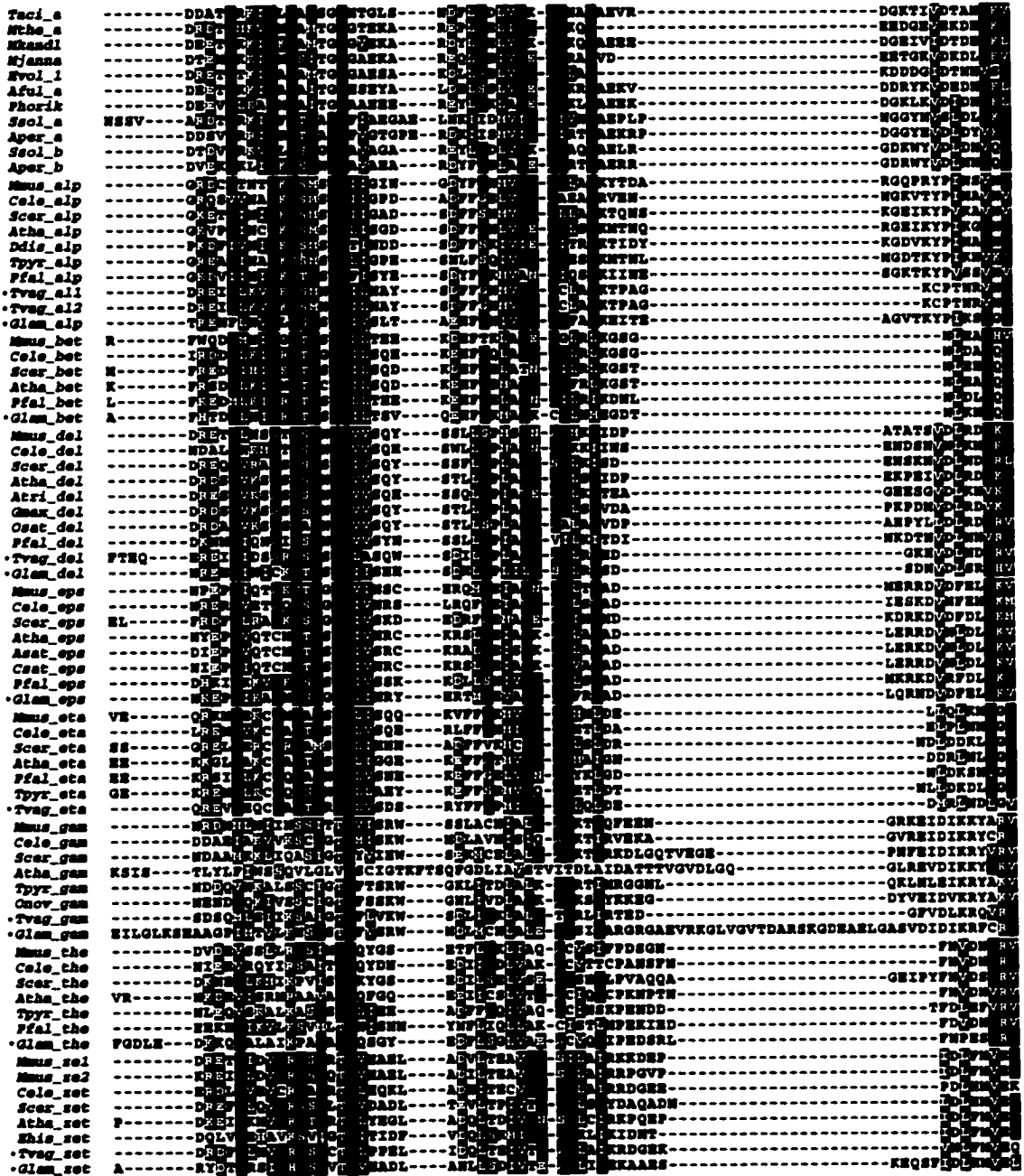


Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences



Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences







Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences



Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences

```

Taci_a  -KQV---KAKV---LRS---TRHNS---V---R---K---AKSKKTFPFGQGGGQQGQGFPGGQNFPEY
Mthe_a  -KIV---KAKV---LRS---TRHNS---V---R---K---AAASSGSSSEKGNHNGGNGGQNFPM
Mstandl -KIV---KAKV---LRS---TRHNS---V---R---K---AARELSEKSEKSEKSEKGGSEF
Mjanna  -KIV---KAKV---LRS---TRHNS---V---R---K---AAEKVKQDEKKGKGGGQGGQDEF
Evol_1  -KIV---KAKV---LRS---TRHNS---V---R---K---AAGDLGQQGTGSDDDDDGGQAFQGMGGGNGGNGGNGGNGGAM
Aful_a  -KIV---KAKV---LRS---TRHNS---V---R---K---AAKGLSEKSEKSEKSEKGGSEDFE
Fhorik  -KIV---KAKV---LRS---TRHNS---V---R---K---AAKLEKSEKSEKSEKGGSEDFEFGSDLD
Scol_a  -KIV---KAKV---LRS---TRHNS---V---R---K---AAAPLSEKSEKSEKSEKGGSEDFEFGSDLD
Aper_a  -KIV---KAKV---LRS---TRHNS---V---R---K---AAAPFKKSEKSEKSEKGGSEDFEFGSDLD
Scol_b  -KIV---KAKV---LRS---TRHNS---V---R---K---AAAGKSEKSEKSEKSEKGGSEDFEFGSDLD
Aper_b  -KIV---KAKV---LRS---TRHNS---V---R---K---AAKSEKSEKSEKSEKGGSEDFEFGSDLD
Mmus_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLEPCKDDKKGSTHNAVESGALDD
Cele_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLDKQEPFGQDDCEA
Scor_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVKDESGQGE
Ddis_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tpyr_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_al1 -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_al2 -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_alp -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_bet -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atri_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Gmax_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Osat_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_del -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Asat_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Csat_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_eps -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tpyr_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Ocov_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_ets -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tpyr_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Ocov_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_gam -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tpyr_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Pfal_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_the -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Mmus_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Cele_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Scor_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Atha_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Ehis_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Tvag_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE
Glam_set -KIV---KAKV---LRS---TRHNS---V---R---K---KLVDFPFPKEDFDE

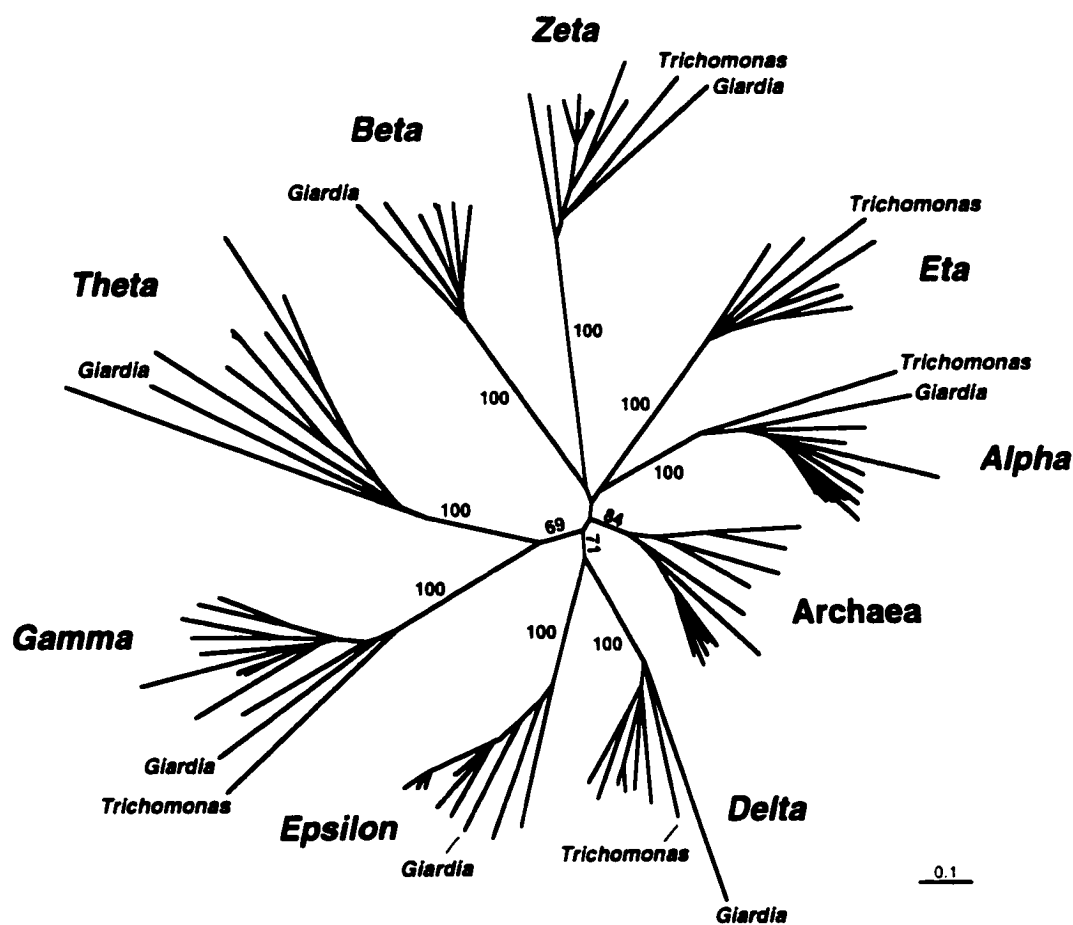
```

Figure 2.1 Alignment of archaeal chaperonins and eukaryotic CCT protein sequences

motifs similar to those described for other chaperonins (Kubota *et al.* 1994), and share significant amino acid identity with all group II chaperonin homologs. The most striking feature of the alignment is the presence of multiple insertions in the *Giardia* CCT sequences that are not found in any CCTs characterized thus far. These insertions generally map to regions of variable length; however, the *Giardia* CCT $\theta$  and  $\epsilon$  sequences possess unique insertions (16 and 9 amino acids, respectively) in a highly conserved region corresponding to a domain present in the bacterial/organellar chaperonins (positions 339-374 of the *E. coli* GroEL sequence; Ditzel *et al.* 1998) but absent from eukaryotic CCTs and archaeal chaperonins. The significance of these insertions (which presumably occurred independently) in terms of chaperonin subunit structure/function is not known.

### Chaperonin phylogeny

The evolutionary relationship between the *Trichomonas* and *Giardia* sequences and other eukaryotic and archaeal chaperonins was examined by constructing phylogenetic trees. Figure 2.2 shows an unrooted neighbor-joining tree produced from an alignment of 11 representative archaeal sequences, 85 eukaryotic CCTs and 260 amino acid positions. Most notably, the *Trichomonas* and *Giardia* sequences form robust clades with each of the eight different CCT paralogs (100% support with all phylogenetic methods; data not shown). This indicates that (1) the gene duplications producing the paralogs predate the divergence of *Trichomonas* and *Giardia* from other eukaryotes, and (2) multiple CCT paralogs have been retained over a large time scale of eukaryotic evolution. It is likely that both *Trichomonas* and *Giardia* possess all eight CCT paralogs. Indeed, a portion of the one CCT subunit gene not isolated from *Giardia*, *CctH*, has recently been sequenced by the *Giardia* genome sequencing project (Smith *et al.* 1998).



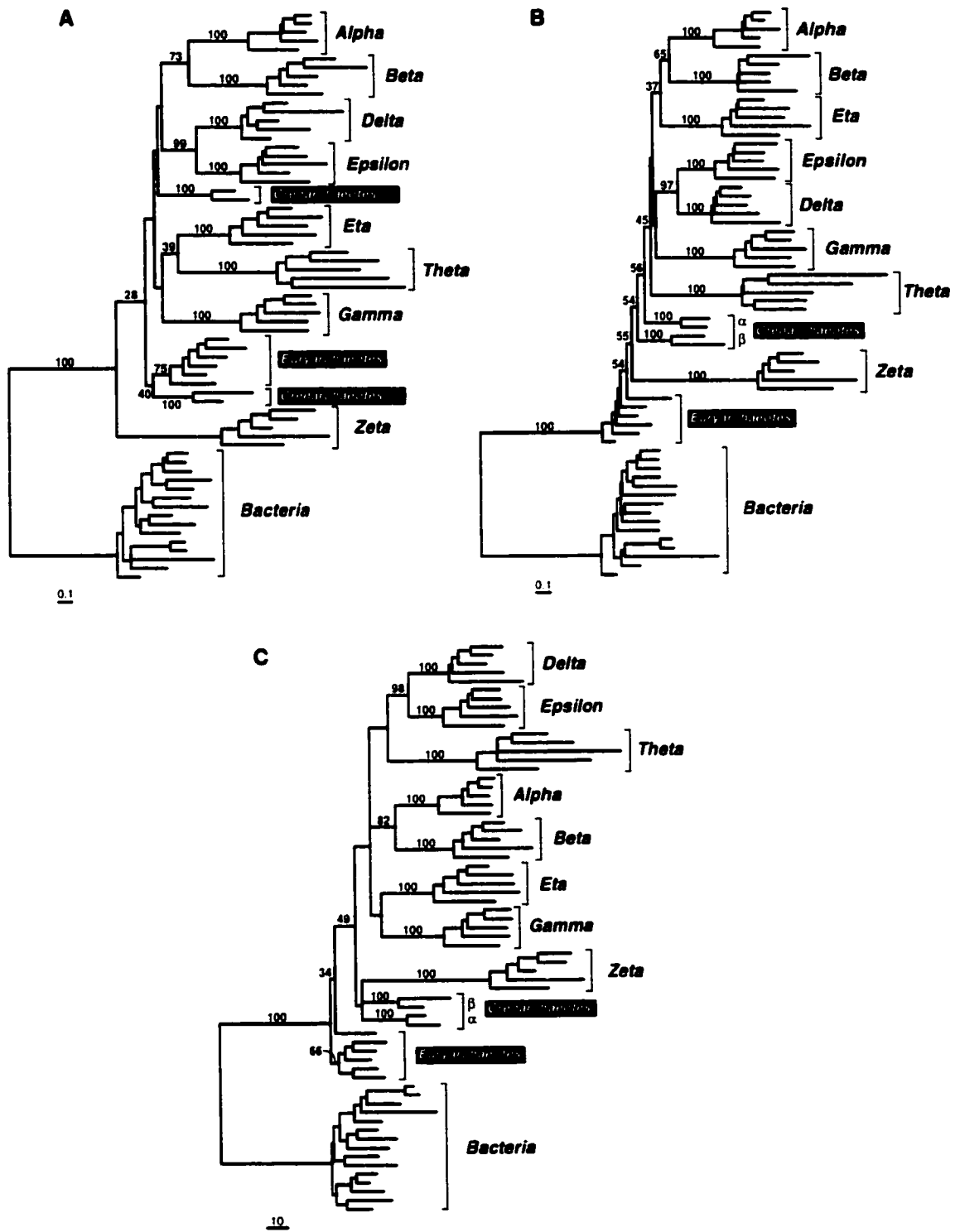
**Figure 2.2** Placement of *Trichomonas* and *Giardia* chaperonins into known CCT subunit families. The tree shown is an unrooted neighbor-joining distance tree of 11 archaeal chaperonins and 85 eukaryotic CCT sequences, inferred from an alignment of 260 amino acid positions. The *Trichomonas* and *Giardia* sequences are labeled. In each CCT subunit clade the *Trichomonas* and/or *Giardia* sequence branches strongly with subunit members from other species. This was also observed in additional phylogenetic analyses. Bootstrap support (100 replicates) for the deepest branches are indicated if over 50%. The scale bar indicates estimated number of amino acid substitutions per site.

To more rigorously address the question of the evolutionary relationship of the CCT paralogs with the archaeal chaperonins, and to determine the position of the bacterial (i.e., group I) root of the group II chaperonin tree, I performed phylogenetic analyses using alignments that contained reduced numbers of taxa and maximal phylogenetic diversity. Surprisingly, when the bacterial chaperonin sequences were included as an outgroup (65 taxa, 227 position alignment), maximum likelihood (ML) analyses produced trees in which the eukaryotic CCT $\zeta$  clade (not archaea) was the deepest branch of the group II chaperonins (Figure 2.3A). ML-distance trees placed the euryarchaeotes as the deepest branch, but as a paraphyletic group separated from the crenarchaeotes by the CCT $\zeta$  clade of eukaryotes (Figure 2.3B). A similar result was obtained in ML analyses using an alignment from which the fastest-evolving sites had been removed (24 sites, 203 total sites; Figure 2.3C). The deepest branches in these phylogenies were not well supported, however, suggesting that CCT $\zeta$  (the longest branch of the CCTs; see Figure 2.2) might be attracted to the long branch of the bacterial outgroup. Clearly, the small number of alignable amino acid positions between the group I and II chaperonins (approximately 200 sites, corresponding primarily to the ATP-binding/hydrolysis motifs) provide little phylogenetic signal with which to address the evolutionary history of the archaeal/eukaryotic chaperonin tree. I therefore focused on the group II chaperonin dataset and attempted to determine the placement of the archaeal chaperonin root to the eukaryotic CCT tree, and the branching order of the various CCT paralogs.

Figure 2.4 shows an ML tree of representative archaeal chaperonins and eukaryotic CCTs (50 taxa, 355 sites). As in Figures 2.2 and 2.3, strong support for the monophyly of all the individual CCT subunit clades is recovered. As well, the CCT $\delta$  and CCT $\epsilon$  paralogs form a well-supported clade with ML, distance,

**Figure 2.3** Phylogenetic analysis of group I and group II chaperonin protein sequences. An alignment of 65 taxa was used, and contained 15 representative group I (bacterial) sequences, 10 archaeal sequences (6 euryarchaeotes, 4 crenarchaeotes) and 5 representative sequences from each of the 8 eukaryotic CCT subunit families. The alignment contained 227 unambiguously aligned amino acid positions, corresponding primarily to the universally conserved ATP-binding/hydrolysis motifs. (A) The maximum likelihood (ML) tree ( $\ln L = -19,783.4$ ) inferred from a heuristic search of 1,000 trees in protML (Adachi and Hasegawa, 1996) using the 65 taxa, 227 site dataset. ML RELL bootstrap values are given above the branches if  $>25\%$ . (B) Fitch-Margoliash distance tree inferred from a distance matrix calculated with a rate heterogeneity model (the JTT +  $\Gamma$  + inv model of amino acid substitution; see Materials and Methods). The 65 taxa, 227 site dataset was also used. ML-distance bootstrap values (500 replicates) are given above the branches. (C) ML tree ( $\ln L = -15,610.14$ ) inferred from a heuristic search of 1,000 trees in protML using a dataset in which the fastest-evolving sites had been removed (65 taxa, 203 total amino acid positions). ML RELL values are provided. The archaeal sequences (euryarchaeotes and crenarchaeotes) are highlighted, as are the 8 different CCT subunits. Scale bars represent the estimated number of amino acid substitutions per site.





**Figure 2.3** Phylogenetic analysis of group I and group II chaperonin protein sequences

**Figure 2.4** Phylogeny of group II chaperonin protein sequences. The tree shown is the maximum likelihood (ML) tree ( $\ln L = -24,644.31$ ) inferred from a heuristic search of 1,000 trees in protML (Adachi and Hasegawa 1996) using an alignment containing 355 unambiguously aligned amino acid positions. The alignment contained 10 archaeal sequences (6 euryarchaeotes, 4 crenarchaeotes) and 5 sequences from each of the 8 eukaryotic CCT paralogs (partial sequences determined in this study [CCT $\gamma$  from *Trichomonas* and *Giardia*, *Giardia* CCT $\delta$ ] were excluded in order to maximize the number of sites). The eukaryotic CCT subunits (paralogs) are highlighted. Bootstrap support for the major branches, as well as for the deepest branch in each CCT clade are indicated where >45%. ML RELL values are given above the branch, ML (with rate heterogeneity model) distance bootstrap values (500 replicates) are below. Gray inset boxes indicate support values for nodes of particular interest (ML, ML RELL values; MD, ML distance bootstrap values; QP, quartet puzzling support values (10,000 quartet puzzling steps); FM, distance (Fitch-Margoliash) bootstrap values). Dashes (--) indicate support values <45%. Scale bar indicates the estimated number of substitutions per amino acid site.

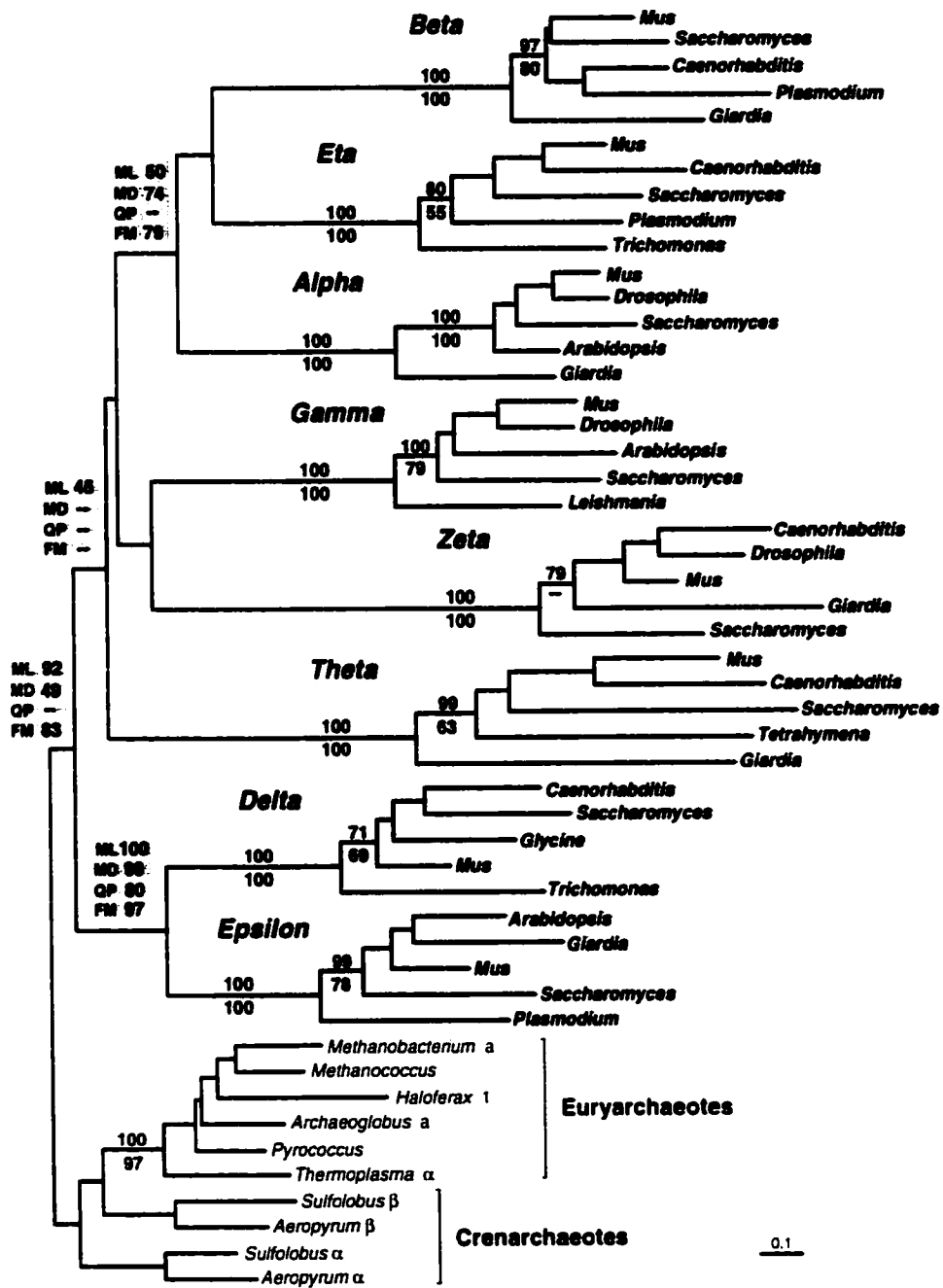
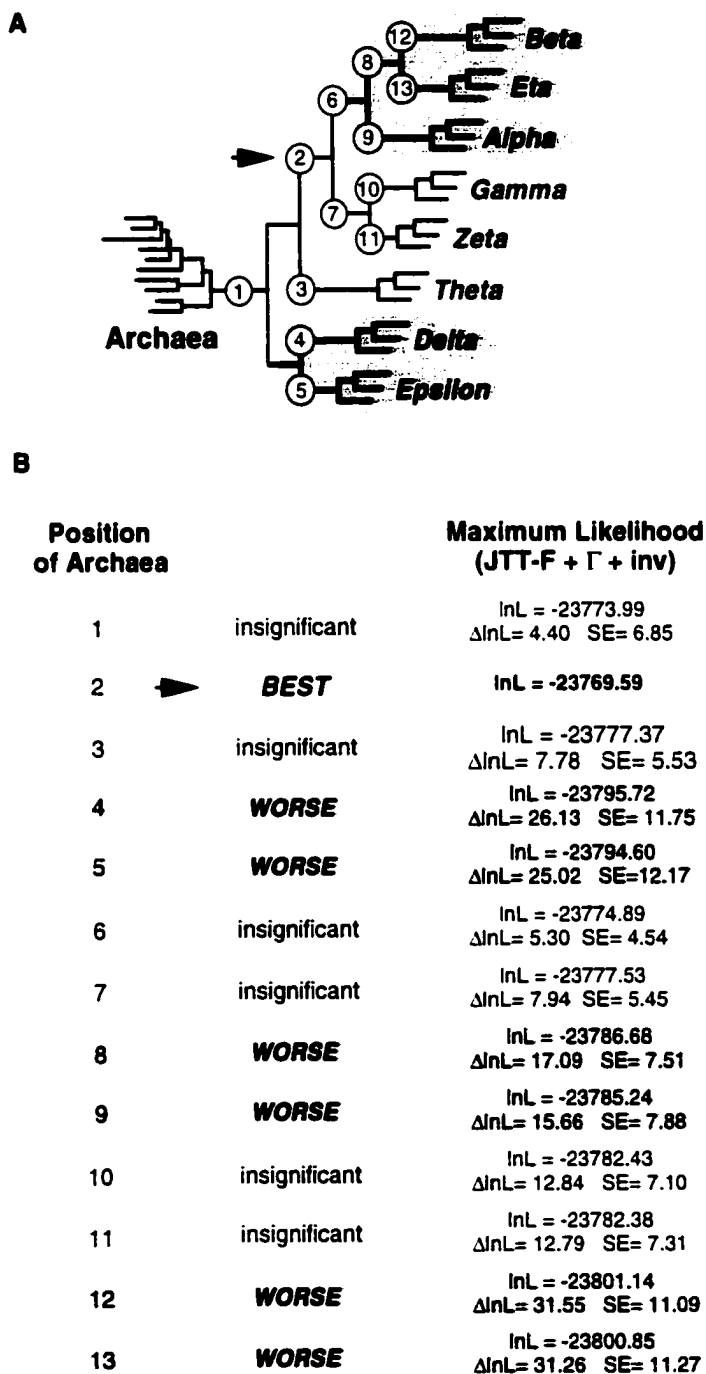


Figure 2.4 Phylogeny of group II chaperonin protein sequences

and parsimony methods (data not shown), as do CCT $\alpha$ ,  $\eta$  and  $\beta$  (although more weakly). For archaea, the clustering of the euryarchaeal sequences together is well supported, while the monophyly of the  $\alpha$  and  $\beta$  paralogs of crenarchaeotes is not. The ML tree shows the crenarchaeal  $\beta$  subunit sequences branching with the euryarchaeotes, suggesting that the  $\alpha/\beta$  paralogy in crenarchaeotes may predate their divergence from euryarchaeotes (this topology was observed with some but not all phylogenetic methods; data not shown). Interestingly, most of the deepest branches of the group II chaperonin tree are poorly resolved, even when the maximum number of alignable amino acid positions are used. The systematic exclusion of individual CCT paralogs from the analyses, most notably CCT $\zeta$  (the longest branch) and CCT $\theta$  (poorly conserved; see below), had little effect on the support for the relationships among the CCT subunits, suggesting that no particular subset of the data was the cause of the unstructured trees (data not shown). I therefore performed Kishino-Hasegawa tests (Kishino and Hasegawa 1989) in PUZZLE to assess the significance of alternative topologies to the ML tree, taking into account among-site rate heterogeneity. In these analyses, the optimal topology was slightly different from the protML tree in Figure 2.4 (which was the second-best tree; 0.64 SE difference), and placed the archaeal root between the CCT $\theta/\delta/\epsilon$  and CCT $\beta/\eta/\alpha/\gamma/\zeta$  clades (Figure 2.5A). Several other rootings were not considered worse at a 5% level of significance (e.g., the archaea as a sister group to CCT $\gamma$ ,  $\theta$  or  $\zeta$ ), but were between 1.2 and 1.8 SEs worse than the best tree (Figure 2.5B). Notably, placements of the archaeal root *within* the CCT $\delta/\epsilon$  and  $\alpha/\beta/\eta$  clades were significantly worse topologies, confirming the results of Figure 2.4, and suggesting that these paralogies occurred more recently in the evolution of CCT.

An additional duplication has also occurred within the evolution of the CCT $\zeta$  (zeta) gene family. Two distinct CCT $\zeta$  subunits (1 and 2) have been

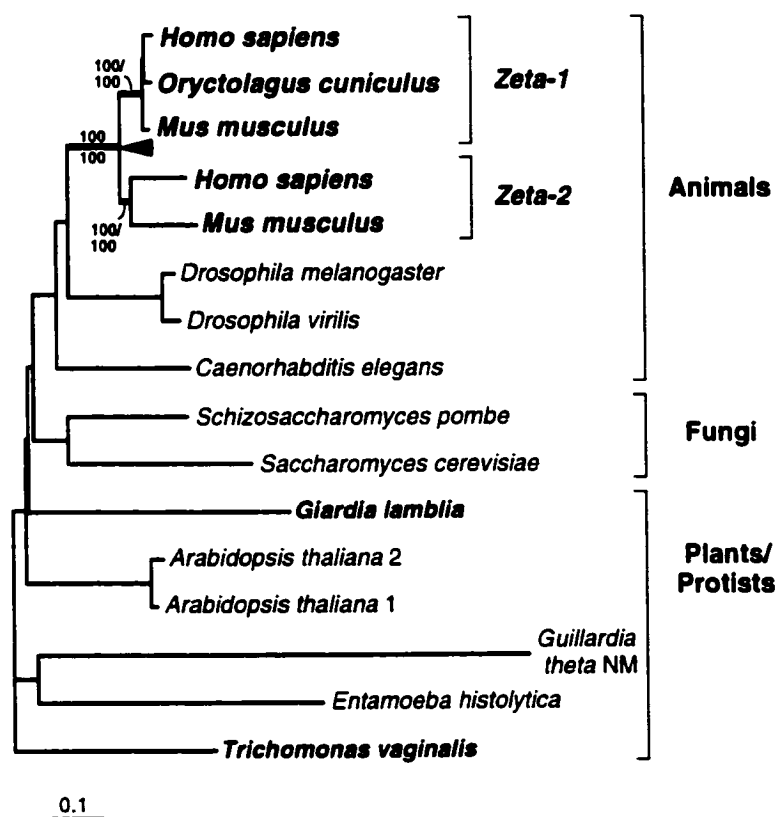


**Figure 2.5** Testing the position of the root of the eukaryotic CCT tree. (A) Schematic of the phylogeny shown in Figure 2.4, circles indicate alternative positions of the archaeal chaperonin root to the eukaryotic CCT tree tested by the method of Kishino and Hasegawa (1989) accounting for among-site rate variation (JTT +  $\Gamma$  + inv model). The optimal placement (node 2) is labeled with an arrow. (B) Significance of tree topologies in (A) over alternatives.  $\Delta$ InL values  $>1.96$  SE were considered significantly worse at the 5% level. Shading corresponds to schematic (A).

characterized in mouse (Hynes, Kubota and Willison 1995; Kubota *et al.* 1997), and the two sequences share approximately 81% amino acid sequence identity (Kubota *et al.* 1997). A phylogenetic analysis of all available CCT $\zeta$  sequences is shown in Figure 2.6. The CCT $\zeta$ -1 and  $\zeta$ -2 sequences in mouse and human are more closely related to each other than to CCT $\zeta$  sequences from other eukaryotes, suggesting that the gene duplication producing them occurred quite recently, likely within mammalian evolution. Interestingly, the CCT $\zeta$ -2 subunit gene is only expressed in testis, unlike the expression pattern observed for other CCTs, which are expressed in all tissues investigated thus far, including testis (Kubota, Hynes and Willison 1995a; Kubota *et al.* 1997). The significance of this observation is as yet unclear, but Kubota *et al.* (1997) suggest that the CCT $\zeta$ -2 subunit may mediate the folding of testis-specific proteins, such as  $\alpha$ -tubulin isoforms, by substituting for the CCT $\zeta$ -1 subunit in the CCT complex. Curiously, genes encoding two different CCT $\zeta$  subunits are also present in the genome of the flowering plant *Arabidopsis thaliana* (Figure 2.6), and duplicates of other CCT subunits are also present in mammalian genomes (data not shown).

### **Rates of evolution differ among CCT subunits**

The phylogenetic trees shown in Figures 2.2 and 2.4 show that the branch leading to the archaeal chaperonins is remarkably short compared to the branches leading to the different CCT subunits. The branch lengths within the various CCT clades also appear variable. To assess the significance of the latter observation, I calculated the percent amino acid identities shared between the mouse CCTs and the *Caenorhabditis*, *Saccharomyces*, *Giardia* and *Trichomonas* sequences, as well as the proportion of constant amino acid residues found in each individual CCT subunit alignment. The results (Figure 2.7) suggest differences in the degree of conservation of the individual CCT subunits. CCT $\theta$



**Figure 2.6** Phylogeny of CCTzeta protein sequences. The tree shown is an unrooted Fitch-Margoliash distance tree inferred from an alignment containing 16 sequences and 470 amino acid positions. The *Trichomonas* and *Giardia* sequences appear in bold. The putative gene duplication is indicated by an arrow, and the CCT zeta-1 and zeta-2 clades are highlighted. Statistical support for this clade is provided (distance bootstrap values and maximum likelihood RELL values, respectively). The scale bar indicates estimated number of amino acid amino acid substitutions per site.

<b>CCT subunit</b>	<b>% Identity with Mouse Ortholog<sup>1</sup></b>				<b>% Constant AA Residues<sup>2</sup></b>
$\alpha$ alpha	67.0	61.7	52.1	48.9*	27.5*
$\beta$ beta	66.7	65.6	58.5	—	37.7
$\delta$ delta	62.4	58.6	46.4*	51.9	36.8*
$\epsilon$ epsilon	67.5	59.9	57.6	—	37.5
$\eta$ eta	64.1	62.1	—	53.1	36.9
$\gamma$ gamma	60.5	60.0	50.3*	47.2*	21.3*
$\theta$ theta	55.0	49.5	41.8	—	14.5
$\zeta$ zeta 1	67.5	57.5	52.5	55.9	31.6
zeta 2	63.3	53.0	50.8	53.6	
	<b>Mus</b>	<b>Caenorhab.</b>	<b>Saccharo.</b>	<b>Giardia</b>	<b>Trichomonas</b>

**Figure 2.7** Rates of evolution differ among CCT subunits. The table shows % amino acid identities shared between mouse and *Caenorhabditis*, *Saccharomyces*, *Trichomonas* and *Giardia* CCT protein sequences, and the proportion of constant amino acid sites in alignments of CCT orthologs. <sup>1</sup>Pairwise amino acid identities were calculated in PAUP\* (Swofford, 1998) using a complete alignment (706 positions) of the eight CCT paralogs. <sup>2</sup>The proportion of constant amino acid residues within individual CCT subunit families were obtained using PUZZLE 4.0 (Strimmer and von Haeseler, 1997) from alignments that included only unambiguously aligned positions, no missing data and maximum taxonomic diversity (animal, fungal, plant and protist sequences). Dashes (—) indicate a sequence is absent for comparison, asterisks (\*) indicate a partial *Trichomonas* or *Giardia* sequence.



(and to a lesser extent CCT $\gamma$ ) appears to be the least conserved subunit, showing the lowest percent identity in all 'within-ortholog' comparisons. As well, only 14.5% of the amino acid residues in the CCT $\theta$  alignment were constant (this number dropped further to 9.5% when the divergent *Plasmodium* CCT $\theta$  sequence was included), compared to 21.3-37.7% constant residues in the other CCT subunit alignments. To statistically assess differences in the substitution rates of the different CCT paralogs, a molecular clock likelihood ratio test was performed with  $n-2$  degrees of freedom in PUZZLE (Strimmer and von Haeseler 1997) on an ML-distance tree of the eight eukaryotic CCTs (40 representative taxa and 355 sites). A molecular clock for the CCT paralogs was strongly rejected with a  $p$ -value  $<0.01$  (data not shown).

### **Regions of variability among the different CCT subunits**

An examination of the alignment of group II chaperonins shown in Figure 2.1 reveals that some regions of the molecule are variable in length as well as sequence. It is significant that these regions do not necessarily correspond to areas of variability within a particular CCT subunit family. In fact, many of the subunits possess insertions that are themselves invariant in sequence across a wide range of eukaryotic species, but are not present in any of the other subunits. To investigate what these observations might mean in terms of CCT subunit structure and function, the variable regions—those areas found to be highly variable in length and/or amino acid sequence—were mapped onto the crystal structure of the archaeal homolog of CCT, the *Thermoplasma acidophilum* thermosome (Ditzel *et al.* 1998).

The results (Figure 2.8) show that while regions of variability (highlighted gray) clearly map to all three 'functional' domains, the intermediate and apical domains appear to possess proportionately more variable sequence than does the



**Figure 2.8** Regions of variability among the different CCT subunits. Areas of the chaperonin protein sequence that were found to be highly variable in length and/or amino acid sequence between the 8 different CCT subunits were mapped onto the crystal structure of the  $\alpha$  subunit of the *Thermoplasma acidophilum* thermosome, the archaeal homolog of CCT (Ditzel *et al.* 1998). (A) Side view. The apical, intermediate and equatorial domains are labeled and color-coded, and the regions of the molecules found to be variable among the CCT subunits are highlighted gray. (B) Same structure as in (A), but viewed from within the central cavity of the thermosome. Secondary structural elements are labeled according to Ditzel *et al.* (1998) (H=helix, S=sheet).

equatorial domain. I found that 38.8, 38.9 and 28.8% of the total amino acid sequence in the apical, intermediate and equatorial domains was variable, and this difference was significant at  $p < 0.05$  in a chi-square test (data not shown). This observation is consistent with that of Kim, Willison and Horwich (1994), who found that the equatorial domain, which contains the ATP-binding site, appeared to be the most highly conserved among the different CCT subunits and between the CCTs and *E. coli* GroEL. It is also significant that the variable regions in these domains map predominantly to the exterior of the molecule. A cluster of variable sequence is clearly evident along most of the 'back-side' of the molecule (e.g., helix H8; Figure 2.8A), corresponding to the outer face of the monomer as it sits in the hetero-oligomer. Regions of variability also map to a loop under helix H11 of the apical domain and to the region surrounding helix H13 (Figure 2.8B). In the thermosome, helices H11 and H13 of adjacent subunits contact one another (Ditzel *et al.* 1998; see below).

Most interesting is the patch of variability that corresponds to the extreme tip of the apical domain (the 'helical protrusion'). In the *T. acidophilum* thermosome, this region appears to be highly flexible: the analysis of different crystal forms shows that the protrusion has the potential to tilt greater than  $20^\circ$  relative to the rest of the apical domain (Klumpp and Baumeister 1998). Bosch, Baumeister and Essen (2000) have crystallized the apical domain of the *T. acidophilum*  $\beta$  subunit and showed that this region adopts a different conformation than that found in the crystal structure of the  $\alpha$  subunit apical domain (Klumpp, Baumeister and Essen 1997). These authors suggest that sequence variations in the corresponding region of the CCT subunits could produce similar structural differences, and that such heterogeneity could mediate subunit-specific roles in substrate recognition. The extreme sequence divergence between the different CCT subunits in this region does make

heterogeneity in secondary and tertiary structure particularly likely. However, the exact role of the helical protrusion in the binding of substrate is unclear. While the electron microscopic data available for the CCT-actin complex (Llorca *et al.* 1999a) show actin binding well below the helical protrusions, tubulin seems to interact with a much broader region of the CCT apical domains, including parts of the helical protrusions (Llorca *et al.* 2000; see below).

### **Subunit-specific 'signatures' in CCT**

In a sequence- and structure-based comparison of group I and group II chaperonins, Kim, Willison and Horwich (1994) noted that the different CCT subunits were remarkably divergent, particularly in their apical domains, but that these regions were nevertheless highly conserved within the different subunit classes. To examine the pattern and degree of conservation in the different CCT subunits more closely, the rate of amino acid sequence evolution was estimated—site by site—across alignments of individual subunits that contained maximal taxonomic diversity (i.e., animal, fungal, plant and protist sequences). When these 'site-rates' were mapped onto an alignment containing all the CCT subunits, three general categories of amino acid sites were apparent: (1) conserved (slowly evolving) and identical amino acid residues present in multiple subunits (e.g., the ATPase domains), (2) conserved but different amino acid residues present in different subunits, and (3) poorly conserved/fast evolving residues (i.e., little or no evolutionary constraint) present in one or multiple subunits. The results are presented in Figure 2.9. Most notably, and consistent with published results (Kim, Willison and Horwich 1994), much of the divergence between the different CCT subunits corresponds to their apical domains, the region involved in the binding of substrate. However, I also detected differences in the degree of conservation and amino acid sequence of

**Figure 2.9** Site-rate analysis of CCT subunits. The *Trichomonas* and *Giardia* CCTs are shown aligned with representative mouse CCT homologs (21 sequences, 706 amino acids). Amino acid positions estimated to be slowly evolving and conserved in two or more CCT subunits are shaded gray; highly conserved residues and amino acid insertions unique to specific CCT subunits are boxed (see text). Dashes indicate gaps in the alignment. Functional domains predicted previously (equatorial, intermediate, apical and 'helical protrusion'; (Ditzel *et al.* 1998; Klumpp, Baumeister, and Essen 1997)) are indicated. Asterisks next to taxon names indicate partial sequences; asterisks under the alignment indicate amino acid residues used for the phylogenetic analysis of group II chaperonins. Taxon abbreviations: Tvag, *Trichomonas vaginalis*; Glam, *Giardia lamblia*; Mmus, *Mus musculus*. Subunit abbreviation: eps, epsilon. Alignments of individual CCT subunits were used to estimate site-by-site evolutionary rates and they contained the following taxa and number of sites: CCT $\alpha$ —338 sites (*Homo*, *Xenopus*, *Arabidopsis*, *Drosophila*, *Dictyostelium*, *Schistosoma*, *Caenorhabditis*, *Tetrahymena*, *Saccharomyces*, *Trichomonas*, *Giardia*), CCT $\beta$ —358 sites (*Homo*, *Saccharomyces*, *Schizosaccharomyces*, *Caenorhabditis*, *Plasmodium*, *Giardia*), CCT $\delta$ —361 sites (*Homo*, *Caenorhabditis*, *Fugu*, *Saccharomyces*, *Schizosaccharomyces*, *Glycine*, *Trichomonas*), CCT $\epsilon$ —360 sites (*Homo*, *Caenorhabditis*, *Drosophila*, *Plasmodium*, *Saccharomyces*, *Avena*, *Cumcumis*, *Arabidopsis*, *Giardia*), CCT $\eta$ —360 sites (*Homo*, *Caenorhabditis*, *Saccharomyces*, *Schizosaccharomyces*, *Plasmodium*, *Tetrahymena*, *Trichomonas*), CCT $\gamma$ —287 and 353 sites (*Homo*, *Caenorhabditis*, *Xenopus*, *Drosophila*, *Arabidopsis*, *Tetrahymena*, *Oxytricha*, *Saccharomyces*, *Schizosaccharomyces*, *Leishmania*, with and without *Trichomonas*, *Giardia*, respectively), CCT $\theta$ —359 sites (*Homo*, *Caenorhabditis*, *Candida*, *Saccharomyces*, *Schizosaccharomyces*, *Tetrahymena*, *Giardia*), CCT $\zeta$ —360 sites (*Homo* (zeta1, zeta2), *Caenorhabditis*, *Drosophila*, *Saccharomyces*, *Schizosaccharomyces*, *Trichomonas*, *Giardia*). Where missing data precluded site-rate calculations (eg. partial *Trichomonas* and/or *Giardia* sequences), amino acid positions were considered slowly evolving if present in all taxa in a particular subunit alignment. Conservative amino acid substitutions were also taken into consideration.



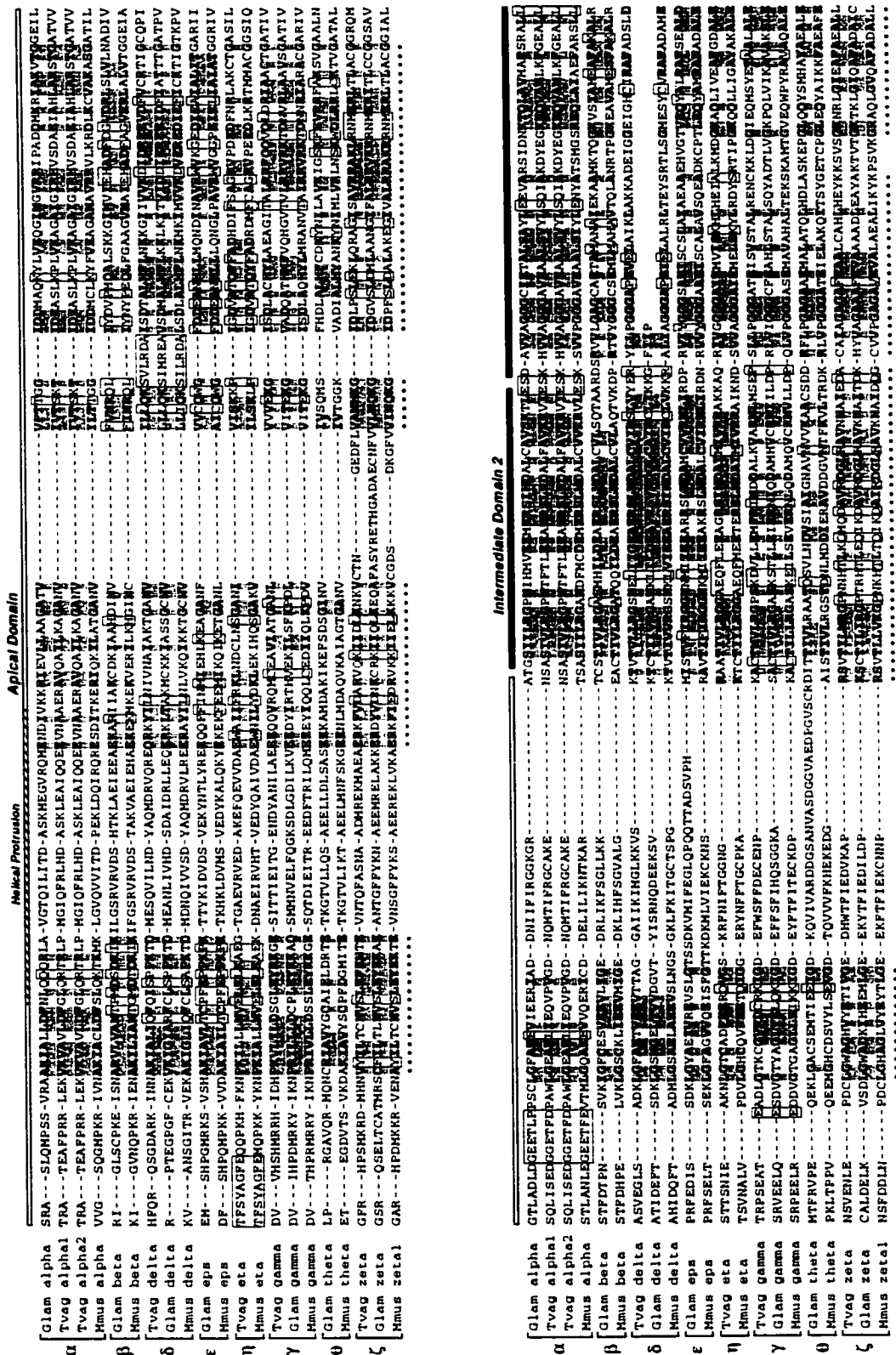


Figure 2.9 Site rate analysis of CCT subunits

Equatorial Domain 2

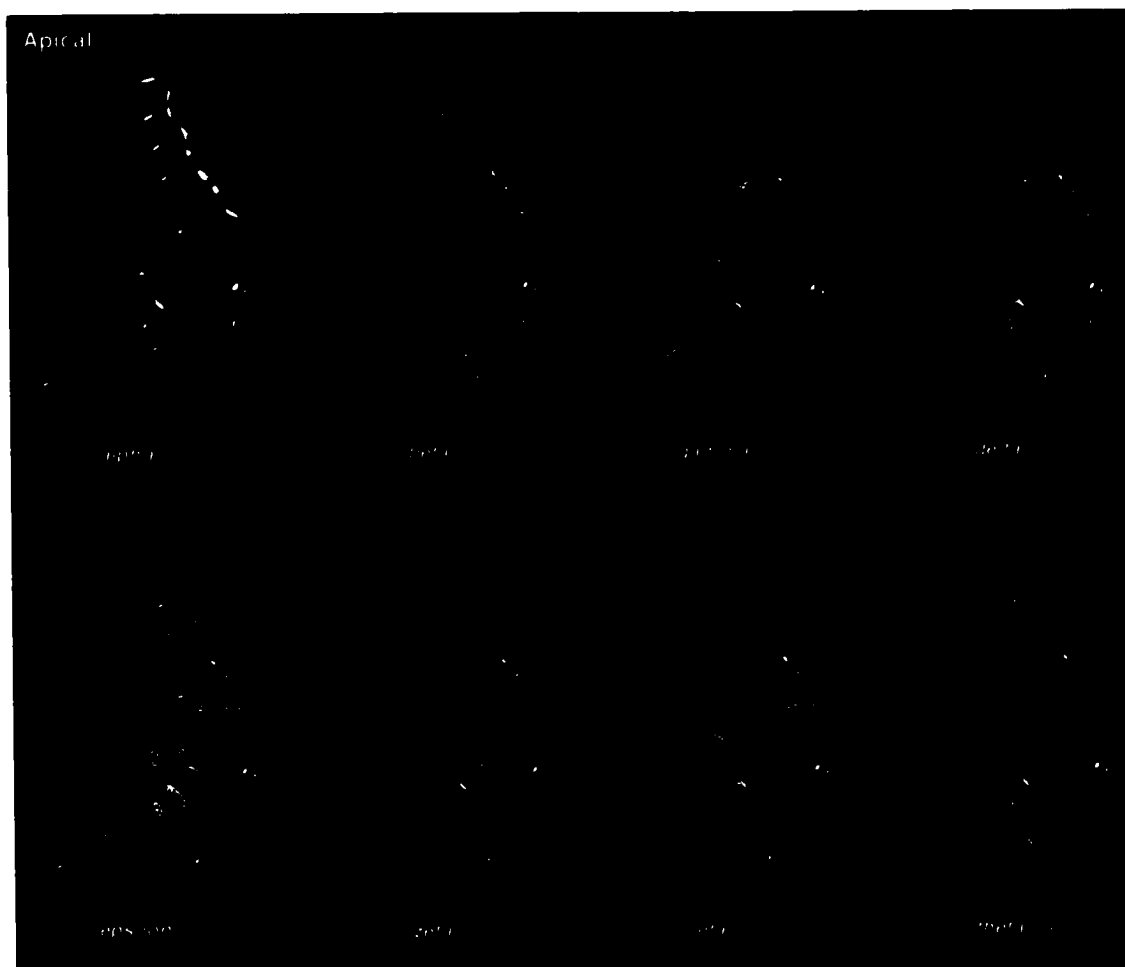
[Glam alpha	NRKQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Tvag alpha1	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus alpha2	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus alpha	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Glam beta	SEYFNHNPQDGEKAVSKKALHNRAG--KTDFGLDHRV--GTVRI--VRRHAGITTEYVCFCHVVLVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus beta	HISVYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Tvag delta	AVYVYKASVGLAPFDVYVQVQVHAEISSG--KCFPAALDLVH--GKIRD--GHKDGVIKPCQ--LSUQVYK
[Mmus delta	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Glam eps	IEELTAAKESIOEIKETLANNVLAQKRTG--SHHLGIDCHO--RMTID--HKEOSVFTLSVVOCHLIVVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus eps	VIEPHALSESHHNSHTOTTEVEVAROVKES--NPAIGIDCLH--KGSID--HOYQHVIFETLIGKCOHSLGQVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Tvag eta	AVYVYKASVGLAPFDVYVQVQVHAEISSG--KCFPAALDLVH--GKIRD--GHKDGVIKPCQ--LSUQVYK
[Mmus eta	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Glam gamma	YVYVYKASVGLAPFDVYVQVQVHAEISSG--KCFPAALDLVH--GKIRD--GHKDGVIKPCQ--LSUQVYK
[Mmus gamma	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Glam theta	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus theta	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Glam zeta	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH
[Mmus zeta1	VYRIYDQVYRSLDATTAEKAVHARAILKETPAEENEKLRHRLDLOH--GVICD--IHVQAGVLEPFRIRIKENSVYVAVNUTLITLIDDSIRLNPEOOOSVODHH

Figure 2.9 Site rate analysis of CCT subunits



the putative ATP-binding domains in the different subunits, as well as the presence of highly conserved 'paralog-specific' motifs present in the equatorial and intermediate domains (Figure 2.9). For example, the CCT $\theta$  subunit has the sequence GDGTN in the amino-terminal equatorial domain, a slight variation on the near-universally conserved motif GDGTT, which forms one of the loops of the ATP-binding pocket. In CCT $\gamma$ , the highly conserved motif NDGAT (found in the amino-terminal equatorial domain of almost all the CCT subunits) has changed to NDGNA. Interestingly, multiple alanine replacements of highly conserved residues within the ATP-binding pocket of CCT $\zeta$  (e.g., GDGTT to AAAAA) had relatively mild effects on cell growth in yeast (Lin *et al.* 1997), unlike the severe phenotypes observed with mutations in the same region of CCT $\alpha$  (Miklos *et al.* 1994; Ursic *et al.* 1994). When the sequences from a wide diversity of eukaryotes are considered, some parts of the ATP-binding domains of the different subunits appear to be less conserved than others, suggesting that their ATPase functions may in fact be redundant.

As was done for the variable regions of the molecule, the positions of the slowly evolving subunit-specific 'signatures' were mapped onto the structure of the  $\alpha$  subunit of the *T. acidophilum* chaperonin (Figures 2.10 and 2.11). Within the conserved 'core' region of the apical domain, many of the CCT subunits possess a cluster of highly conserved, subunit-specific residues on the inside face of the apical domain, just below the helical protrusion (Figure 2.10). Significantly, this is precisely the region of the apical domain that appears to be in direct contact with actin and tubulin in CCT-substrate complexes (Llorca *et al.* 2000; Llorca *et al.* 1999a). CCT $\beta$  and  $\epsilon$ , two of the three subunits implicated in the binding of actin (Llorca *et al.* 1999a), have more subunit-specific signatures in this region than the other subunits, while CCT $\theta$  has almost none (Figure 2.10). As well, CCT $\alpha$  and CCT $\zeta$  have unique insertions in this area, very near helix H11 (data not shown).



**Figure 2.10** Subunit-specific 'signatures' in individual CCT apical domains. Isolated apical domains are shown, one for each of the eight CCT subunits ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$  and  $\theta$ ). Amino acid residues unique to a particular CCT subunit apical domain ('signatures'; see text) are highlighted yellow. As in Figure 2.8, regions of the sequence found to be variable between the different CCT subunits are shaded gray. The domain orientation and color coding matches that in Figure 2.8B. Secondary structural elements are labeled according to Ditzel *et al.* (1998) (H=helix).



**Figure 2.11** Subunit-specific signatures in individual CCT intermediate and equatorial domains. Isolated intermediate (A) and equatorial (B) domains are shown, one for each of the eight CCT subunits ( $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$  and  $\theta$ ). Domain color coding matches that in Figure 2.8. Amino acid residues unique to a particular CCT subunit are highlighted yellow. Variable regions are shaded gray. The domain orientation is the same as that in Figure 2.8B, and secondary structural elements are labeled according to Ditzel *et al.* (1998) (H=helix, S=sheet).

This region is not indicated as 'variable' in Figures 2.8 and 2.10 due to the fact that the CCT $\alpha$  and CCT $\zeta$  insertion boundaries are highly conserved with respect to the sequence of the other CCTs and the archaeal chaperonins.

It is interesting that many of the highly conserved subunit-specific amino acids present in the apical domain are charged residues, not hydrophobic ones like those implicated in substrate binding in the *E. coli* GroEL system (Figure 2.9; Kim, Willison and Horwich 1994; Klumpp, Baumeister and Essen 1997; Willison 1999). This could reflect a fundamentally different mechanism of substrate recognition for the group II chaperonins than that observed for GroEL, one based on hydrogen bonding and electrostatic interactions instead of hydrophobicity (Klumpp, Baumeister and Essen 1997; Willison 1999). However, as noted by Klumpp, Baumeister and Essen (1997), the helical protrusions of the CCT subunits may contain a sufficient number of hydrophobic residues to mediate substrate binding through hydrophobic interactions. As mentioned previously, the CCT-tubulin complex (Llorca *et al.* 2000) is consistent with this possibility.

Several areas of the equatorial domain also appear to be 'hotspots' for unique amino acid substitutions. For example, all eight subunits possess unique residues in the vicinity of a small turn between helix H14 and strand S20 (Figure 2.11A). In the *T. acidophilum* thermosome, this region (together with turn segments between helices H7 and H8, and strands S7 and S8) contacts the equatorial domain of the adjacent subunit (Ditzel *et al.* 1998). It is thus likely that subunit-specific motifs in these regions (as well as those in the vicinity of helices H11 and H13 in the apical domain; above) are at least partly responsible for the assembly of a CCT complex with a unique arrangement of subunits. Interestingly, the CCT $\zeta$  subunit possessed a large number of unique residues in these areas: a string of unique substitutions map to the 'outer' surface of helix H14 and to strand S7 (Figure 2.11A), both of which face the central cavity in the

thermosome structure (Ditzel *et al.* 1998). The significance of this observation, in terms of chaperonin structure and function, is unclear.

## DISCUSSION

### **Eukaryotic chaperonin evolution: deep paralogy**

Gene duplications, gene conversions and gene losses make the reconstruction of ancient molecular events difficult. Group II chaperonins are a striking example—lineage-specific gene duplication, gene conversion and gene loss has occurred in archaea, and a remarkably ancient and complex paralogy exists in eukaryotes. The data in this chapter bear on several aspects of the origin and evolution of the completely hetero-oligomeric CCT in eukaryotes and on the origin of the eukaryotic cell itself.

The CCT genes presented here from *Trichomonas* and *Giardia*, two of the most divergent eukaryotes presently known, show strong affinity for each of the eight CCT subunit families found in 'higher' eukaryotes. The gene duplications producing the different subunits clearly occurred very early in the evolution of the eukaryotic cell, and it is unlikely that the loss of any one of the CCT paralogs could, at this stage, be tolerated. The essential nature of at least six (and likely all) of the eight CCT genes in yeast (Lin *et al.* 1997; Stoldt *et al.* 1996), and their seemingly universal distribution in the diverse eukaryotic lineages examined here speak to that constraint. It has been suggested (Willison and Horwich 1996) that CCT evolved from an eight-fold symmetric chaperonin complex like that in the crenarchaeote *Pyrodictium occultum* (Phipps *et al.* 1991, 1993), based on the near-universal distribution of eight-membered ring structures among group II chaperonins (*Sulfolobus* being the only exception; Marco *et al.* 1994). The  $\alpha$  and  $\beta$  subunits of crenarchaeotes (the deepest paralogy in archaea; Archibald, Logsdon

and Doolittle 1999) do not branch preferentially with particular subsets of CCT paralogs, however, as would be expected if the paralogy predated the divergence of crenarchaeotes and eukaryotes: there is no sense in which particular archaeal chaperonin paralogs are more closely related to some CCT paralogs than to others. While the ancestral chaperonin complex in eukaryotes was likely composed of eight-membered rings, it appears that CCT became hetero-oligomeric independent of the chaperonin complexes in archaea.

Attempts to resolve the relative branching order of the CCT paralogs and thus determine the order in which CCT 'acquired' so many different subunits were met with limited success. Unlike other paralogous 'eukaryote-specific' gene families such as actins and tubulins, which have very distantly related prokaryotic homologs, the eukaryotic CCTs have relatively close archaeal homologs to serve as an outgroup. Phylogenetic analyses suggest that the CCT $\delta/\epsilon$  and CCT $\alpha/\eta/\beta$  clades are the product of more recent gene duplications; however, the exact placement of the archaeal root on the CCT tree remains unclear. The tree topology shown in Figure 2.4 illustrates one possible scenario of CCT subunit diversification, but the results of Kishino-Hasegawa (KH) tests (Figure 2.5) suggest that others are also compatible with the data. This lack of resolution may be due to the fact that numerous gene duplications (and sequence divergences) occurred in quick succession, or that chromosome or whole genome duplications produced a dramatic increase in the number of CCT genes simultaneously. Alternatively, the molecules themselves may simply be 'saturated', having accumulated a sufficient number of amino acid substitutions as to make accurate reconstruction of their evolutionary history impossible.

It should be mentioned that the KH test, as implemented here, is in some cases biased towards the rejection of 'sub-optimal' trees. A modified version of the test recently developed by Shimodaira and Hasegawa (1999) aims to

eliminate this problem, but at present there is considerable debate over its proper implementation (Goldman, Anderson and Rodrigo 2000; Shimodaira and Hasegawa 1999). With this in mind, the results presented in Figure 2.5 should be interpreted with caution, particularly with respect to the rejection of the archaeal root falling within the CCT $\alpha$ / $\eta$ / $\beta$  clade. While the bootstrap support for the CCT $\delta$ / $\epsilon$  grouping was always high (even when extremely conservative alignments were used; see Figure 2.3), the CCT $\alpha$ / $\eta$ / $\beta$  cluster was generally less well supported and was not observed with all datasets and phylogenetic methods. Despite these methodological concerns, the data suggest that CCT underwent intermediate stages of hetero-oligomerism, perhaps similar to the degree observed in present-day archaeal chaperonin complexes, and that the CCT $\delta$  and  $\epsilon$  (and possibly  $\alpha$ ,  $\beta$ , and  $\eta$ ) subunits represent a more recent divergence in eukaryotic chaperonin evolution.

Kubota *et al.* (1994) suggested that all CCT subunits should be present in all eukaryotes, and estimated a divergence time of two billion years for the different CCT paralogs, based on the assumption that the amino acid substitution rate of each CCT subunit family has been constant. The data presented here are consistent with the former prediction, but indicate that a clock-like rate of sequence divergence for each of the eight CCT paralogs is clearly not the case. I observed striking differences in the degree of conservation of the individual CCT subunits, as well as paralog-specific highly conserved sequence motifs (Figures 2.1 and 2.9). CCT $\theta$  appears to be the least conserved subunit, and may have reduced/different functional constraints. The results of recent biochemical studies (Liou, McCormack and Willison 1998; Liou and Willison 1997) support this notion: compared to the other CCTs, unique subunit-subunit binding properties were observed for CCT $\theta$  *in vitro*, as was a much reduced level of CCT $\theta$  mRNA relative to the other CCT genes (Liou and Willison 1997). From this

perspective, and in light of the phylogenetic analyses presented here, amino acid identity comparisons which suggest that the eight CCT subunits are approximately equally related to each other (Kubota *et al.* 1994; Kubota, Hynes and Willison 1995a) are misleading.

### **Co-evolution of chaperonin and substrate?**

Why are there so many CCT paralogs? Willison and co-workers have speculated that the duplication and differentiation of the different CCT subunits early in eukaryotic evolution was concurrent with (and facilitated) the evolution of the cytoskeleton (Willison and Horwich 1996; Willison and Kubota 1994). Unlike GroEL, which appears to service a broad range of substrates in the bacterial cytoplasm (Houry *et al.* 1999), CCT is thought to be more 'specialized'. As mentioned previously, actins and tubulins, the major cytoskeletal proteins of eukaryotic cells, appear to be the predominant substrates of CCT (Kubota, Hynes and Willison 1995a; Willison and Kubota 1994), although others continue to be discovered (Farr *et al.* 1997; Feldman *et al.* 1999; Melki *et al.* 1997; Srikakulam and Winkelmann 1999; Won *et al.* 1998). Using cryoelectron microscopy, Llorca *et al.* (1999a, 2000) have recently provided strong evidence for interactions between actins and tubulins and the apical domains of specific CCT subunits within the central chamber of CCT. It is thus likely that the cytoskeletal proteins and the various CCT subunits have co-evolved with one another (Hartl 1996; Willison 1999; Willison and Horwich 1996). While the data presented here suggest that CCT went through intermediate stages of hetero-oligomerism, the process by which such hetero-oligomerism initially arose, and how CCT ultimately became completely hetero-oligomeric is still a mystery. There are theoretically 5040 possible combinations of subunits in CCT (Liou and Willison 1997), yet all the



available data suggest a unique arrangement of subunits exists *in vivo* (Llorca *et al.* 2000, 1999a; Liou and Willison 1997).

In the previous chapter, I proposed a model for the origin of hetero-oligomerism in archaeal chaperonin complexes that emphasized co-evolution between duplicate *subunits*, as opposed to subunit and substrate. Simply put, a series of mutations in the subunit-subunit contact regions of one subunit followed by compensatory changes in a duplicate subunit could produce a tendency toward the assembly of an ordered arrangement of subunits within a chaperonin ring, as in the *T. acidophilum* thermosome. Such a process could eventually lead to obligatory hetero-oligomerism, even in the absence of 'specialized' roles for the duplicate subunits in protein folding.

With respect to the origin of hetero-oligomerism in CCT, the issue becomes a question of which evolved first: subunit-specific roles in protein folding or an ordered arrangement of subunits within a chaperonin ring? Given that CCT-actin and CCT-tubulin interactions are both subunit-specific and geometry-dependent, the evolution of a unique arrangement of the subunits relative to one another would have had to precede (or at least be concurrent with) the 'specialization' of the subunits themselves. It is certainly clear that the different CCTs have diverged substantially from one another, not only in their substrate-binding apical domains, but in regions involved in ATP binding and in subunit-subunit contacts (Figures 2.9, 2.10 and 2.11; see also Kim, Willison and Horwich 1994). Genetic studies in yeast support this notion. Lin *et al.* (1997) performed a comprehensive mutational analysis of the CCT $\zeta$  subunit (CCT6p in *Saccharomyces*) and found this subunit to be sensitive to mutations in the equatorial, intermediate and apical domains. Presumably, some of these mutations were deleterious not because of deficiencies in substrate recognition or folding, but because of an inability of the CCT $\zeta$  subunit to interact with the other

subunits to properly assemble the hetero-oligomeric complex. In this sense, it seems appropriate to view a particular subunit's function in terms of its contribution to the proper formation of the hetero-oligomeric CCT particle as well as to the binding of substrate(s).

Llorca *et al.* (1999a) showed that actin sits in the central cavity of CCT in a 1,4 configuration. CCT $\delta$  interacts with the small domain of actin while either CCT $\beta$  or  $\epsilon$  contacts the large domain. Given this fact, one might expect CCT $\beta$  and  $\epsilon$  to be quite similar. However, if the actin-CCT $\delta$  interaction is isomorphous in both arrangements, the CCT $\beta$  and  $\epsilon$  binding sites on the large domain of actin would have to be in different locations. Indeed, the phylogenetic analyses presented here show that CCT $\beta$  and  $\epsilon$  are no more closely related to one another than to any of the other subunits (in fact, CCT $\delta$  and  $\epsilon$  appear most closely related; Figure 2.4). Further, the two subunits don't appear to share a great degree of sequence similarity in their putative substrate binding domains, no more than is shared with any other subunit. It therefore seems likely that the actin-CCT $\beta$  and actin-CCT $\epsilon$  interactions are very different. Similarly, no detectable correlation was observed between the various subunits implicated in binding the different domains of tubulin. From the perspective of substrate diversity, it is interesting that  $\alpha$ - and  $\beta$ -tubulin appear to interact with CCT in an identical fashion (Llorca *et al.* 2000), despite the fact that they share only approximately 40% sequence identity (see Ritco-Vonsovici and Willison 2000 for discussion; the diversification of the  $\alpha$ -,  $\beta$ - and  $\gamma$ -tubulin homologs occurred early in eukaryotic evolution (Keeling and Doolittle 1996; Roger 1999), similar to the pattern observed for the CCT subunits themselves). The exact nature of the specificity between the apical domains of the CCT subunits and their various substrates is thus likely to be extraordinarily complex, perhaps involving a combination of hydrophobic and non-hydrophobic interactions (Leroux and Hartl 2000; Ritco-Vonsovici and

Willison 2000). The subunit-specific residues identified here for CCT should make excellent targets for mutational studies aimed at probing the structure and function of the CCT subunits, in terms of their contribution to substrate binding, ATPase activity and the overall assembly and function of CCT.

Interestingly, the evolutionary pattern of the group II chaperonins bears strong resemblance to that of the proteasome, a barrel-shaped proteolytic complex found in archaea, in the eukaryotic cytosol and in some bacteria. Archaea possess single  $\alpha$  and  $\beta$  subunits (Baumeister *et al.* 1998), while eukaryotes possess seven  $\alpha$  and seven  $\beta$  paralogs (subunits), one for each position in the seven-membered  $\alpha$  and  $\beta$  rings (Groll *et al.* 1997). In both chaperonins and proteasomes, the evolution of single hetero-oligomeric particles, instead of multiple distinct homo-oligomeric ones, suggests that co-evolution between duplicate subunits has been a significant factor in shaping their architectures.

It is clear that gene duplication and gene loss have been, and still are, prominent forces in archaeal and eukaryotic chaperonin evolution. A 'recent' CCT gene duplication in mammals (CCT $\zeta$ -1, -2; Kubota *et al.* 1997), a probable *Sulfolobus*-specific paralogy in crenarchaeotes (Archibald, Logsdon and Doolittle 1999), and the presence of multiple copies of CCT paralogs in *Trichomonas* (and undoubtedly many other eukaryotes) indicate that chaperonin gene duplication is an ongoing process. In the previous chapter I presented phylogenetic evidence for 'recent' gene loss in the two different euryarchaeal species. An even more striking case can be inferred for yeast CCTs. Genome sequence analyses in *Saccharomyces* suggest that a whole genome duplication may have occurred after its divergence from *Kluyveromyces* (Wolfe and Shields 1997); given the fact that CCT was already completely hetero-oligomeric at this time (i.e., yeast had at least eight CCT genes), the presence of exactly eight CCT genes in the present-day

yeast genome (Stoldt *et al.* 1996) indicates that multiple CCT duplicates have been lost.

The evolutionary forces influencing the retention of duplicate chaperonin genes are less obvious. What is becoming clear is that many of the complex paralogies unique to eukaryotic genomes (e.g.,  $\alpha$ - and  $\delta$ - DNA polymerases (Edgell, Malik and Doolittle 1998),  $\alpha$ - and  $\beta$ -tubulins (Keeling and Doolittle 1996) and RNA polymerases I, II and III (Stiller, Duffield and Hall 1998) were present early in eukaryotic evolution. The CCT gene family examined here is the most extreme example thus far. A tendency towards highly paralogous gene families (and more 'complex' macromolecular machinery) in eukaryotes compared to prokaryotes may reflect fundamental differences in the ways in which prokaryotic and eukaryotic genomes evolve. Larger genomes with multiple linear chromosomes should reduce the probability of gene conversion between recent (unlinked) duplicates, and, in general, more easily accommodate duplicate genes (offsetting the effects of random gene loss). As well, chromosomal or whole genome duplications provide a ready mechanism for doubling the number of paralogs present in a genome. Inherent differences in the mechanisms and frequency of gene/chromosome/genome duplication and gene conversion/loss could influence the initial retention of duplicate genes as much as the positive selection for new paralog-specific functions.

## CHAPTER III

### Molecular Chaperones Encoded by a Reduced Nucleus— the Cryptomonad Nucleomorph

This chapter includes work published in Archibald, J. M., T. Cavalier-Smith, U.-G. Maier, and S. Douglas. 2001. Molecular chaperones encoded by a reduced nucleus—the cryptomonad nucleomorph. *J. Mol. Evol.* (in press).

#### INTRODUCTION

Molecular chaperones function in most eukaryotic cellular compartments. In addition to mediating the proper folding of nascent proteins, molecular chaperones are involved in the disassembly of oligomeric protein complexes, directing the conformational maturation of signal-transducing molecules and facilitating the degradation of unstable proteins. Molecular chaperones also play an important role in guiding the translocation of proteins across organellar membranes. HSP70, HSP90 and the chaperonins are three of the most ubiquitous and well-studied classes of ATP-dependent molecular chaperones, and together share the common feature of recognizing their target substrates *via* interactions with hydrophobic regions exposed on non-native proteins (see Bukau *et al.* 2000; Bukau and Horwich 1998; Johnson and Craig 1997 for review).

The correct folding and transport of nascent proteins is especially complicated in organisms such as cryptomonads, photosynthetic algae that acquired their plastids by secondary endosymbiosis. Cryptomonads are unusual in that they possess an additional cytoplasmic compartment (the periplastid space), which is the remnant cytosol of a red algal endosymbiont. Within the periplastid space is a diminutive nucleus (the nucleomorph) that encodes mostly

genes for its own expression as well as a few needed by the plastid (Gilson, Maier and McFadden 1997; McFadden *et al.* 1997). The periplastid space is surrounded by a dual-membrane system called the 'chloroplast endoplasmic reticulum', or CER (Gibbs 1979), the outer envelope of which is continuous with the host nuclear envelope. In these organisms, nucleus-encoded gene products destined for the plastid must therefore traverse an extra pair of membranes (Cavalier-Smith 2000).

An early model for protein trafficking in cryptomonads proposed that nucleus-encoded polypeptides translated on host ribosomes present on the CER were transported to the plastid by vesicles budding from the periplastid membrane (derived from the plasma membrane of the former symbiont) and fusing with the outer plastid envelope membrane (Gibbs 1979; Gibbs 1981). However, other roles for such vesicles have been suggested (Cavalier-Smith 1999), and it is more likely that nucleus-encoded polypeptides are translocated across the CER and the periplastid membrane (by a still unknown mechanism) directly into the periplastid space (Cavalier-Smith 2000; Deane *et al.* 2000; Ishida, Green and Cavalier-Smith 2001). Molecular chaperones that function in the periplastid space are therefore probably involved in (1) mediating the initial translocation process itself, (2) facilitating the proper folding of newly imported and newly synthesized proteins destined to function in the periplastid space, and (3) assisting the translocation of nucleus- and nucleomorph-encoded polypeptides through both membranes of the plastid envelope.

The gene for one such chaperone, a cytosolic HSP70, has been found in the nucleomorph genome of the cryptomonad alga *Rhodomonas salina* (*Pyrenomonas salina*; Hofmann *et al.* 1994; Rensing and Maier 1994). Hsp70 genes have also been isolated and sequenced from the plastid (Wang and Liu 1991) and the nucleus (Rensing *et al.* 1997), but not the nucleomorph, of another cryptomonad,

*Guillardia theta*. Genes encoding the organellar chaperonin cpn60 have been isolated and sequenced from both the plastid (Douglas and Penny 1999) and nucleomorph (Wastl *et al.* 1999) of *G. theta*, and presumably mediate the folding of plastid-targeted proteins.

The cryptomonad periplastid space—essentially a highly reduced eukaryotic cytoplasm—contains very few of the normal cell constituents. Apart from the plastid and the nucleomorph itself, only starch grains, 80S ribosomes and smooth periplastid vesicles are ultrastructurally visible; it lacks mitochondria, peroxisomes and lysosomes or large vacuoles, as well as any visible rough endoplasmic reticulum (ER), Golgi complex or cytoskeleton. The nucleomorph was generally thought to divide amitotically (McKerracher and Gibbs 1982; Morrall and Greenwood 1982), as electron-microscopic studies failed to detect the presence of cytoskeletal elements normally associated with a mitotic apparatus (Gillott and Gibbs 1980; McKerracher and Gibbs 1982; Morrall and Greenwood 1982). However, the recent identification of alpha-, beta-, and gamma-tubulin genes (the major protein components of microtubules) in the *G. theta* nucleomorph genome (Keeling *et al.* 1999; Zauner *et al.* 2000) strongly suggests that some microtubule-related structures are present. Other than HSP70 (Hofmann *et al.* 1994; Rensing and Maier 1994), evidence for molecular chaperones that might function in the periplastid space of cryptomonads is currently lacking. In this chapter I present an analysis of *G. theta* nucleomorph-encoded genes for the cytosolic HSP70 and HSP90 classes of chaperone, eight different subunits of the cytosolic chaperonin, CCT, as well as a heat-shock transcription factor (HSF) known to be involved in the expression of heat shock proteins.

As discussed in the previous chapter, the predominant *in vivo* substrates of CCT seem to be tubulins and actins (Kubota *et al.* 1994; Willison and Horwich

1996). Guiding the proper folding of the nucleomorph-encoded tubulins may therefore be an important function for CCT in *G. theta*. Missing from the *G. theta* nucleomorph genome are genes for the ER homologs of HSP70 and HSP90 as well as the co-chaperones HSP40/dnaJ and hip, hop and prefoldin, which assist HSP70, HSP90 and CCT, respectively. This suggests that these proteins are no longer needed or are encoded in the host nuclear genome and targeted into the periplastid space.

HSP90 is the most abundant cytosolic protein in eukaryotic cells (Jakob and Buchner 1994) and, together with other proteins, is responsible for the conformational maturation of signal-transducing proteins such as steroid-hormone receptors and protein kinases (Bohen, Kralli and Yamamoto 1996). Interestingly, HSP90 is also known to associate with the centrosome and participate in cell-cycle control in animals, probably by facilitating interactions with centrosomal proteins (Lange *et al.* 2000) such as gamma-tubulin, with which it co-localizes (Zarzov, Boucherie and Mann 1997). HSP90 has been shown to play a role in spindle-pole-body duplication in *Saccharomyces cerevisiae* (Lange *et al.* 2000) and to associate with HSP70 to form a multichaperone (Scheufler *et al.* 2000). The presence and properties of the *G. theta* nucleomorph HSP70 and 90 homologs described here, together with the discovery of gamma-tubulin and the centrosomal protein ranbpm (Zauner *et al.* 2000), raise the possibility that all four proteins interact to form a relict mitotic apparatus of red algal origin in the cryptomonad nucleomorph, even though cytological evidence for nucleomorph mitosis has not yet been found.



## RESULTS

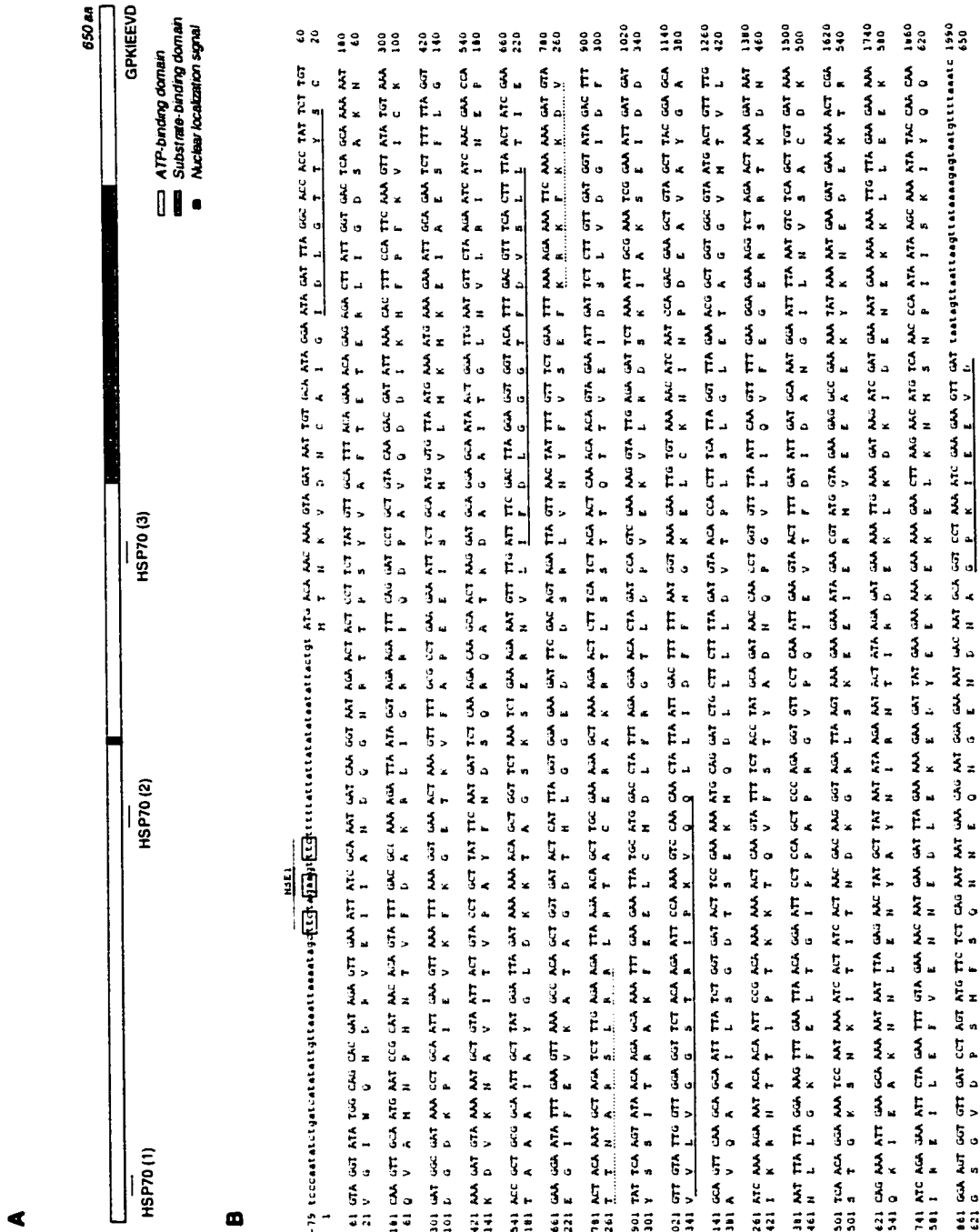
### The *Guillardia theta* nucleomorph genome

The sequencing of the *Guillardia theta* nucleomorph genome is described by Zauner *et al.* (2000) and Douglas *et al.* (2001). Putative genes were identified by BLASTP, BLASTX and BLASTN computer searches of the public databases (<http://www.ncbi.nlm.nih.gov/BLAST>) and the web tool pSort (<http://psort.nibb.ac.jp/>) was used to identify targeting signals in the inferred protein sequences.

### HSP70

A single *hsp70* gene was found on chromosome 1 of the *G. theta* nucleomorph genome (Zauner *et al.* 2000). To obtain additional sequences for comparative study and phylogenetic analysis, I searched the public databases by BLAST (Altschul *et al.* 1990) using this sequence as a query. HSP70 protein sequences (both cytosolic and ER forms) were obtained for a variety of animals, fungi, plants and protists, as well as bacteria. The amino acid sequence inferred from the *G. theta hsp70* gene was added to an HSP70 alignment manually based on globally conserved regions (see Materials and Methods), as were the HSP70s obtained from the databases. An alignment of 51 eukaryotic HSP70 sequences and 526 unambiguously aligned amino acid positions was used for phylogenetic reconstruction.

The *G. theta hsp70* gene encodes a predicted protein of 650 amino acids in length (72.485 kDa), which contains the highly conserved motif GPKIEEVD at its carboxy terminus (Figure 3.1). It also possesses all three of the conserved consensus patterns for HSP70 proteins at positions 12-19, 204-217 and 341-355 (Figure 3.1A and B). Computer analysis using the search tool pSort revealed that



**Figure 3.1** The *hsp70* gene from the *Guillardia theta* nucleomorph. (A) Schematic of HSP70 showing the domain structure. (B) Nucleotide sequence of *hsp70*. The inferred amino acid sequence is shown beneath the nucleotide sequence in upper case letters. The conserved consensus patterns for the HSP70 family are underlined and the highly conserved motif at the carboxyl terminus is shown in bold and underlined. The bipartite nuclear localization signal is indicated by dotted underlining and the heatshock element upstream of the coding sequence is boxed.

the *G. theta* HSP70 possesses a putative bipartite nuclear localization signal similar to those described by others (Dingwall and Laskey 1991; Rensing and Maier 1994; Robbins *et al.* 1991). A perfect heat shock element (TTCnnGAAnnTTC) necessary for the binding of heat shock transcription factor is located at positions -41 to -29 relative to the initiator methionine (Figure 3.1B).

A phylogenetic analysis of HSP70 is presented in Figure 3.2. Most notably, the *G. theta* sequence clusters with cytosolic HSP70s to the exclusion of the ER homologs, and branches strongly with the nucleomorph sequence from another cryptomonad, *Rhodomonas salina*. Together, the nucleomorph sequences branch at the base of a clade containing the glaucocystophyte *Cyanophora paradoxa*, the green alga *Chlamydomonas reinhardtii*, and a variety of land plants (moderately supported with protML and distance methods, but not resolved with quartet puzzling). Given that the cryptomonad endosymbiont is thought to be of red algal origin (e.g., Cavalier-Smith *et al.* 1996; Douglas and Turner 1991), this result is particularly interesting in light of the debate over the origin(s) of primary plastids (see Palmer 2000 for review). While some data support the notion of a common origin of red algae and green plants (e.g., Burger *et al.* 1999), an analysis of the nuclear RNA polymerase II (*RPB1*) did not (Stiller and Hall 1997). More recent analyses by Moreira, Le Guyader and Phillippe (2000), including a concatenation of 13 nuclear genes, showed strong evidence for the monophyly of these two groups. Overall, the topology of the HSP70 tree presented here is very similar to that obtained in a recent comprehensive analysis by Germot and Phillippe (1999). Interestingly, the *G. theta* nuclear HSP70 branched near the base of the cytosolic HSP70 tree, and not near the algal/plant grouping, as was observed in previous analyses (Rensing *et al.* 1997). The two cryptomonad nucleomorph HSP70s are remarkably well conserved, and possess relatively

**Figure 3.2** Phylogenetic analysis of HSP70 protein sequences. The maximum likelihood (ML) tree ( $\ln L = -15,731.55$ ) from a heuristic search of 2000 trees in protML (Adachi and Hasegawa 1996) is shown; ML branch-lengths were inferred using a rate heterogeneity model (JTT-F +  $\Gamma$ ; see Materials and Methods) in PUZZLE (Strimmer and von Haeseler 1997). 526 unambiguously aligned amino acid positions were used. Cytosolic and endoplasmic reticulum (ER) HSP70s are represented, and the *Guillardia theta* nucleomorph and nuclear HSP70 sequences are highlighted. Support values above the branches are ML RELL values and are indicated if > 50%. Gray boxes contain support values for nodes of particular interest (FM, distance (Fitch-Margoliash) bootstrap values; QP, quartet puzzling support values, ML, ML RELL values). The scale bar indicates the inferred number of amino acid substitutions per site.

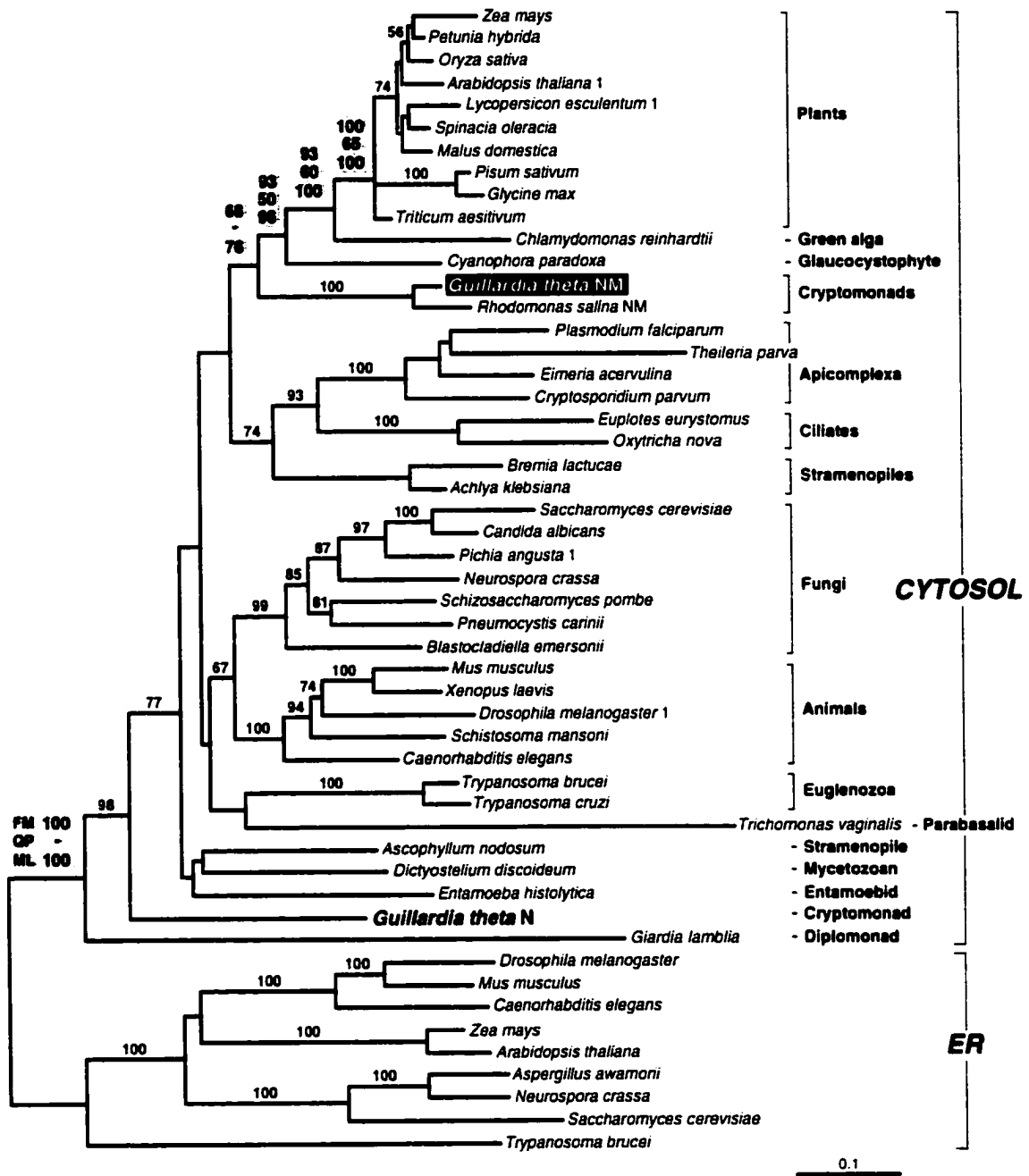


Figure 3.2 Phylogenetic analysis of HSP70 protein sequences

short branches compared to many of the other protist HSP70s (e.g., the apicomplexa, the ciliates, *Trichomonas vaginalis* and *Giardia lamblia*).

## HSP90

An *hsp90* (*hsp82*) gene was also present on chromosome 1 of the *G. theta* nucleomorph genome (Zauner *et al.* 2000). As was done for HSP70, HSP90 sequences were identified in the public databases by BLAST (Altschul *et al.* 1990), using the *G. theta* sequence as a query. Cytosolic and ER HSP90s were obtained for a wide diversity of bacteria, animals, fungi, plants and protists. For the red alga *Porphyra yezoensis*, 12 independent sequences encoding fragments of a cytosolic HSP90 were obtained from the EST database (Nikaido *et al.* 2000). These sequences were assembled using Sequencher, and two portions of nearly overlapping open reading frame were identified. A single EST covered a 500-nt portion of the 5' end of the gene (GenBank accession number AV432848), and the remaining 11 ESTs (Accession numbers AV431882, AV432126, AV433084, AV436237, AV429660, AV429526, AV429525, AV429680, AV436049, AV429756 and AV429533) were assembled to form a contig of approximately 1500 nucleotides in length (between one and five-fold coverage), encoding 504 amino acids of the carboxy-terminal half of the molecule. No frameshifts were detected in the coding region, and only those portions of the sequence unambiguously assigned as HSP90 coding sequence were used in phylogenetic analyses. A protein sequence alignment of 58 sequences, including bacterial HSP90s, was used in preliminary phylogenetic experiments, and a final set of 37 eukaryotic (cytosolic and ER) sequences was used for most analyses. The alignment contained 543 amino acid positions.

The *G. theta* nucleomorph-encoded HSP90 is 684 amino acids long (79.245 kDa). It contains the highly conserved motif MEAVD at its carboxy terminus

(Figure 3.3), and possesses an additional motif characteristic of HSP90s (YSNKEIFLRE) at position 23–32. It also has a nuclear localization signal (KKKKK) at position 232-236. Two near-perfect heat shock elements are located at positions –66 to –49 and –46 to –34 relative to the initiator methionine.

As was the case for HSP70, the *G. theta* nucleomorph HSP90 is very well conserved and is clearly cytosolic, clustering strongly with the other cytosolic homologs in phylogenetic analyses (Figure 3.4). As expected, the *G. theta* sequence forms a highly supported group with the *Porphyra yezoensis* HSP90 (assembled from multiple EST sequences; see above), consistent with previous accounts of a red algal ancestry for the cryptomonad nucleomorph genome (Cavalier-Smith *et al.* 1996; Douglas *et al.* 1991). Together, these sequences form a weakly supported clade with HSP90s from a variety of land plants. Again, this result is consistent with recent data suggesting the common ancestry of red algae and green plants (Moreira, Le Guyader and Phillippe 2000). Most surprising was the observation that the sisterhood of animals and fungi, supported by protein and SSUrRNA phylogenies, as well as a 12 amino acid insertion in EF1- $\alpha$  (see Baldauf 1999 for recent review), was not recovered. Few phylogenetic studies have been performed on HSP90, largely due to limited taxon sampling. Gupta (1995), using parsimony and distance-based methods of tree reconstruction, first noted that this molecule failed to show a relationship between animals and fungi, and instead showed a weak relationship between animals and plants. My analyses also failed to resolve an animal-fungal connection, despite the use of maximum likelihood methods (including taking into account among-site rate variation), a more conservative alignment (543 unambiguously aligned amino acid positions), and increased taxon sampling. To investigate the possibility that the fungal sequences were branching at the base of the cytosolic HSP90 tree artifactually due to 'long branch attraction', I performed phylogenetic analyses with and





**Figure 3.4** Phylogenetic analysis of HSP90 protein sequences. The tree shown ( $\ln L = -15,585.36$ ) is the topology of the best maximum likelihood (ML) tree from a heuristic search of 1000 trees in protML (Adachi and Hasegawa, 1996), using 543 unambiguously aligned amino acid sites. ML branch-lengths were inferred under a rate heterogeneity model (JTT-F +  $\Gamma$  + inv; see Materials and Methods) in PUZZLE (Strimmer and von Haeseler, 1997). Cytosolic and endoplasmic reticulum (ER) forms of HSP90 are indicated, and the *Guillardia theta* nucleomorph sequence (highlighted) strongly clusters with the cytosolic homologs. Support values above the branches are ML RELL values, and are given if > 50%. Gray boxes contain support values for nodes of particular interest (FM, distance (Fitch-Margoliash) bootstrap values; QP, quartet puzzling support values, ML, ML RELL values). The scale bar indicates the inferred number of substitutions per amino acid site.

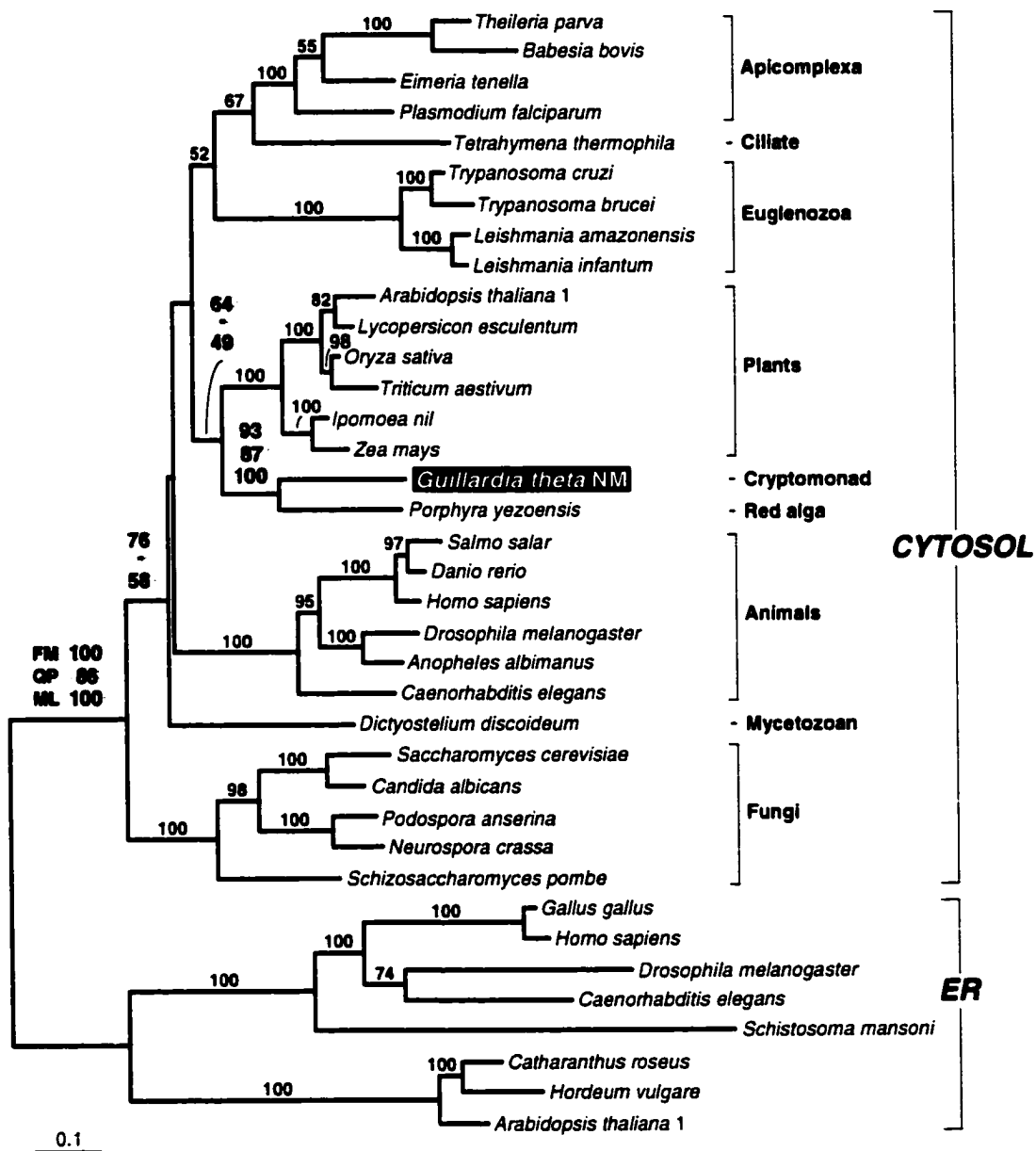


Figure 3.4 Phylogenetic analysis of HSP90 protein sequences

without a bacterial outgroup, and with the selective removal of various long-branch taxa (e.g., divergent fungi). However, the basal position of the fungi was a consistent pattern in all analyses (data not shown). It does seem significant that the animal HSP90s also branch near the base of the cytosolic tree (Figure 3.4): the apparent paraphyly of the animals and fungi in HSP90 phylogenies may therefore be due to the misplacement of the root.

### Heat shock transcription factor (HSF)

A gene encoding a heat shock transcription factor was found on chromosome 1 of the *G. theta* nucleomorph genome. Using this sequence as a probe, I searched the public databases for additional HSFs from diverse eukaryotes. A large set of animal, fungal and plant HSF-like sequences were obtained (60 sequences), many of which were duplicates or isoforms encoded in the same genome. HSF sequences were aligned using CLUSTAL W (Higgins and Sharp 1988) and the alignment was adjusted manually.

The *G. theta* nucleomorph HSF is, to my knowledge, the first protist sequence to be described. Most noticeably, the predicted polypeptide is a mere 185 amino acids long (22.260 kDa), much shorter than typical eukaryotic HSFs which are between 400 and 800 amino acids in length (Figure 3.5A). Whether this drastically reduced size is a unique feature of the *G. theta* HSF, or is characteristic of protist HSFs as a whole, is currently unknown. No transactivation domain, typically found in the carboxy-terminal half of eukaryotic HSFs, is detectable in the HSF open reading frame or anywhere else in the completely sequenced nucleomorph genome (Douglas *et al.* 2001). Two of the three consensus patterns for HSF-type DNA-binding domain proteins ('HSF1' and 'HSF2') are well-conserved (Figure 3.5C); the 'HSF3' consensus pattern within the oligomerization domain (Figure 3.5B) shows only limited sequence identity although the leucine

**Figure 3.5** The heat shock transcription factor gene from the *Guillardia theta* nucleomorph. (A) Schematic showing the domain structure of the 185-amino-acid *Guillardia theta* nucleomorph heat shock transcription factor (HSF) protein, compared to select animal, fungal and plant HSFs. The DNA-binding and oligomerization domains are highlighted. (B) The oligomerization domain of HSF showing leucine zipper motif (shaded vertical bars) and 'HSF3' consensus pattern. (C) Protein sequence alignment of the DNA-binding domain of select HSF sequences showing 'HSF1' and 'HSF2' consensus patterns and putative nuclear localization signal (boxed). Dashes indicate gaps in the alignment and dots indicate an identical amino acid residue to that in the reference sequence (*Homo* HSF1).

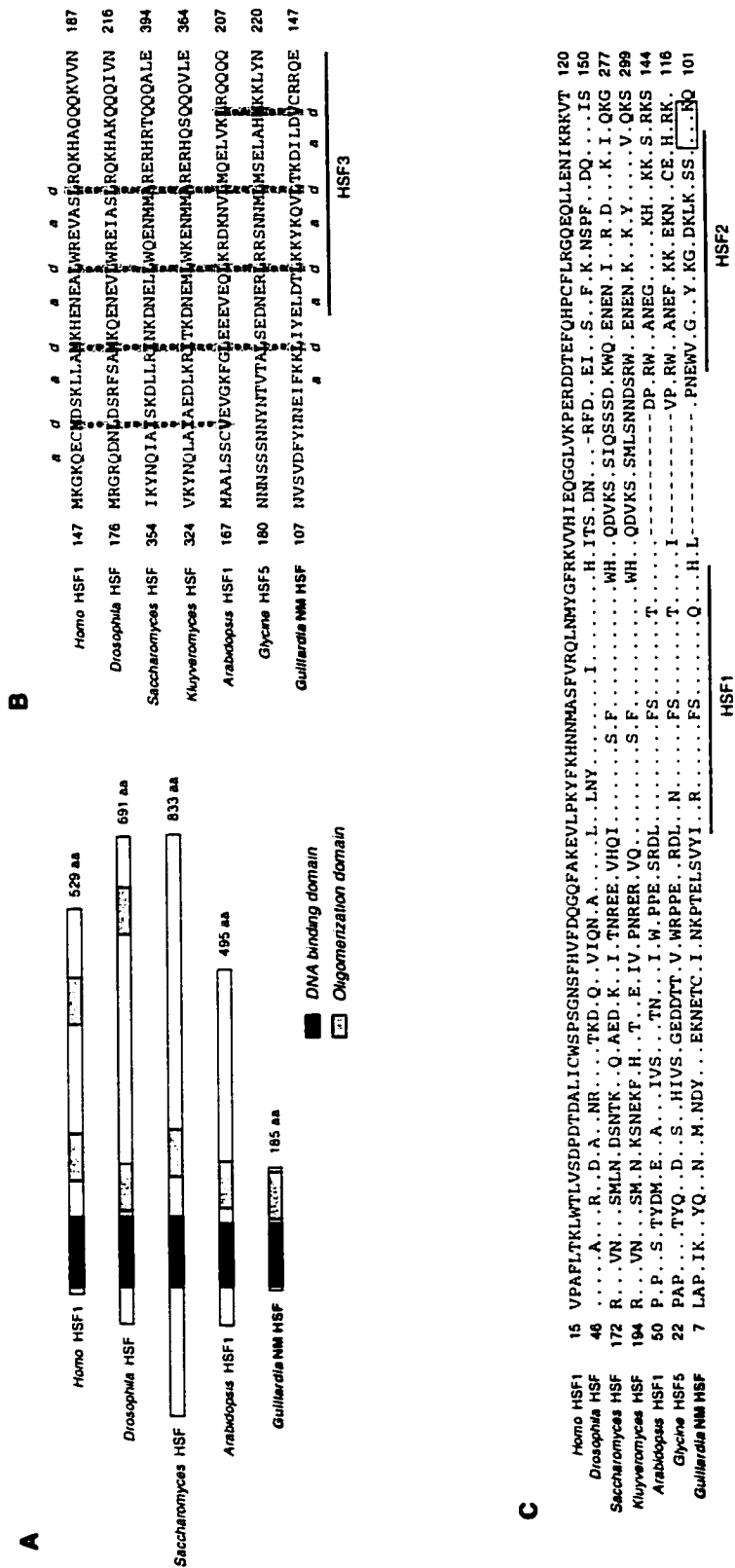


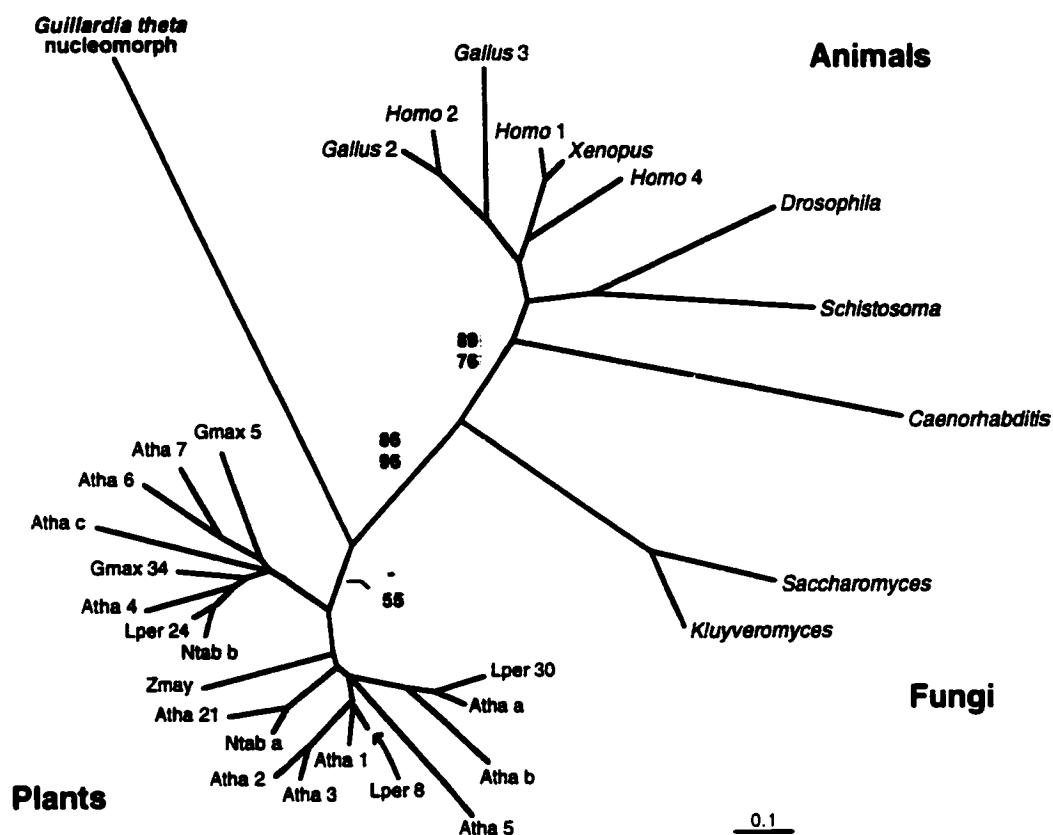
Figure 3.5 The heat shock transcription factor gene from the *Guillardia theta* nucleomorph

zipper motif is preserved. A putative nuclear localization signal (KRKK) is present at the carboxyl-terminal end of the DNA-binding domain.

A phylogenetic analysis of the DNA-binding domain of selected HSFs (Figure 3.6) reveals rampant gene duplication in the evolutionary history of this protein family. Due to its small size (approximately 100 amino acids), the DNA-binding domain contains little phylogenetic signal. However, the analyses do show a well supported separation of the animal and fungal HSFs from the plants and the single protist sequence (86 and 96%, quartet puzzling and ML-distance bootstrap, respectively), consistent with an insertion/deletion present in the DNA-binding domain of the animal and fungal sequences (Figure 3.5C). The analyses suggest that an early gene duplication occurred in land plant evolution, as two distinct, but weakly supported clusters of HSF sequences are present (Figure 3.6). More recent duplications also appear to have occurred, as more than a dozen distinct HSF-like sequences were identified in *Arabidopsis thaliana* alone, and multiple paralogs were present in a variety of other plants (data not shown). Duplications have also happened in animal evolution. The single *G. theta* HSF sequence still retains several conserved motifs characteristic of other DNA-binding proteins (see above), but was a fairly long branch compared to other HSFs in phylogenetic analyses (Figure 3.6).

### The chaperonin CCT

Genes encoding the  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ ,  $\eta$  and  $\theta$  subunits of the eukaryotic chaperonin CCT were found distributed among all three chromosomes of the *G. theta* nucleomorph genome (Douglas *et al.* 2001). Unlike the *G. theta* HSP70 and HSP90 genes, none of the CCTs contained recognizable heat shock promoter elements. The amino acid sequences inferred from the CCT genes were readily alignable with other eukaryotic homologs, but were extremely divergent in



**Figure 3.6.** Phylogenetic analysis of heat shock transcription factor protein sequences. The analysis was performed using an alignment of 31 heat shock transcription factors (HSFs) and 89 unambiguously aligned amino acid positions in the DNA-binding domain. The tree shown was inferred with a distance (Fitch-Margoliash) algorithm from a maximum likelihood distance matrix calculated with a rate heterogeneity model (JTT-F +  $\Gamma$  +inv; see Materials and Methods). Statistical support for the major branches is shown (quartet puzzling support values (top) and ML-distance bootstrap values (bottom)). The *Guillardia theta* nucleomorph HSF sequence is in bold face. The scale bar represents the inferred number of amino acid substitutions per site. Abbreviations for plant HSFs: Gmax, *Glycine max*; Atha, *Arabidopsis thaliana*; Lper; *Lycopersicon peruvianum*; Zmay, *Zea mays*; Ntab, *Nicotiana tabacum*. GenBank accession numbers for plant HSFs: Gmax 5, S59539; Gmax 34, S59538; Atha 1, g7428781; Atha 2, CAB63800; Atha 3, CAA74397; Atha 4, CAB16764; Atha 5, AAF16564; Atha 6, CAB63802; Atha 7, CAB63803; Atha 21, AAC31792; Atha a, AAC31222; Atha b, CAB41311; Atha c, AAB84350; Lper 8, P41153; Lper 24, P22335; Lper 30, P41152; Zmay, S57633; Ntab a, BAA83711; Ntab b, BAA83710.

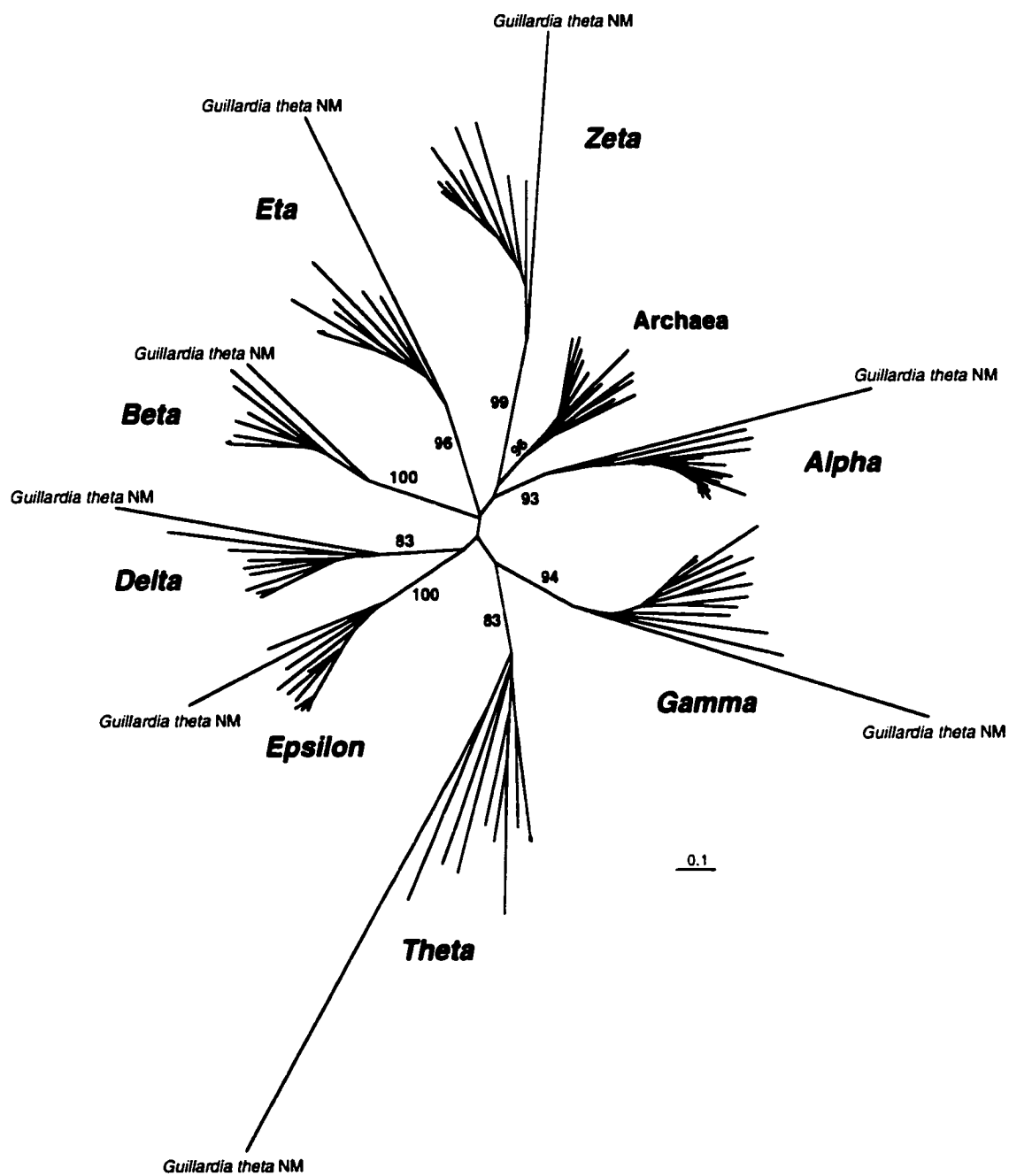
sequence (see below). As well, most of the subunits were shorter than other CCTs at their amino and carboxy termini.

A phylogenetic analysis of the CCT subunits and their archaeal homologs is presented in Figure 3.7, and clearly illustrates the remarkable divergence of the *G. theta* nucleomorph sequences relative to those of other eukaryotes. All eight sequences branch with their respective orthologous groups with high statistical support and with all methods (data not shown), but are characterized by extremely long branch lengths. To investigate this further, I examined the amino acid composition of the *G. theta* CCTs and found that, compared to CCT sequences from a diversity of other eukaryotes, the nucleomorph sequences possessed highly biased amino acid compositions. I used PUZZLE 4.0 (Strimmer and von Haeseler 1997) to perform Chi-square tests for amino acid composition bias on each of the eight CCT subunits and found that the *G. theta* sequence failed in each case (data not shown). Notably, the sequences appeared to be biased towards asparagine residues in cases where a unique substitution had occurred in the *G. theta* sequence, likely reflecting the extreme A-T bias of the nucleomorph genome. Analysis of codon usage of nucleomorph genes as whole indicated a bias of greater than 80% towards codons ending in A or T (S. Douglas, pers. comm.). Figures 3.8-3.11 show the *G. theta* CCT $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$  sequences aligned with a representative sample of their respective orthologs (alignments of the  $\epsilon$ ,  $\zeta$ ,  $\eta$  and  $\theta$  subunits appear in Appendix B).

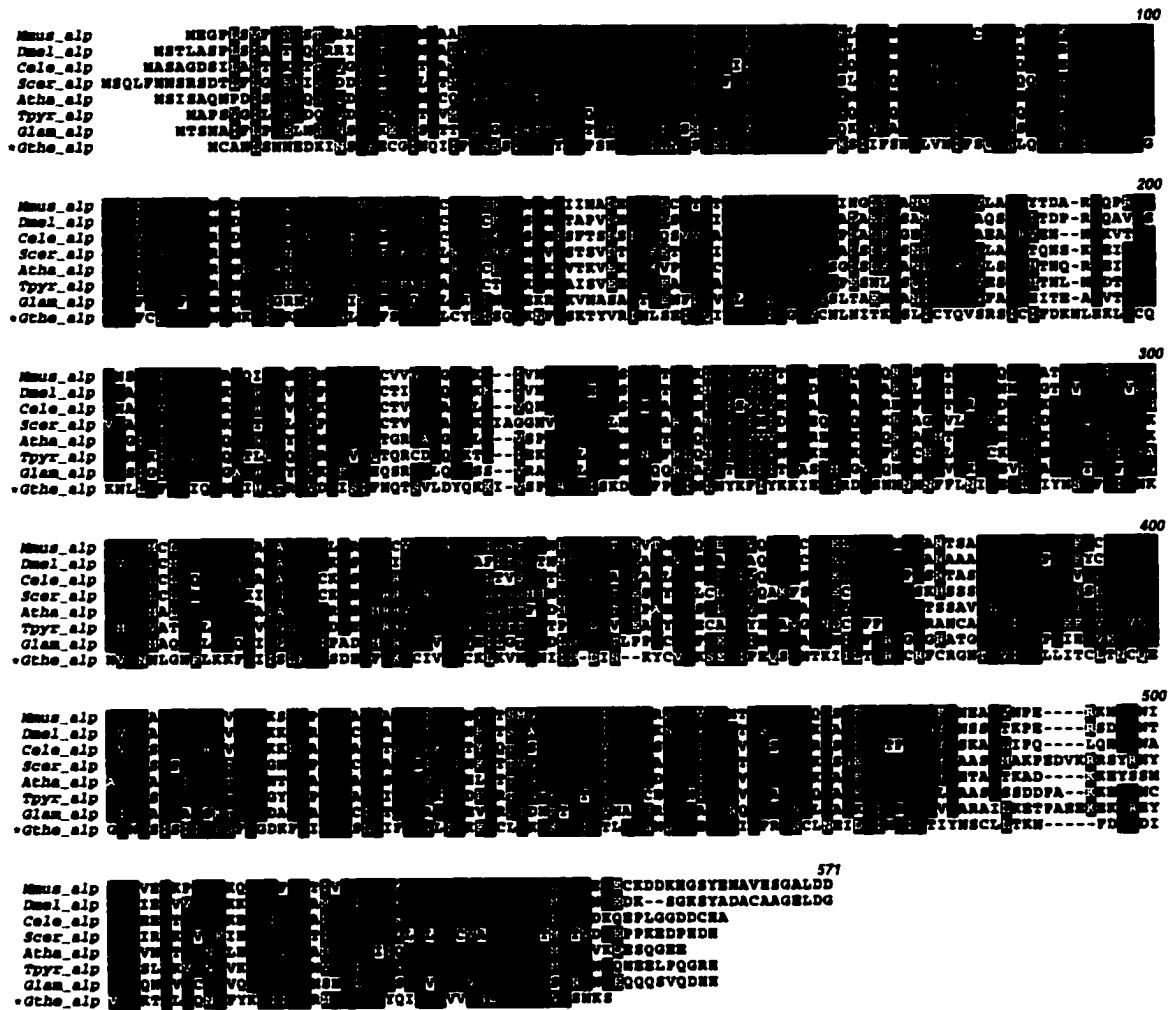
While significant differences have been observed in the rates of evolution among the different CCT subunits themselves (see Chapter II; Archibald, Logsdon and Doolittle 2000), no particular CCT subunit in *G. theta* appeared noticeably more divergent than the others (Figure 3.7). This is in contrast to the pattern observed for the *G. theta* tubulins, where gamma-tubulin was found to be



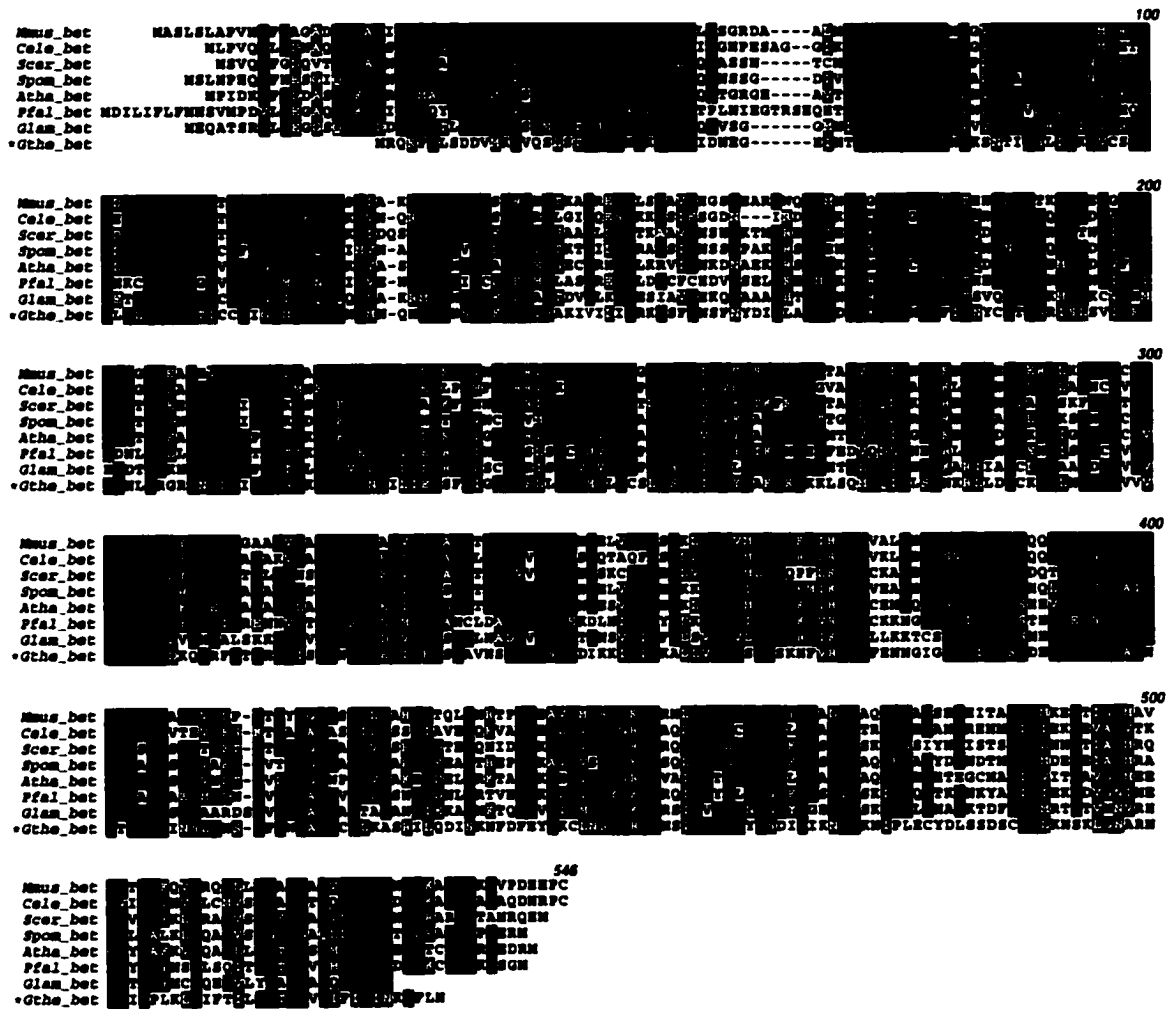
**Figure 3.7** Phylogenetic analysis of archaeal and eukaryotic chaperonin protein sequences. The tree shown is a comprehensive neighbor-joining (NJ) distance tree of 103 group II chaperonin sequences, constructed from an alignment of 251 unambiguously aligned amino acid residues. Distance (NJ) bootstrap support for the monophyly of the archaeal chaperonins and the different CCT subunits are given (calculated from 100 resampling replicates). The archaeal chaperonins and the eight different CCT subunits (paralogs) are highlighted, as are the *Guillardia theta* nucleomorph CCT subunit sequences. The scale bar indicates the inferred number of substitutions per amino acid site.



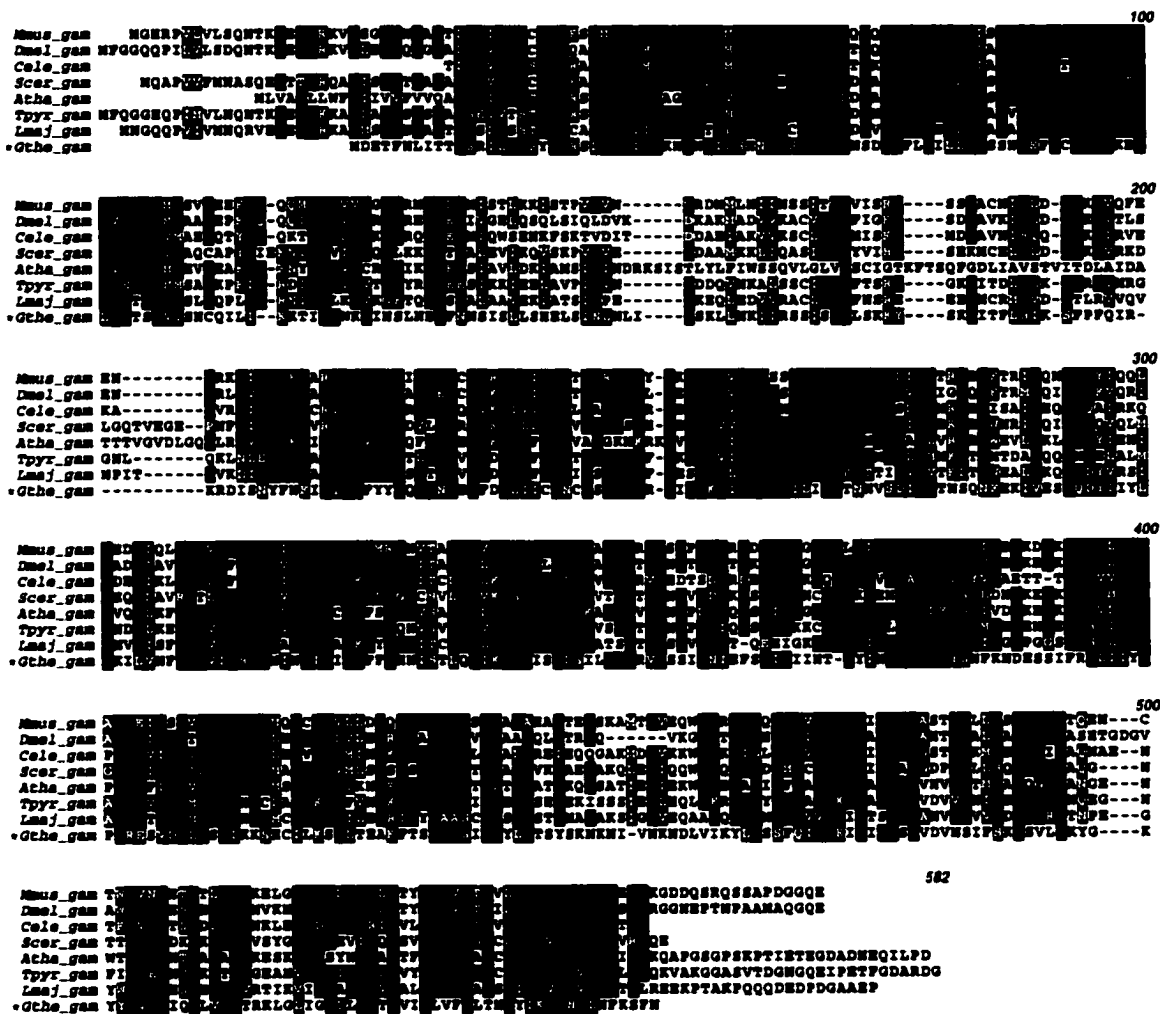
**Figure 3.7** Phylogenetic analysis of archaeal and eukaryotic chaperonin protein sequences



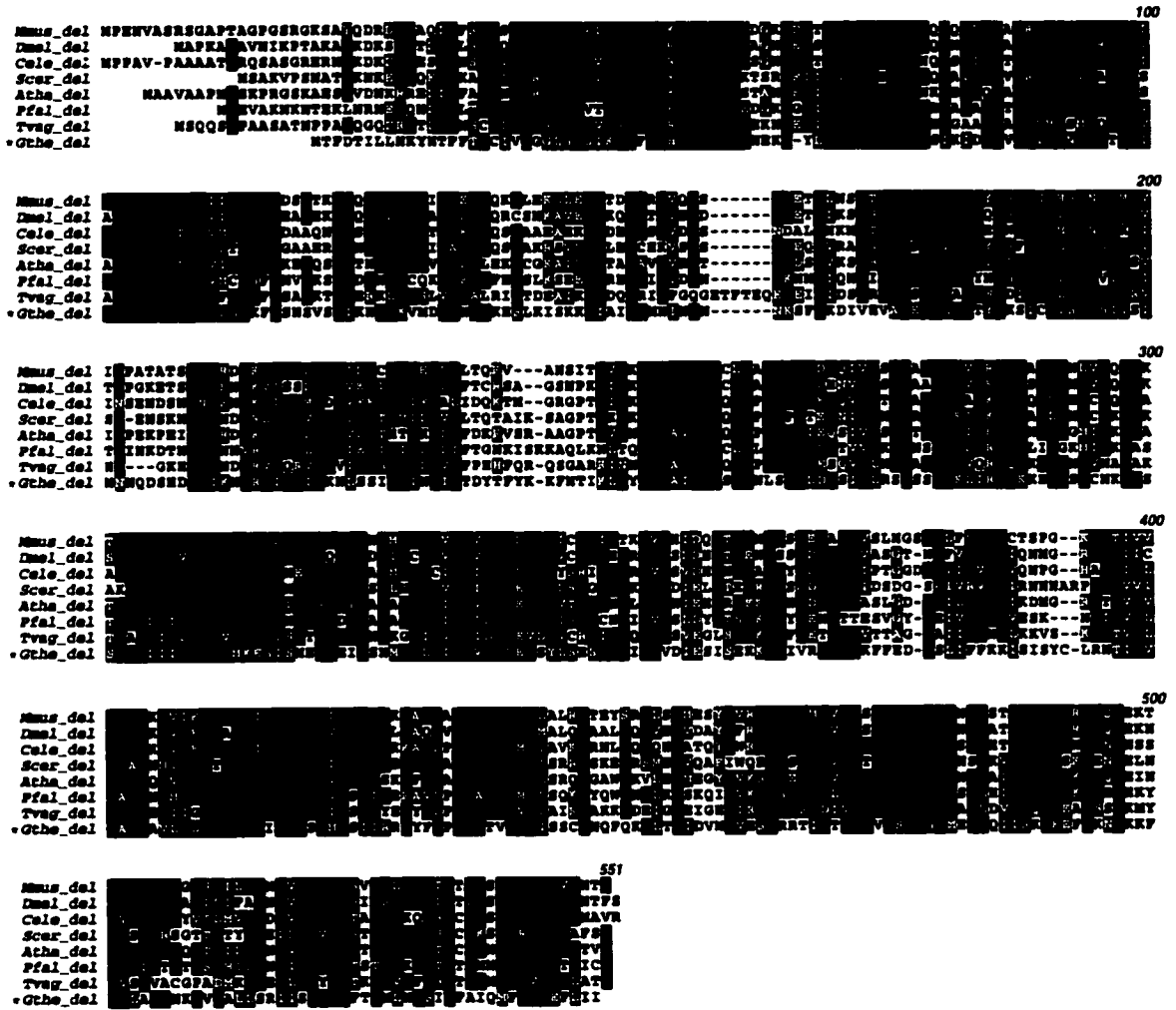
**Figure 3.8** Alignment of select CCT $\alpha$  protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Tpyr*, *Tetrahymena pyriformis*; *Glam*, *Giardia lamblia*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *alp*, CCT $\alpha$ .



**Figure 3.9** Alignment of select CCTbeta protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Spom*, *Schizosaccharomyces pombe*; *Atha*, *Arabidopsis thaliana*; *Pfal*, *Plasmodium falciparum*; *Glam*, *Giardia lamblia*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *bet*, CCTbeta.



**Figure 3.10** Alignment of select CCTgamma protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Tpyr*, *Tetrahymena pyriformis*; *Lmaj*, *Leishmania major*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *gam*, CCTgamma.



**Figure 3.11** Alignment of select CCTdelta protein sequences. The *Guillardia theta* nucleomorph sequence is highlighted by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Pfal*, *Plasmodium falciparum*; *Tvag*, *Trichomonas vaginalis*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *del*, CCTdelta.

extraordinarily divergent relative to the moderately divergent alpha- and beta-tubulins (Keeling *et al.* 1999).

## DISCUSSION

The presence or absence of functionally well-characterized proteins in the periplastid space of cryptomonads may suggest which features it might still share with a 'typical' eukaryotic cytosol, and which features have been severely reduced, modified or lost entirely. As was the case for the *G. theta* tubulin genes (Keeling *et al.* 1999), the nucleomorph-encoded molecular chaperones described here have important implications for our current understanding of the cell biology of the cryptomonad periplastid complex.

The nucleomorph genome of *G. theta* encodes several of the major families of molecular chaperones necessary for protein folding, transport and degradation. Consistent with its dramatically reduced size, single intronless genes are present for HSP70, HSP90, the heat shock transcription factor HSF, and each of the CCT subunits. This is in contrast to the situation in vertebrates and plants, where multiple genes for heat shock proteins and their transcription factors exist. Heat shock elements are found upstream of both heat shock protein genes (but in none of the CCT subunit genes), suggesting that this pathway of gene regulation is present. Although both heat shock proteins are similar in size to those found in other eukaryotes, the predicted HSF polypeptide is only about half the size of HSFs from other eukaryotes, and the CCT subunits have slightly reduced amino- and carboxy-termini. Apparently the nucleomorph genome has dispensed with everything but the bare essentials—active sites and functional domains—but retained some of the same regulatory mechanisms.

The activation of heat shock genes is known to be rapid yet transient and is mediated by HSF in response to elevated temperatures and chemical or physiological stress. Although shorter than homologs from other eukaryotes, the *G. theta* HSF possesses both a DNA-binding domain and an oligomerization domain containing a leucine zipper motif (Figure 3.5), suggesting that the homotrimerization required for binding to heat shock elements in the DNA can occur. The nuclear-localization signal (KRKK) suggests that HSF is active in a nuclear environment, in this case most likely the nucleomorph. Under physiological conditions and during recovery from stress, HSP70 has been shown to negatively regulate HSF transcriptional activity by binding to the HSF transactivation domain (Shi, Mosser and Morimoto 1998). This HSP70/HSF complex may then interact with HSP90 (or the HSP90 multichaperone complex) resulting in the inability of HSF to form trimers and activate transcription (Zuo *et al.* 1998). Under conditions of stress, the majority of cellular chaperone is bound to denatured protein, leaving less to bind to and repress HSF, thus allowing transcription of HSP genes to be stimulated. It is also thought that intramolecular interaction between HSF amino- and carboxy-terminal coiled coil domains keeps the protein in an inactive state (Zuo *et al.* 1994). Interestingly, no recognizable transactivation domain or carboxy-terminal coiled coil domain is present in the nucleomorph HSF, suggesting that it is regulated by a different mechanism. Without the carboxy-terminal domain, the *G. theta* HSF may be permanently active at some level.

The nucleomorph-encoded HSP70 and HSP90 proteins from *G. theta* possess the hallmarks of typical cytosolic heat shock proteins. HSP70 proteins typically consist of a highly conserved amino-terminal ATPase domain of 44-kDa and a 25-kDa carboxy-terminal domain that contains a conserved substrate-binding domain of 15-kDa and a less-conserved 10-kDa domain (Figure 3.1A).



The ATPase activity of HSP70 is stimulated by HSP40—interestingly, no *hsp40* gene could be found on the nucleomorph chromosomes, suggesting that this component might be imported from the host compartment. The HSP70/HSP40/substrate complex is thought to be stabilized by the binding of another protein co-factor, Hip, to the ATPase domain (for review see Frydman and Höhfeld 1997). The carboxy terminus of HSP70 contains eight residues that are highly conserved among all eukaryotes (GPTIEEVD; GPKIEEVD in *G. theta*; Figure 3.1) and are responsible for interacting with an adaptor protein, Hop, that facilitates the assembly of the Hsp70-Hsp90 multichaperone (Scheufler *et al.* 2000). HSP70 has also been found in the centrosome in animals as well as in the cytoplasm and the nucleus, where it binds HSF; the presence of a nuclear localization signal in the *G. theta* HSP70 suggests that it has the potential to localize to the nucleomorph.

HSP90 is known to interact with HSP70 *via* an adapter that binds the C-terminal sequence MEEVD (Scheufler *et al.* 2000); this sequence is MEAVD in the *G. theta* HSP90 (Figure 3.3). In animals, HSP90 is a conserved core centrosomal component, and in yeast it is known to play a role in spindle pole duplication. Although a spindle apparatus and microtubules have not been seen in dividing cryptomonad nucleomorphs (McKerracher and Gibbs 1982; Meyer 1987; Morrall and Greenwood 1982), they could easily have been overlooked because of low contrast in the rather dense nucleomorph and their probably exceedingly transient duration in such a small nucleus: only a few very short microtubules would be needed for a fully functional spindle. It is thus possible that the nucleomorph-encoded HSP90, HSP70, gamma-tubulin and ranbpm proteins interact to form a simplified centrosome and nucleate the assembly of alpha- and beta-tubulin to form a minute spindle responsible for segregating the nucleomorph chromosomes; this could be tested by immunocytochemistry.

Because nucleomorphs divide like most fungal nuclei, by a closed division with no nuclear-envelope breakdown (McKerracher and Gibbs 1982; Meyer 1987; Morrall and Greenwood 1982), the putative spindle will probably be intranuclear as in their red algal ancestors (Schornstein and Scott 1982). The presence of a nuclear-localization sequence on Hsp90 raises the possibility that the centrosome may lie within the nucleomorph envelope, not outside it opposite a polar fenestra as in red algae (Schornstein and Scott 1982). This would be consistent with the apparent absence of a polar fenestra during nucleomorph division (McKerracher and Gibbs 1982; Meyer 1987; Morrall and Greenwood 1982), and its non-detection by electron microscopy in the periplastid space.

I have demonstrated drastic differences in the rates of protein sequence evolution among the various molecular chaperones encoded in the *G. theta* nucleomorph genome. While all of the eight *G. theta* nucleomorph CCT sequences fall into the previously described CCT subunit families with high statistical support, they are characterized by extremely long branches in phylogenetic analyses and by highly biased amino acid compositions (see Results). In stark contrast, the nucleomorph HSP70 and HSP90 sequences are remarkably well conserved. The phylogenies presented in Figures 3.2 and 3.4 show that these sequences have very short branch lengths compared to other cytosolic homologs. In fact, the *G. theta* HSP70 nucleomorph sequence has a shorter branch than the HSP70 encoded in the host nuclear genome. In contrast to the eight CCT subunits, neither HSP70 or HSP90 show biased amino acid compositions in Chi-square tests compared to their respective cytosolic homologs.

Such radical differences in the degree of conservation of HSP70 and HSP90 on one hand, and the eight CCT subunits on the other, presumably reflect differing degrees of functional constraint. However, in the absence of detailed

information on the exact functions of these chaperones in the periplastid space, it is difficult to know what the reasons for such differences might be. It does seem significant that HSP70 is known to be extensively involved in protein translocation (in addition to general protein folding), a process that should be evolutionarily conserved in highly membranous cells such as cryptomonad algae. In contrast, the chaperonin CCT appears to be exclusively involved in the folding of newly-translated proteins (Willison and Horwich 1996); a reduction in the number of CCT substrates present and functioning in the periplastid space could explain the remarkable divergence of the CCT subunits. Extreme divergence in the evolution of the CCT substrates themselves (due to decreased/different functional constraints) could also influence the evolution of the CCTs—in effect, the co-evolution of chaperonin and substrate. It is interesting that one confirmed set of CCT substrates likely to function in the periplastid space of *G. theta*, the tubulins, are also divergent to varying degrees (Keeling *et al.* 1999). A homolog of the cytoskeletal element actin, the other well-characterized substrate of CCT, is not present in the *G. theta* nucleomorph genome. It remains to be seen whether this cytoskeletal component is imported or has been lost.

The complete lack of co-chaperones encoded in the nucleomorph is quite unexpected—if gene transfer from the nucleomorph to the host genome is more or less a random process, one would not expect all of the major chaperones to be nucleomorph-encoded, and all co-chaperones to be in the host nucleus. It is thus likely that some of the co-chaperones and interacting factors have been lost altogether during the extreme reduction of the nucleomorph genome. In the case of CCT, the loss of co-chaperonin(s) (if possible) would surely influence the evolutionary constraints on the various functional regions of the molecules. A fuller understanding of the reasons for such a drastic increase in evolutionary rates awaits a more detailed knowledge of the full range of *in vivo* CCT

substrates in a 'typical' cell cytosol and the ability to compare these substrates to those still functioning in the periplastid space itself. Data on which co-chaperones are truly nucleus-encoded and functioning in the periplastid space, and which ones have been lost, will be essential. The cryptomonad endosymbiont may ultimately tell us about the minimal chaperone system necessary for maintaining basic eukaryotic cellular processes.

## CHAPTER IV

### The Chaperonin Genes of 'Jakobid' Flagellates: Implications for Early Eukaryotic Evolution

New sequences have been deposited in Genbank under accession numbers AF322043-AF322050.

#### INTRODUCTION

The 'jakobid' flagellates are free-living, mitochondriate, heterotrophic protists (O'Kelly 1993; O'Kelly 1997; O'Kelly and Nerad 1999). They include the families Jakobidae, genus *Jakoba* (Patterson 1990), Histionidae, genera *Histiona* (Voight 1901) and *Reclinomonas* (Flavin and Nerad 1993), and Malawimonadidae, genus *Malawimonas* (O'Kelly and Nerad 1999). The jakobids have figured prominently in hypotheses about the origin and early evolution of eukaryotes, originally due to the presence of ultrastructural features shared with certain amitochondriate lineages (retortamonads and, indirectly, diplomonads; O'Kelly 1993). Small subunit ribosomal RNA (SSUrRNA) and protein phylogenies often place the amitochondriates at or near the base of the eukaryotic tree (e.g., Hashimoto *et al.* 1994; Leipe *et al.* 1993; Sogin *et al.* 1989; Stiller, Duffield and Hall 1998). O'Kelly (1993) hypothesized that if amitochondriate protists were the earliest diverging eukaryotic lineages, then the earliest diverging *mitochondriate* eukaryotes should share ultrastructural features with these groups.

The sequencing of jakobid mitochondrial genomes has provided startling insight into mitochondrial evolution and has further suggested a pivotal role for the jakobids in our understanding of the evolution of eukaryotes. These organisms possess the most bacterial-like mitochondrial genomes characterized

thus far (Gray, Burger and Lang 1999; Gray *et al.* 1998; Lang *et al.* 1997; Palmer 1997). Most striking is the mitochondrial DNA (mtDNA) of *Reclinomonas americana*, which encodes 97 genes, more than is present in any other mitochondrial genome. Several of these genes have never before been found encoded in mtDNA, including a gene for a bacterial translation factor (*tufA*), a putative cytochrome oxidase assembly protein (*cox11*), a secretion pathway protein (*secY*), and genes for four subunits of a bacterial-type RNA polymerase (*rpoA-D*; single-subunit phage-type RNA polymerases are thought to function in all other known mitochondria) (Gray *et al.* 1998; Lang *et al.* 1997). Operon-like ribosomal protein gene clusters similar to those found in bacteria are also present. Based on the retention of ancestral features in their mitochondrial genomes, the jakobids could be among the earliest-diverging eukaryotic lineages.

More recently, the jakobids have been suggested to be members of a much larger assemblage of protists, the 'excavate taxa'. The excavates are a diverse group of amitochondriate, mitochondriate, and hydrogenosomal lineages that share as their uniting feature the presence of a ventral feeding groove (Patterson, Simpson and Weerakoon 1999; Simpson 1999). In addition to the 'core jakobids' (*Jakoba*, *Reclinomonas* and *Histiona*) and *Malawimonas*, the excavates include the heteroloboseans, diplomonads, retortamonads, *Trimastix* and *Carpediemonas* (Simpson 1999). While ultrastructural data suggest that these organisms may share a common 'excavate' (i.e., feeding groove-bearing) ancestor, there is no consensus view on the relationships amongst the various excavate taxa, or their relationship to other mitochondriate, amitochondriate, and hydrogenosome-containing groups. Currently, views on the origin and evolution of eukaryotes are in a state of flux. Many of the putatively deep-branching and primitively amitochondriate protist lineages (including some of the excavates) are now thought to be derived from mitochondrion-bearing ancestors (see Roger 1999 for

recent review). Further, the ability of current phylogenetic methods to accurately reconstruct the deepest branches of phylogenetic trees has come into question (Hirt *et al.* 1999; Philippe and Germot 2000; Stiller and Hall 1999, and references therein). Philippe and Adoutte (1998) suggested that a 'big bang' occurred at the base of eukaryotes, and that the major cladogenetic events in eukaryotic evolution occurred in quick succession. Thus there is at present no clear picture as to which protist groups—if any—actually represent early diverging lineages.

The cytosolic chaperonin CCT has the potential to be a useful molecule for eukaryotic phylogeny. Individual CCT subunits are over 500 amino acids in length and are highly conserved, both desirable properties for a phylogenetic marker. The presence of eight distinct CCT paralogs also makes it possible to perform multiple reciprocal rootings of the phylogenetic tree of eukaryotes. Unfortunately, the CCTs are relatively poorly sampled compared to more commonly used phylogenetic markers such as SSUrRNA and tubulin. Focussing on the  $\alpha$  subunit of CCT, I therefore sought to broaden the diversity of taxonomic representation, and add to the *Trichomonas vaginalis* and *Giardia lamblia* CCTs described in Chapter II. I isolated CCT $\alpha$  genes from the jakobids *Reclinomonas americana* (strains 50394 and 50283) and *Malawimonas jakobiformis*, as well as a CCT $\delta$  gene from *M. jakobiformis*. I also amplified CCT $\alpha$  from two heteroloboseans, *Naegleria gruberi* and *Acrasis rosea*, from the euglenozoan *Trypanosoma brucei*, and from the parabasalid *Monocercomonas* sp.

Surprisingly, and unlike most protein-coding genes in protists, the jakobid CCT genes possessed numerous spliceosomal introns—between five and seven in ~ 1-1.4 kb of coding sequence. Rooted and unrooted phylogenetic analyses of CCT $\alpha$  recover many of the eukaryotic groups resolved with other phylogenetic markers, and suggest that *R. americana* (but not the other jakobid, *M. jakobiformis*) is distantly related to the Heterolobosea and the Euglenozoa. Most surprising

was the fact that *Giardia lamblia* is not early-diverging in CCT $\alpha$  phylogenies, but instead forms a weak but consistent clade with *T. brucei*, *R. americana*, and the heteroloboseans. The high density of introns present in the CCT $\alpha$  and CCT $\delta$  genes of the jakobid flagellates, and the presence of intron positions shared between jakobids, animals, fungi and plants, suggests that many of the intron-sparse/-lacking protist lineages have lost spliceosomal introns throughout their evolutionary history.

## RESULTS

### CCT sequences

A degenerate PCR approach was used to isolate CCT genes from protist gDNAs. Near full-length CCT $\alpha$  genes were amplified from two strains of the jakobid *Reclinomonas americana* (ATCC 50394 and 50283) using the CCT-9-for/CCT-11-rev primer pair. For *Malawimonas jakobiformis* (ATCC 50310), CCT $\alpha$  was obtained with the primer pair CCT-2-for/CCT-11-rev. A somewhat smaller fragment of the CCT $\delta$  gene was also isolated from *M. jakobiformis* using the CCT-9-for/CCT-7-rev primer set. Repeated attempts to amplify CCT $\alpha$  from *Jakoba libera* (another jakobid) were unsuccessful, despite the use of all possible primer combinations under a wide range of conditions. PCR reactions using additional degenerate primers designed with a strong bias towards G+C at the 3<sup>rd</sup> codon position (see below) also failed.

The CCT $\alpha$  gene was also isolated from two heteroloboseans (*Acrasis rosea* and *Naegleria gruberi*) with the CCT-9-for/TF-9-rev primer set. For the parabasalid *Monocercomonas* sp., CCT $\alpha$  was obtained using the forward primer CCT-9-for in combination with two reverse primers, CCT-4-rev and TF-9-rev. 90-95% of the CCT $\alpha$  gene was amplified with most primer combinations.



For *Trypanosoma brucei*, a small fragment of the CCT $\alpha$  gene was found by searching genomic data from unfinished microbial genomes (<http://www.tigr.org>). Based on this sequence an exact-match reverse primer (Tbru.CCTa.R) was designed and used in combination with two degenerate forward primers (CCT-2-for and CCT-9-for) to amplify most of the CCT $\alpha$  coding sequence from *T. brucei* gDNA. From the sequence of five independent clones, two different CCT $\alpha$  sequences were apparent, one of which matched the original fragment present in the *T. brucei* genome data. The ambiguities between the two clone types were almost all synonymous substitutions, suggesting the presence of multiple copies of the CCT $\alpha$  gene in *T. brucei*.

CCT sequences from *Plasmodium falciparum* were identified by BLAST (Altschul *et al.* 1997) at the *PlasmoDB* website (<http://www.plasmodb.org/>). CCT $\alpha$  was obtained from chromosome 11 sequence, while CCT $\delta$  was found in genomic sequence from chromosome 13.

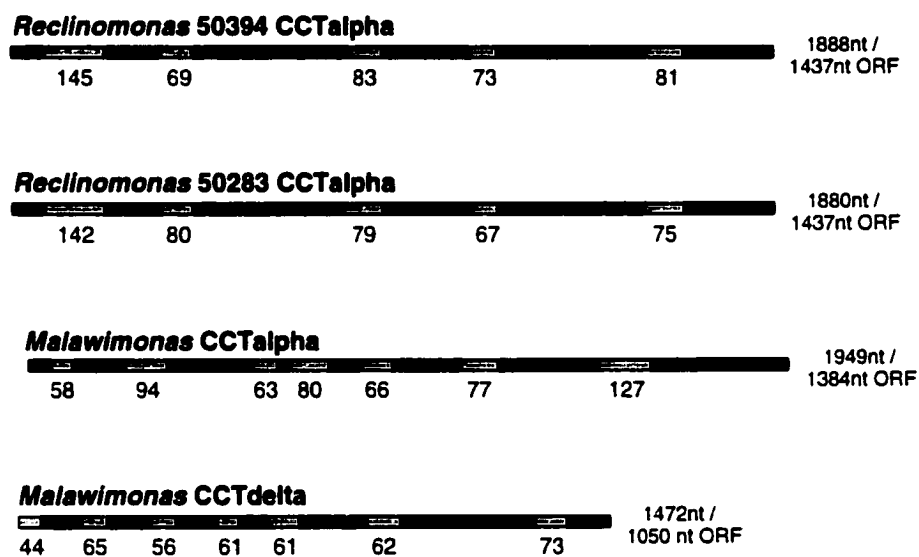
### **Jakobid CCT genes possess numerous spliceosomal introns**

The jakobid CCT genes described here are among the first nuclear protein-coding genes to be characterized from these protists. The most striking feature of the *Reclinomonas americana* and *Malawimonas jakobiformis* CCTs is the presence of multiple spliceosomal introns. The CCT $\alpha$  gene from *R. americana* possessed five introns (ranging from 67-145 nt in length, with some size heterogeneity between homologous introns in the two strains), while the *M. jakobiformis* CCT $\alpha$  contained seven introns (58-127 nt long). The *M. jakobiformis* CCT $\delta$  gene possessed seven introns between 61 and 73 nt in length, despite being a shorter fragment of coding sequence than that obtained for CCT $\alpha$  (~ 1 kb of ORF). When the putative introns were removed, the inferred protein sequences were readily alignable with orthologs from a wide range of other eukaryotes with no size heterogeneity.

All the introns possessed 'standard' 5'-GT...AG-3' intron boundaries, with the exception of a single intron in one of two CCT $\delta$  clones from *M. jakobiformis*, which possessed a 'CT' at the 5' intron-exon boundary. It is thus unclear whether this represents a PCR-generated artifact or a legitimate non-canonical 5' intron boundary. A summary of the sizes and positions of the introns found in the *R. americana* and *M. jakobiformis* CCT genes is shown in Figure 4.1. With the exception of a single intron in the *Acrasis rosea* CCT $\alpha$ , none of the non-jakobid CCT $\alpha$  genes sequenced in this study contained introns.

The sequencing of independent clones suggested the presence of multiple copies of the CCT $\alpha$  gene in *R. americana* (50283 and 50394), *Naegleria gruberi*, *Trypanosoma brucei* (see above), and *Monocercomonas* sp., as well as the CCT $\delta$  gene in *M. jakobiformis*. Between the different clone types, most of the observed substitutions were synonymous; however, slightly different amino acid sequences were inferred from different clones of the *Monocercomonas* sp. CCT $\alpha$  and *M. jakobiformis* CCT $\delta$  genes. When introns were present, some size heterogeneity between homologous introns in the different clones was also observed. The presence of multiple copies of CCT genes was confirmed for CCT $\alpha$  in *R. americana* 50394 and for the *M. jakobiformis* CCT $\delta$  gene by Southern hybridization (data not shown). Overall, a strong G + C bias was present in the *R. americana* and *M. jakobiformis* CCT genes, particularly at the 3<sup>rd</sup> position. Edgcomb *et al.* (2001) observed a similar bias in the jakobid tubulin genes.

Alignments of all available CCT $\alpha$  and CCT $\delta$  protein sequences are shown in Figures 4.2 and 4.3, respectively. The high degree of sequence identity shared between the newly sequenced jakobid, heterolobosean, *Trypanosoma* and *Monocercomonas* CCTs and those in other eukaryotes is readily apparent.



**Figure 4.1** Spliceosomal introns in the *Reclinomonas* and *Malawimonas* CCT genes. The schematic shows the positions and sizes of the introns found in the *Reclinomonas americana* and *Malawimonas jakobiformis* CCTalpha genes and the *M. jakobiformis* CCTdelta gene.

**Figure 4.2** Alignment of CCTalpha protein sequences. The alignment contains 28 sequences and 573 amino acid positions. Sequences determined in this study are highlighted by asterisks. Amino acid residues present in at least 60% of the sequences are shaded black and chemically similar amino acids (if present in  $\geq 60\%$  of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Crig*, *Cricetulus griseus*; *Hsap*, *Homo sapiens*; *Rnor*, *Rattus norvegicus*; *Ppal*, *Paleosuchus palpebrosus*; *Mdom*, *Monodelphis domestica*; *Xlae*, *Xenopus laevis*; *Drer*, *Danio rerio*; *Dmel*, *Drosophila melanogaster*; *Sman*, *Schistosoma mansoni*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Spom*, *Schizosaccharomyces pombe*; *Ddis*, *Dicyostelium discoideum*; *Atha*, *Arabidopsis thaliana*; *Tpyr*, *Tetrahymena pyriformis*; *Pfal*, *Plasmodium falciparum*; *Glam*, *Giardia lamblia*; *R394*, *Reclinomonas americana* ATCC 50394; *R283*, *Reclinomonas americana* ATCC 50283; *Aros*, *Acrasis rosea*; *Ngru*, *Naegleria gruberi*; *Tbru*, *Trypanosoma brucei*; *Mjak*, *Malawimonas jakobiformis*; *Tvag*, *Trichomonas vaginalis*; *Mono*, *Monocercomonas* sp.

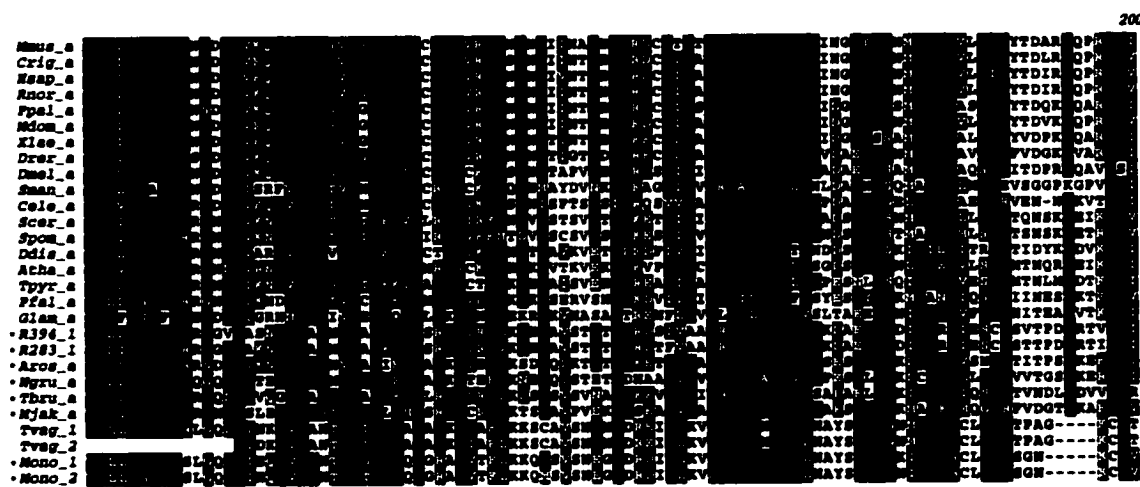


Figure 4.2 Alignment of CCTalpha protein sequences

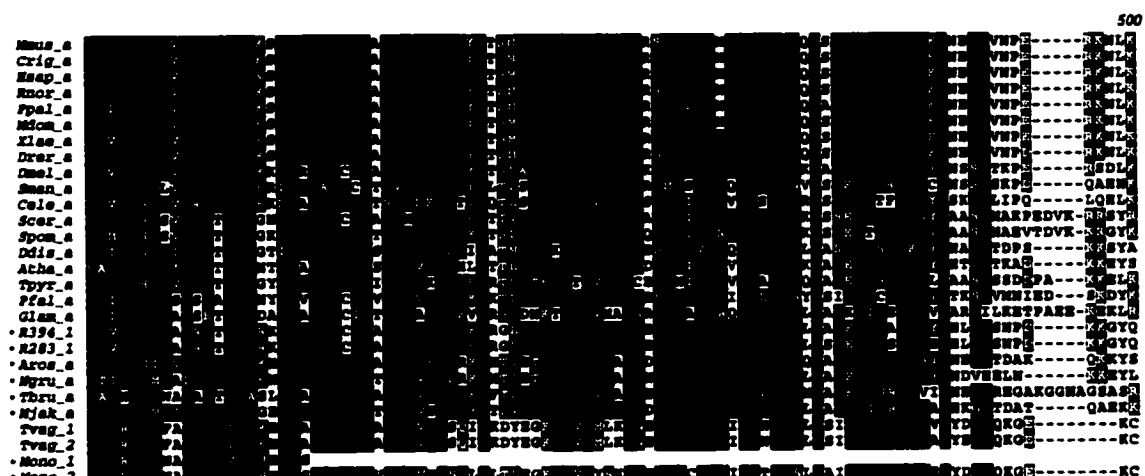


Figure 4.2 Alignment of CCTalpha protein sequences

**Figure 4.3** Alignment of CCTdelta protein sequences. The *Malawimonas jakobiformis* sequence determined in this study is highlighted with an asterisk. The alignment contains 18 CCTdelta sequences and 572 amino acid positions. Amino acid residues present in at least 60% of the sequences in the alignment are shaded black and chemically similar amino acids (if present in > 60% of the sequences) are highlighted gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Hsap*, *Homo sapiens*; *Frub*, *Fugu rubripes*; *Ggal*, *Gallus gallus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Spom*, *Schizosaccharomyces pombe*; *Ncra*, *Neurospora crassa*; *Atri*, *Aedes triseriatus*; *Gmax*, *Glycine max*; *Osat*; *Oryza sativa*; *Atha*, *Arabidopsis thaliana*; *Pfal*, *Plasmodium falciparum*; *Lmaj*, *Leishmania major*; *Tvag*, *Trichomonas vaginalis*; *Glam*, *Giardia lamblia*; *Mjak*, *Malawimonas jakobiformis*.

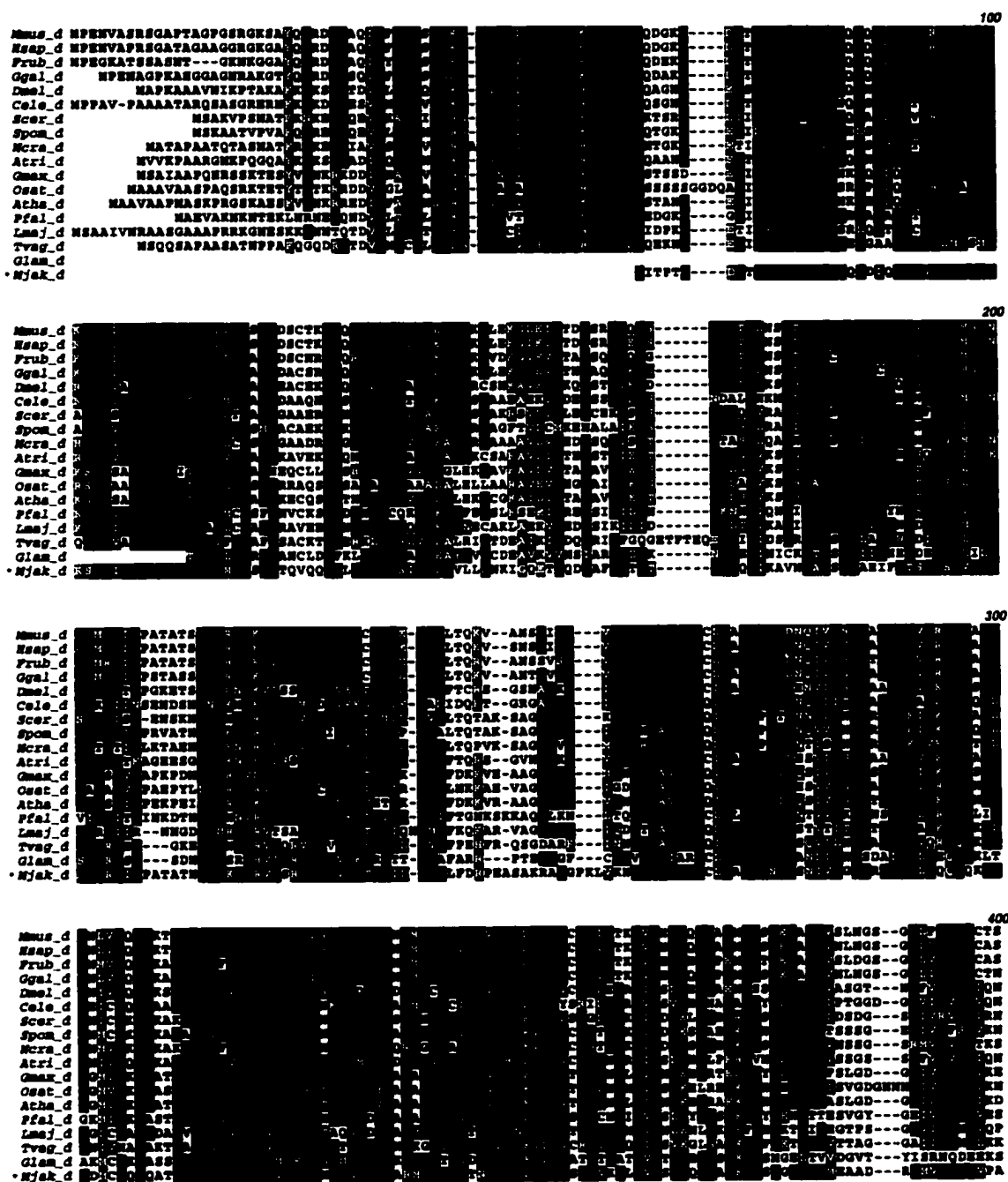


Figure 4.3 Alignment of CCTdelta protein sequences



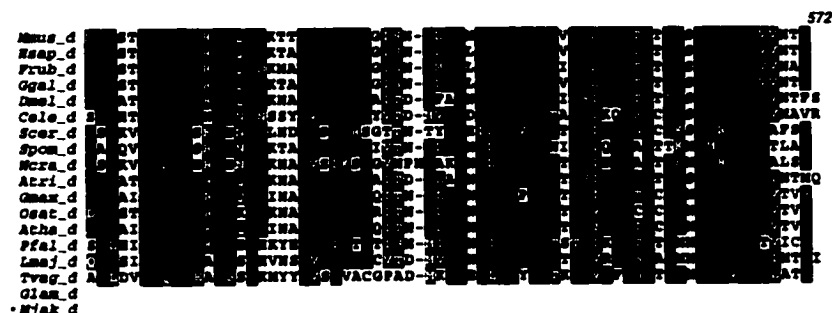
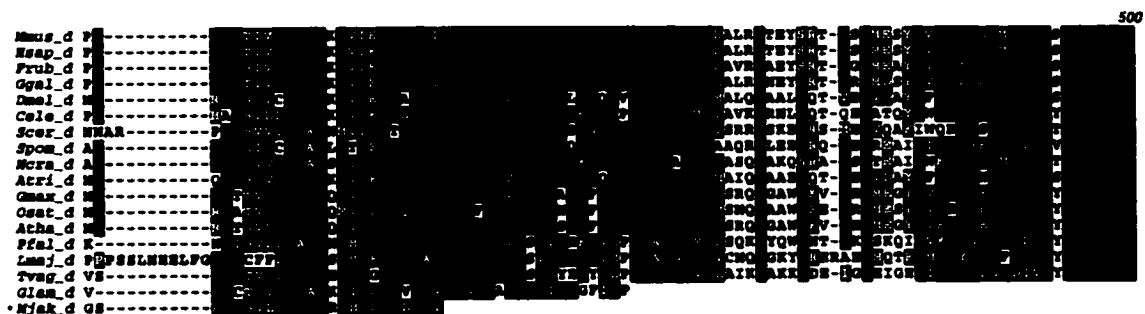
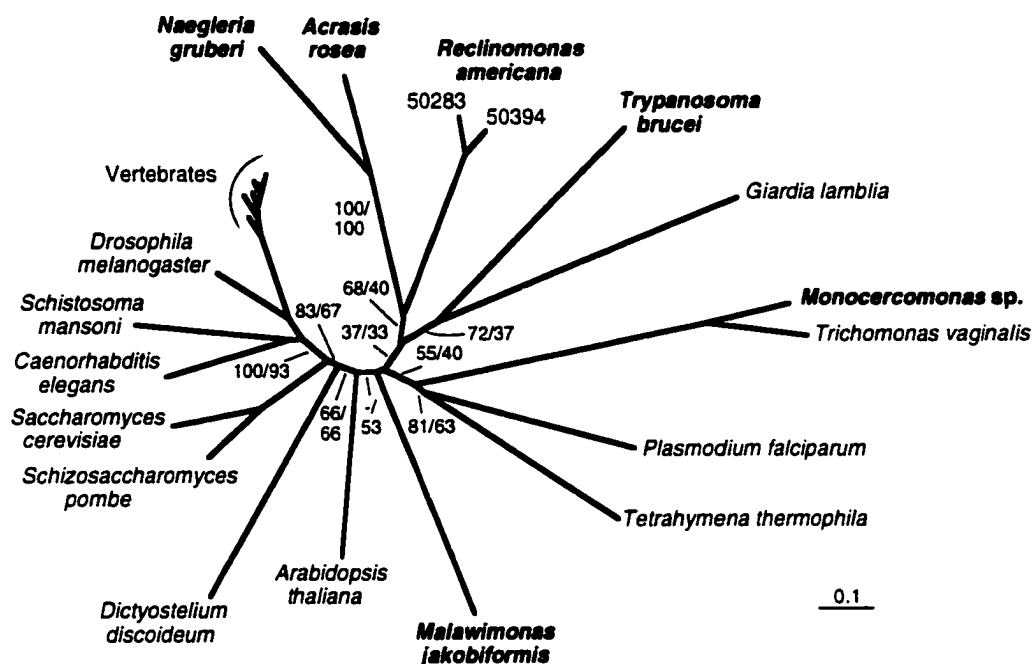


Figure 4.3 Alignment of CCTdelta protein sequences

### CCT $\alpha$ phylogeny

On the basis of mitochondrial gene content (Lang *et al.* 1997) and ultrastructure (O'Kelly 1993), the jakobids could be among the deepest-branching mitochondriate eukaryotes. However, other than alpha- and beta-tubulin genes (Edgcomb *et al.* 2001), no data for jakobid nuclear genes have been available to test this hypothesis. In an attempt to determine the relationships among the jakobid flagellates and their relationship to other eukaryotes, I performed phylogenetic analyses on the CCT $\alpha$  dataset. I also performed multiple rooted analyses using the seven other CCT paralogs to root the CCT $\alpha$  tree.

Figure 4.4 shows a maximum likelihood (ML)-distance phylogeny inferred from CCT $\alpha$  protein sequences. In many ways, the topology is similar to those obtained with more widely used phylogenetic markers such as SSUrRNA, actin, EF-1 $\alpha$ , and alpha- and beta-tubulin. For example, the sisterhood of animals and fungi is moderately well supported in the CCT $\alpha$  tree with all phylogenetic methods, consistent with the results obtained from an ever-increasing wealth of molecular data (see Baldauf 1999 for recent review). As well, the single mycetozoan representative in the CCT $\alpha$  dataset, *Dictyostelium discoideum*, branches weakly but consistently at the base of animals and fungi, in agreement with other molecular phylogenies suggesting that the Mycetozoa are an outgroup to the animal-fungal clade (Baldauf 1999). The alveolates, represented by the ciliate *Tetrahymena pyriformis* and the apicomplexan parasite *Plasmodium falciparum*, receive moderate support as a monophyletic grouping, and, as expected, the two heterolobosean sequences (*Acrasis rosea* and *Naegleria gruberi*) branch together with strong support and with all phylogenetic methods. The parabasalid CCT $\alpha$ s (*Monocercomonas* sp., sequenced here and *Trichomonas vaginalis*, described in Chapter II) are characterized by relatively long branches, similar to phylogenies constructed with other molecules.

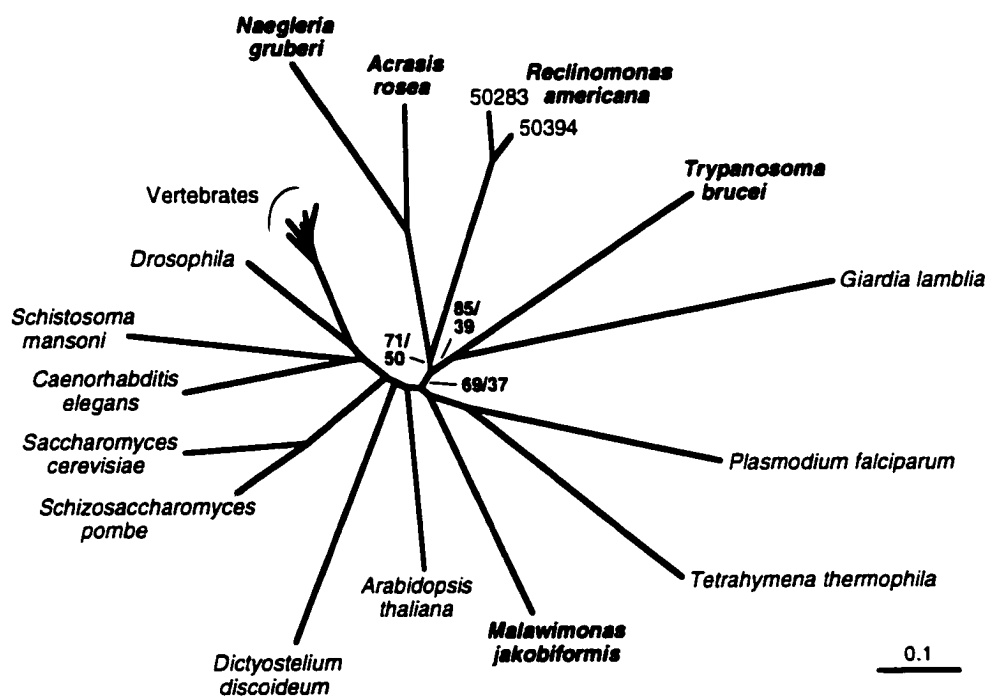


**Figure 4.4** Phylogenetic analysis of CCTalpha protein sequences. The tree shown is a distance (Fitch-Margoliash) tree inferred from a maximum likelihood (ML) distance matrix calculated with a rate heterogeneity model (JTT-F +  $\Gamma$  + inv, pinvar=0.08,  $\alpha$ =1.23; see Materials and Methods). The tree was constructed using a dataset containing 26 taxa and 424 unambiguously aligned amino acid positions. Sequences determined in this study are in bold. Support values for important nodes are ML RELL values and ML-distance bootstrap values, respectively. The scale bar indicates the inferred number of substitutions per site.

The CCT $\alpha$  topology differs from other molecular phylogenies of eukaryotes in several interesting ways. The placement of the alveolates near the parabasalids is not seen with most phylogenetic markers, but is reminiscent of phylogenies of actin, where the ciliate sequences are paraphyletic and branch near the base of the tree (e.g., Bricheux and Brugerolle 1997). As well, the amitochondriate diplomonad *Giardia lamblia* branches with the mitochondriate euglenozoan *Trypanosoma brucei*. Together, these taxa form a weakly supported but consistently observed clade with the heteroloboseans (*N. gruberi* and *A. rosea*) and the two *R. americana* strains. Despite the fact that the *G. lamblia* CCT $\alpha$  is a fairly long branch, this relationship was still observed in the majority of rooted analyses (see below). The two jakobids, *R. americana* and *M. jakobiformis*, showed no affinity for one another in the CCT $\alpha$  tree. While the phylogenetic position of *M. jakobiformis* with respect to the other protists was quite unstable, *R. americana* consistently branched with the two heteroloboseans (as in Figure 4.4), and only occasionally with *T. brucei*.

I tested the effect of removing various long branch taxa on the support for the topology obtained with the full CCT $\alpha$  dataset, as well as the effect of rooting the tree with each of the seven other CCT paralogs. When the parabasalids were removed, the overall topology was the same except for the placement of *M. jakobiformis*, which moved to a position adjacent to the alveolates (Figure 4.5). Interestingly, the support for *G. lamblia* branching with *T. brucei* increased, as did the support for the *R. americana* / Heterolobosea grouping. The removal of the next longest branch taxa, *G. lamblia*, produced a similar result (data not shown).

When CCT $\alpha$  was rooted with each of the seven other CCT paralogs, different topologies were often obtained from different datasets, and from the same dataset analyzed with different phylogenetic methods. However, the parabasalids were usually the deepest branch in the CCT $\alpha$  tree, and the

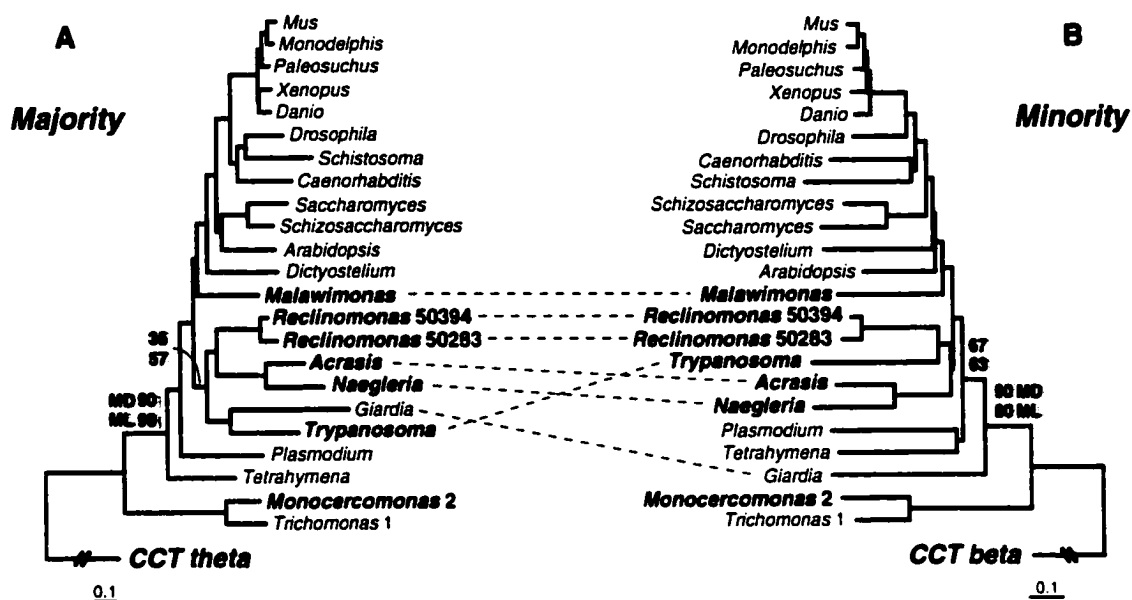


**Figure 4.5** CCTalpha phylogeny with the parabasalid sequences removed. The tree shown is a distance tree inferred from a maximum likelihood distance matrix calculated with a rate heterogeneity model (JTT-F +  $\Gamma$  + inv, pinvar=0.08,  $\alpha$ =1.09). Sequences determined in this study are in bold. The tree was constructed using a dataset containing 24 taxa and 424 unambiguously aligned amino acid positions. Support values for nodes of particular interest are ML RELL values and ML-distance bootstrap values, respectively. The scale bar indicates the inferred number of substitutions per amino acid site.

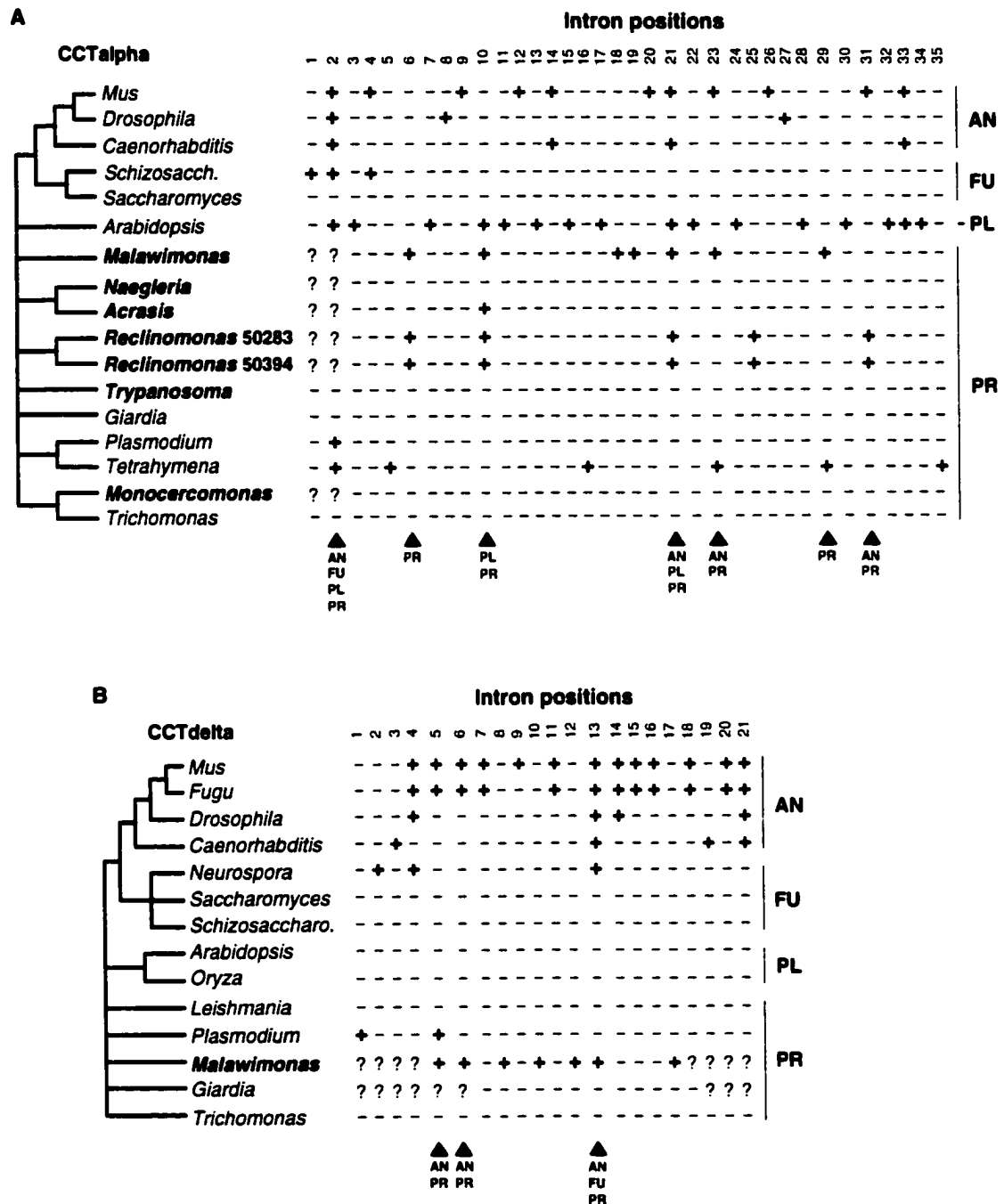
alveolates were often the next deepest branch. Figure 4.6A shows the CCT $\alpha$ - $\theta$  tree, which represents the most commonly obtained topology with the various methods and datasets. Significantly, *G. lamblia* was *not* attached to the long branch of the outgroup in most analyses, but grouped with *T. brucei*. Overall, the internal topology was very similar to that obtained in the unrooted analyses (Figures 4.4 and 4.5). Figure 4.6B shows the CCT $\alpha$ - $\beta$  tree, where the *G. lamblia* sequence branches near the base of CCT $\alpha$ , and *R. americana* branches with *T. brucei* and not with the heteroloboseans.

### Intron phylogeny

The density of spliceosomal introns in the jakobid CCT genes is extraordinarily high—on par with that observed in vertebrate genes (Logsdon 1998). To compare the diversity of intron positions in the jakobid CCTs to those in animals, fungi, plants and other protists, I surveyed all available CCT $\alpha$  and CCT $\delta$  genomic sequences. The results are presented in Figure 4.7. From the 17 CCT $\alpha$  sequences for which genomic sequence data was available, 35 distinct intron positions were observed. For CCT $\delta$ , 14 sequences were available for comparison, and 21 intron positions were found. Interestingly, three out of the five intron positions present in the *R. americana* CCT $\alpha$  gene were shared with *M. jakobiformis*, despite the fact that the two groups showed no affinity for one another in phylogenies of CCT $\alpha$  (above) or tubulin (Edgcomb *et al.* 2001). Of these three, one was shared with *Arabidopsis thaliana* and animals (position 21) and another with *A. thaliana* and the heterolobosean *Acrasis rosea* (position 10). One intron (position 2) was conserved between protists (*Plasmodium falciparum* and *Tetrahymena pyriformis*), plants, fungi and animals (Figure 4.7). For CCT $\delta$ , two introns were shared between protists and animals (positions 5 and 6), and a third between *M. jakobiformis*, one of three fungi (*Neurospora crassa*), and animals



**Figure 4.6** Rooted analyses of CCTalpha. (A) *Majority*. The CCTalpha-theta tree showing the CCTalpha topology most often observed when rooted with another CCT paralog. The tree shown is the best maximum likelihood (ML) tree from a heuristic search of 1000 trees in protML ( $\ln L = -11,253.17$ ). The tree was constructed from an alignment containing 334 unambiguously aligned amino acid positions. (B) *Minority*. The CCT-alpha-beta tree, an example of an alternate CCTalpha topology. The tree shown is the ML distance tree inferred with the Fitch-Margoliash algorithm from a ML distance matrix calculated with a rate heterogeneity model (JTT-F +  $\Gamma$  + inv, pinvar=0.08,  $\alpha=1.49$ ). The alignment contained 381 sites. Sequences determined in this study are in bold. Support values are only provided for several nodes of interest (MD, ML-distance bootstrap values (500 replicates), ML, ML RELL values). Dashed lines between (A) and (B) highlight differences in the deepest branches of the two CCTalpha topologies. The scale bars indicate the inferred number of substitutions per site.



**Figure 4.7** CCTalpha and CCTdelta intron phylogenies. (A) Phylogenetic distribution of 35 known CCTalpha intron positions. The tree shown is a consensus topology of rooted and unrooted phylogenies of CCTalpha protein sequences (Figures 4.4, 4.5 and 4.6) showing only strongly supported relationships. A plus (+) sign indicates the presence of an intron, a minus indicates intron absence, and a question mark indicates missing data. Intron positions are numbered and potentially conserved intron positions are highlighted with an arrow. (B) Phylogenetic distribution of 21 known CCTdelta intron positions. Topology shown is a consensus of the the CCTdelta phylogenetic analyses (data not shown). Intron positions are numbered. A plus (+) sign indicates the presence of an intron, a minus indicates intron absence, and a question mark indicates missing data. Potentially conserved intron positions are indicated with an arrow.



(position 13). Despite the general high abundance of introns in plant protein-coding genes (Logsdon 1998)—16 introns were found in the *Arabidopsis* CCT $\alpha$  gene—neither of the plant CCT $\delta$ s (*Arabidopsis* or *Oryza*) contained any introns. This likely represents a case of ‘recent’ retro-integration of a cDNA, although in the absence of additional data, it is difficult to say how recent. A large number of unique intron positions are also present in both datasets. 24 of 35 CCT $\alpha$  introns, and nine of 21 CCT $\delta$  introns were present in a single species (considering the two *R. americana* strains as a single group), although for CCT $\alpha$ , 12 of the 24 introns were unique to *Arabidopsis*. Two of the seven CCT $\alpha$ , and four of the seven CCT $\delta$  introns in *M. jakobiformis* were not present in any other taxa.

## DISCUSSION

Until recently, attempts to determine the phylogenetic affinities of the jakobid flagellates to one another and to other protists have relied solely on their ultrastructural features. Within the jakobids, the families Jakobidae and Histionidae appear most similar (the ‘core jakobids’), while the third family, Malawimonadidae, may be only distantly related to the other two (O’Kelly 1993; O’Kelly and Nerad 1999; Simpson 1999). The analyses presented here are among the first to characterize nuclear protein-coding genes from the jakobid flagellates, and are the first to assess the utility of the cytosolic chaperonin CCT $\alpha$  as a marker for eukaryotic phylogeny. While the absence of a CCT $\alpha$  sequence from *Jakoba libera* (see Results) precludes any direct comparisons between members of the three jakobid families, my analyses do indicate that *Reclinomonas americana* and *Malawimonas jakobiformis* show no detectable phylogenetic affinity for one another. This result is consistent with phylogenies constructed from concatenated mitochondrial proteins (Burger *et al.* 1999) and with a recent

analysis of jakobid alpha- and beta-tubulins (Edgcomb *et al.* 2001). In the CCT $\alpha$  phylogenies *M. jakobiformis* showed no affinity for any particular protist group, while *R. americana* branched weakly but consistently with the heteroloboseans (see below).

On the whole, CCT $\alpha$  phylogenies show a great deal of congruence with more commonly used phylogenetic markers: animals and fungi appear to be each others closest relatives, the mycetozoan *Dictyostelium* appears as an immediate outgroup to the animal/fungal clade, and the alveolates and Heterolobosea are monophyletic groups. Most interesting was the placement of the diplomonad *Giardia lamblia*. Almost without exception, rooted SSUrRNA and protein phylogenies place the diplomonads at or near the base of the eukaryotic tree (Roger 1999). In most rooted analyses of CCT $\alpha$ , *Giardia* was not positioned at base of the eukaryotic tree, but was nested within mitochondrion-containing groups. This is consistent with the suggestion that diplomonads have lost their mitochondria secondarily (Roger 1999; Roger *et al.* 1998). The analyses presented here illustrate *the* problem with deep phylogeny, long branch attraction (e.g., Germot and Philippe 1999; Philippe and Germot 2000; Stiller and Hall 1999). It is clear that the two parabasalids in the CCT $\alpha$  dataset have the longest branches of all the taxa—they also emerge consistently at the base of the CCT $\alpha$  tree in rooted analyses. For this reason, the deepest branches of the rooted CCT $\alpha$  phylogenies presented here should be viewed with caution.

The high density of introns in the jakobid CCT genes raises the question of their origin. Were they 'recently' acquired, or do they represent the retention of 'old' introns? Completely independent of the issue of the origin of spliceosomal introns themselves (i.e., 'introns-early' versus 'introns-late'; current data favor the latter [Kwiatowski *et al.* 1995; Logsdon 1998; Logsdon *et al.* 1995]), is the question of the diversity and antiquity of introns *within* eukaryotes. For the most

part, protist genes are intron-sparse, and many of the lineages that have figured prominently in hypotheses of early events in eukaryotic evolution (e.g., diplomonads and parabasalids) seem to lack introns entirely (Logsdon 1998). Remarkably few intron positions are known to be conserved between animals, fungi and plants (Fast, Logsdon and Doolittle 1999; Palmer and Logsdon 1991), and fewer still between animals, fungi, plants and protists. Surprisingly, the CCT $\alpha$  dataset contains three potentially 'old' introns (positions 2, 21 and 23; Figure 4.7), and there are possibly another two in CCT $\delta$  (positions 5 and 13). Unfortunately, the CCT $\alpha$  dataset is biased against the presence of 'old' introns due to the fact that the sole fungal representatives, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, are known to be relatively intron-sparse compared to other fungi. The intron density in *Saccharomyces cerevisiae* is only 0.1 per kb (Logsdon 1998), and the majority of introns are found at the extreme 5' ends of genes, suggesting that cDNA-mediated intron loss has occurred (Fink 1987). In the CCT $\delta$  dataset, three introns were found in the *Neurospora crassa* CCT $\alpha$  (*N. crassa* has a higher intron density), and two of these were shared with animals. The CCT $\delta$  dataset is similarly biased. No introns were found in the *S. cerevisiae* or *Schizosaccharomyces pombe* CCT $\delta$  genes, and despite the fact that the intron density in plant protein-coding genes is generally quite high (Logsdon 1998), neither of the plant CCT $\delta$  sequences (*Arabidopsis* or *Oryza*) contained a single intron. With few protist introns available for comparison, it is difficult to make too strong a prediction as to whether the potentially 'old' jakobid introns are due to the retention of ancestral intron positions or are the result of parallel insertions. More CCT $\alpha$  sequences from each of the major eukaryotic groups will be needed to resolve this issue.

It is certainly clear from both datasets that recent intron gain has occurred in the jakobid CCTs and in the CCTs of other eukaryotes. The *Arabidopsis thaliana*

CCT $\alpha$  gene alone contained 12 introns in positions not found in any other taxa, and the CCT $\alpha$  from the ciliate *Tetrahymena pyriformis* contained another three (Figure 4.7, Results). For the jakobid CCT $\alpha$ s, one of the *R. americana* introns and two of the *M. jakobiformis* introns were located in unique positions, as were four of seven *M. jakobiformis* CCT $\delta$  introns. Data from the jakobid tubulin genes also support this notion. As predicted from ultrastructure, the two 'core jakobids', *R. americana* and *J. libera*, branched strongly together in beta- and alpha-/beta-tubulin phylogenies. Yet despite this apparent close relationship, their respective beta-tubulins each possessed a single intron located in different positions, neither of which is found in other known beta-tubulin genes (Edgcomb *et al.* 2001).

As for intron loss, the data are more ambiguous. In general, this uncertainty stems from the lack of a robust phylogeny of the major protist groups upon which intron gain/loss scenarios can be evaluated. There are, however, several cases where intron loss can be inferred from a known or suspected phylogenetic relationship. For example, the apicomplexan *P. falciparum*, whose CCT $\alpha$  contains one intron, branches with the intron-rich ciliate *T. pyriformis* (six introns). A stronger case for intron loss can be inferred for the Heterolobosea. Consistent with the low intron density observed thus far for heterolobosean protein-coding genes (Logsdon 1998), only one intron was found in the CCT $\alpha$  gene from *Acrasis rosea*, and none were present in the *Naegleria gruberi* sequence. However, the intron in the *A. rosea* CCT $\alpha$  is shared with the two *R. americana* strains, *M. jakobiformis*, and *Arabidopsis* (intron position 10; Figure 4.7), suggesting that it may predate the divergence of plants and these protists.

In general, it seems significant that in cases where only one or a few introns are present in the CCT genes (e.g., *Acrasis*, *Plasmodium*, *Schizosaccharomyces*; Figure 4.7) they are located near the 5' end (in CCT $\alpha$ , intron

positions 1-10 are located in the first 20% of the coding sequence while introns 1-5 in CCT $\delta$  are present in the first 11% of the gene). As mentioned previously, this is also observed for the vast majority of introns in yeast genes; a process of reverse-transcriptase-mediated intron loss has been suggested to account for such a non-random distribution (Fink 1987). As for the amitochondriates, although no introns have been described in genes from diplomonads or parabasalids to date, indirect evidence for their existence in the parabasalid *Trichomonas vaginalis* has come from the discovery of a homolog of PRP8, a highly conserved protein component of the spliceosome (Fast and Doolittle 1999). A gene whose product has significant similarity to PRP8 is also present in the near complete genome of the diplomonad *Giardia lamblia* (Smith *et al.* 1998) (identified by searching the high-throughput genome sequence database at NCBI). Taken as a whole, the data suggest that the intron-sparse/-free nature of most protist genes is a derived feature.

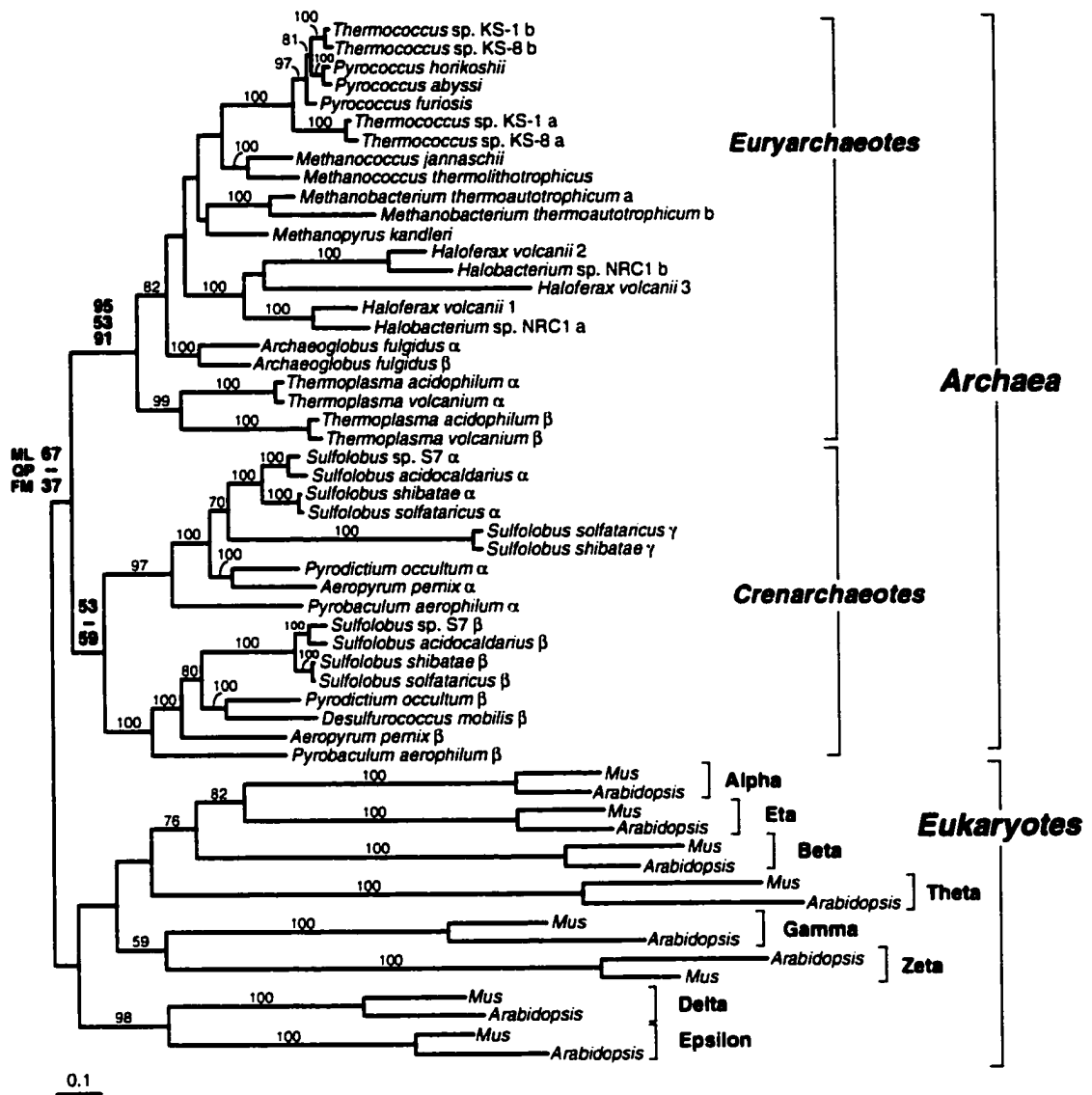
The weak but consistent affinity of *Reclinomonas americana* for the heteroloboseans in CCT $\alpha$  phylogenies (Figures 4.4, 4.5 and 4.6) is interesting for two reasons. First, it is consistent with beta-tubulin and combined alpha-beta-tubulin trees (Edgcomb *et al.* 2001). Second, it is consistent with mitochondrial gene-content data. The mtDNA of the heterolobosean *Naegleria gruberi* contains (in addition to many genes shared with plant and protist mtDNAs) two genes thus far only identified in jakobid mitochondrial genomes (*atp3* and *cox11*; M. W. Gray, pers. comm.; <http://megasun.bch.umontreal.ca/ca.ogmp/projects/ngrub/gen.html>). If the Heterolobosea and the core jakobids (represented here by *R. americana*) are in fact related, intron loss can be more confidently inferred for the heterolobosean CCT $\alpha$  genes.

While the grouping of *R. americana* with the Heterolobosea is consistent with beta-tubulin and combined alpha/beta-tubulin phylogenies (Edgcomb *et al.*

2001) and mitochondrial gene-content data, it is at odds with mitochondrial cristal morphology. Like *Naegleria gruberi* and the euglenozoan *Trypanosoma brucei*, the jakobid *M. jakobiformis* possesses discoidal cristae, yet it is the Histonid *R. americana*, which possesses tubular mitochondrial cristae, that shows the most affinity for the Heterolobosea and *T. brucei* in the CCT $\alpha$  and tubulin trees. While cristal morphology has often been taken to be an evolutionarily stable character, these data suggest that it may not always be reliable for tracking the deepest divisions in eukaryotic evolution. More generally, the significance of the retention of ancestral features in jakobid mtDNAs, in terms of their relationship to other eukaryotes, is as yet unclear. The *M. jakobiformis* mtDNA possesses three genes (*rpl18*, 19, 31) that have thus far not been observed in non-jakobid protist mtDNAs (<http://megasun.bch.umontreal.ca/ogmp/projects/mjako/gen.html>), yet all phylogenetic analyses performed thus far on both mitochondrial and nuclear genes indicate that the jakobids are not a monophyletic group. Similarly, *R. americana* and *M. jakobiformis* share three introns in their CCT $\alpha$  genes despite appearing to be unrelated in CCT $\alpha$  phylogenies.

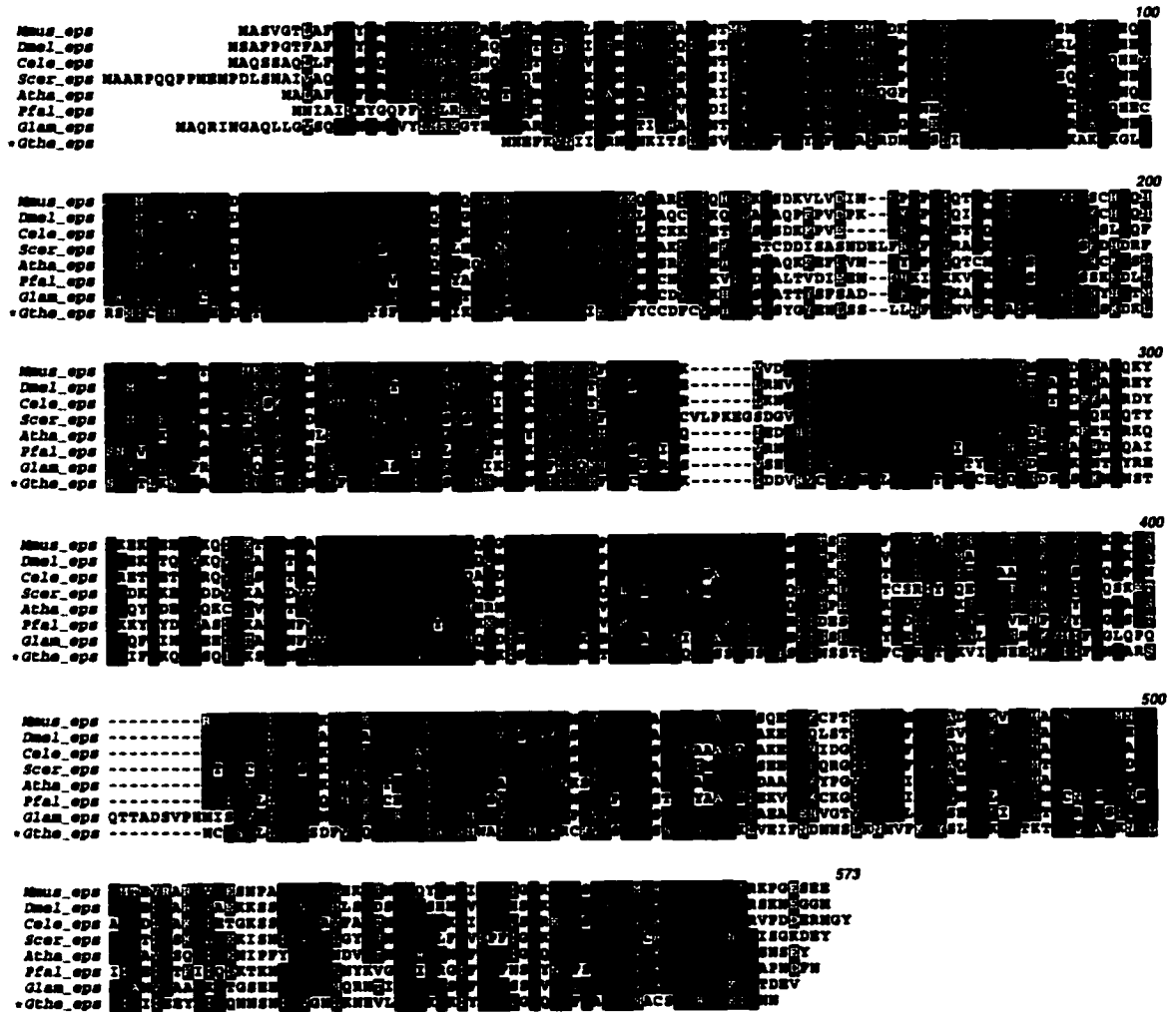
How intron-rich will jakobid nuclear genes turn out to be? At present, such a small sample size makes it impossible to predict. Nevertheless, the data presented here paint a very different picture of spliceosomal intron evolution than has generally been assumed, one in which intron loss has been a significant factor in shaping eukaryotic nuclear genomes.

## Appendix A: Supplementary data



**Figure A.1** Rooting the archaeal chaperonin tree. The tree shown is the maximum likelihood tree ( $\ln L = -19,834.0$ ) inferred from a heuristic search of 1000 trees in protML (Adachi and Hasegawa, 1996). The alignment contained 40 archaeal sequences, 16 representative eukaryotic CCTs (2 from each of the eight CCT paralogs) and 348 unambiguously aligned amino acid positions. Bootstrap support (ML RELL values) for all branches is given if  $> 50\%$ . Support for nodes of particular interest is also provided (ML, ML RELL values; QP, quartet puzzling support values; FM, Fitch-Margoliash (distance) bootstrap values). The scale bar represents the estimated number of amino acid substitutions per site.

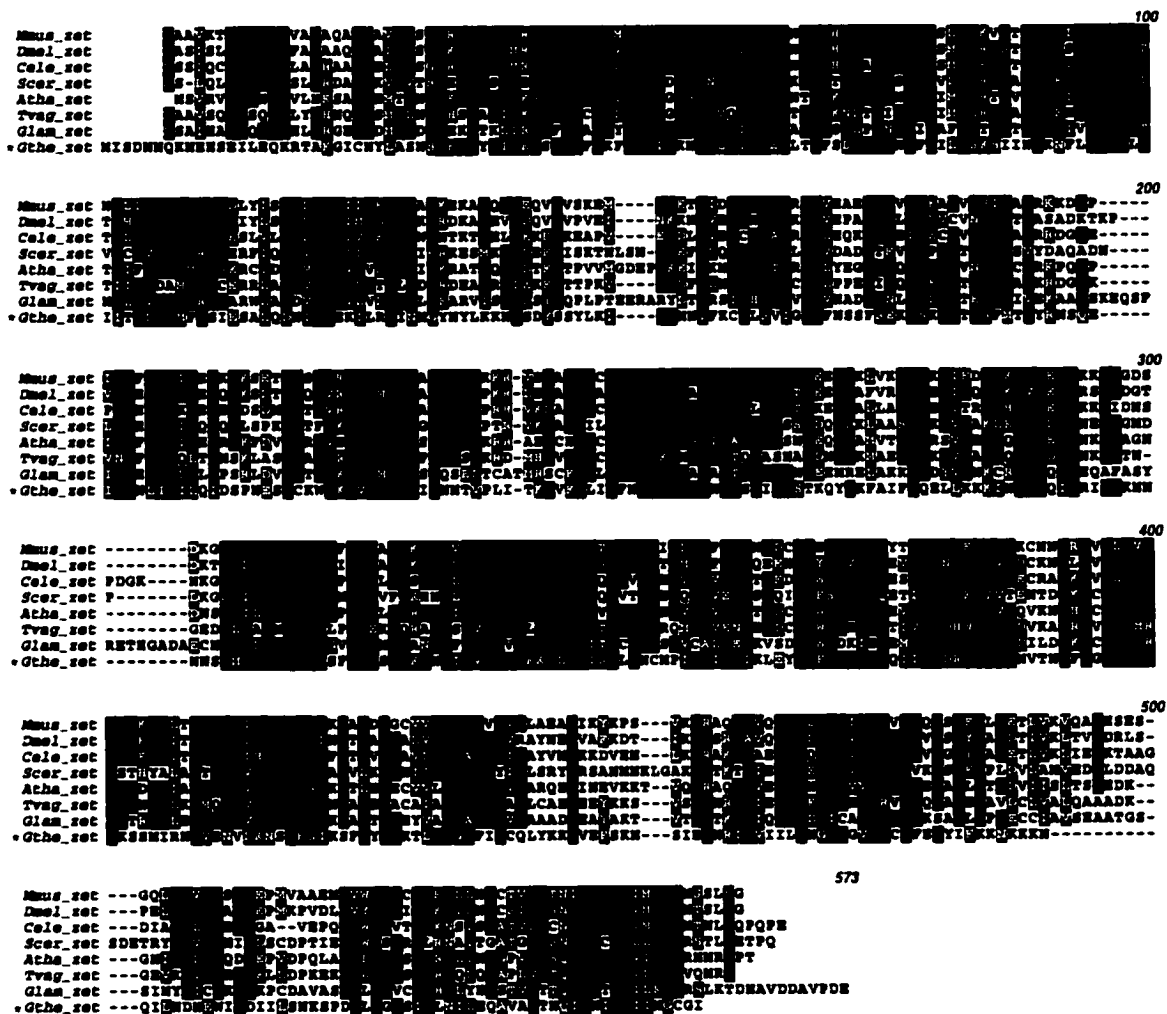
## Appendix B: Supplementary alignments



**Figure B.1** Alignment of select CCTepsilon protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Pfal*, *Plasmodium falciparum*; *Glam*, *Giardia lamblia*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *eps*, CCTepsilon.

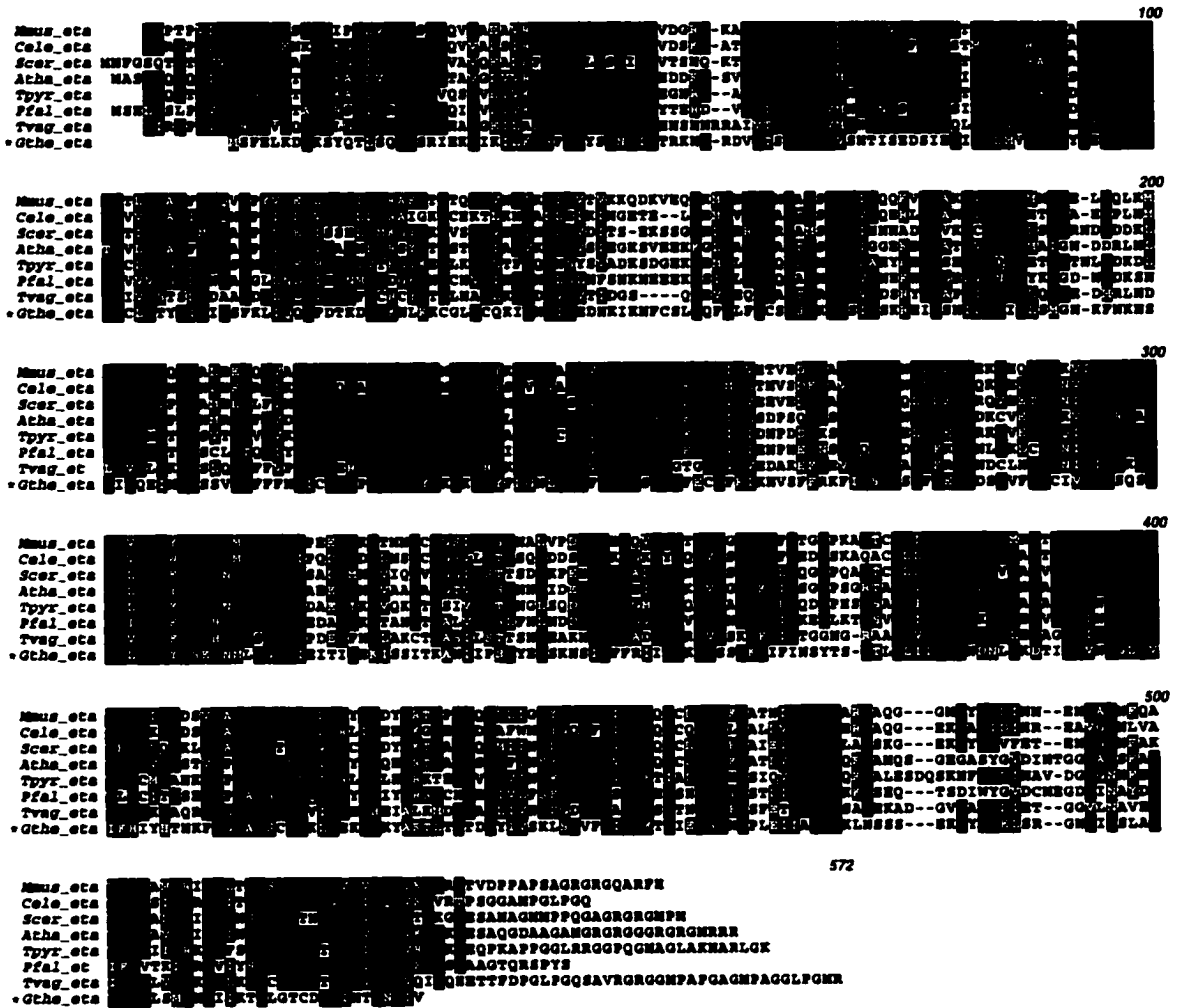


## Appendix B: continued



**Figure B.2** Alignment of select CCTzeta protein sequences. The *Guillardia theta* nucleomorph sequence is highlighted by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Tvag*, *Trichomonas vaginalis*; *Glam*, *Giardia lamblia*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *zet*, CCTzeta.

Appendix B: continued



**Figure B.3** Alignment of select CCTeta protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Tpyr*, *Tetrahymena pyriformis*; *Pfal*, *Plasmodium falciparum*; *Tvag*, *Trichomonas vaginalis*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *eta*, CCTeta.

Appendix B: continued

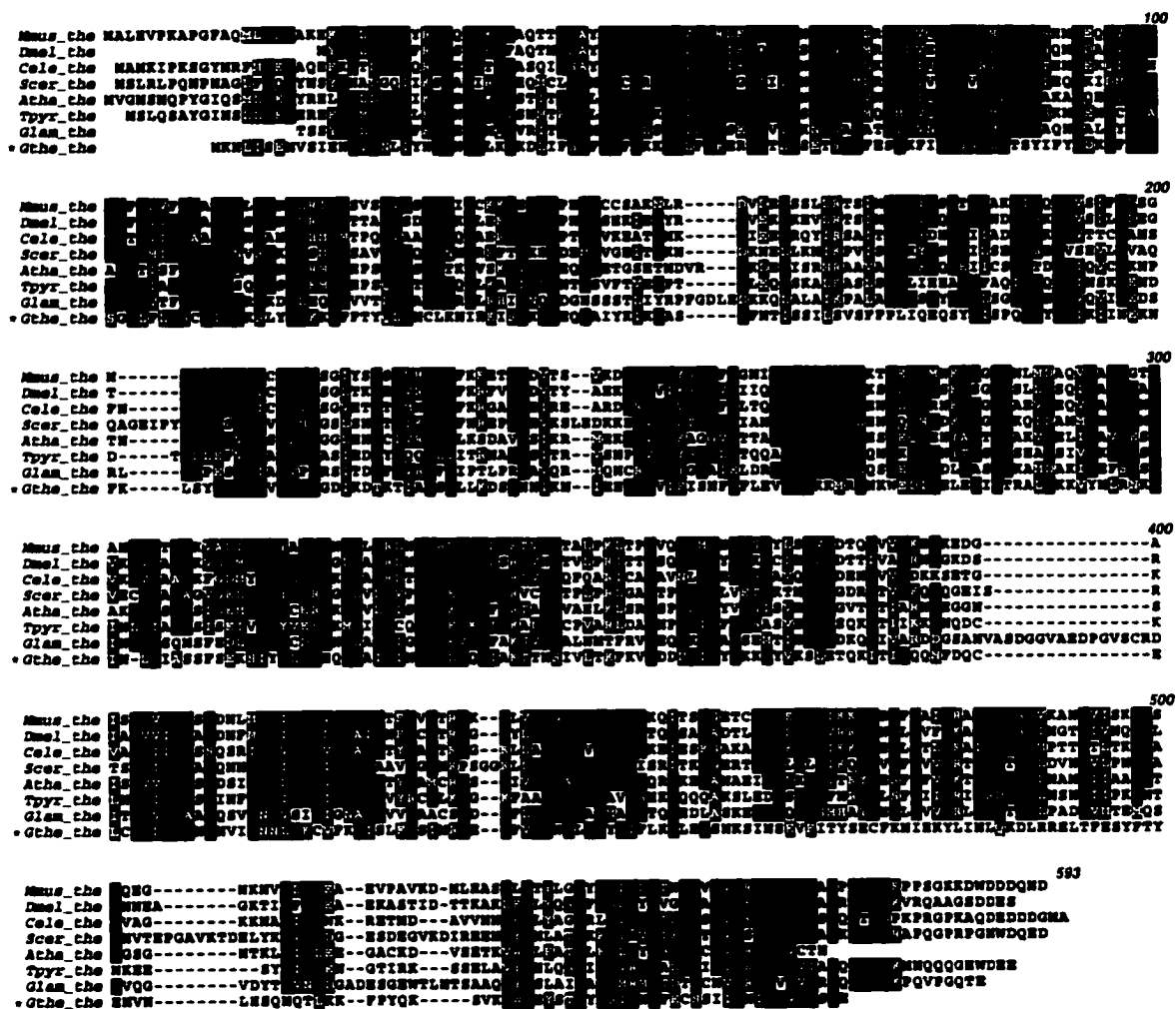


Figure B.4 Alignment of select CCTtheta protein sequences. The *Guillardia theta* nucleomorph sequence is indicated by an asterisk. Amino acid residues present in at least 60% of the sequences are shaded black; chemically similar residues (if present in >60% of the sequences) are shaded gray. Taxon abbreviations: *Mmus*, *Mus musculus*; *Dmel*, *Drosophila melanogaster*; *Cele*, *Caenorhabditis elegans*; *Scer*, *Saccharomyces cerevisiae*; *Atha*, *Arabidopsis thaliana*; *Tpyr*, *Tetrahymena pyriformis*; *Glam*, *Giardia lamblia*; *Gthe*, *Guillardia theta* nucleomorph. Subunit abbreviation: *the*, CCTtheta.

## REFERENCES

- Adachi, J. M., and M. Hasegawa. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* **28**: 1-150.
- Ahmad, S., and R. S. Gupta. 1990. Cloning of a Chinese hamster protein homologous to the mouse t-complex protein TCP-1: structural similarity to the ubiquitous 'chaperonin' family of heat-shock proteins. *Biochim. Biophys. Acta* **1087**: 253-255.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
- Altschul, S. F., and E. V. Koonin. 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**: 444-447.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389-3402.
- Andra, S., G. Frey, M. Nitsch, W. Baumeister, and K. O. Stetter. 1996. Purification and structural characterization of the thermosome from the hyperthermophilic archaeum *Methanopyrus kandleri*. *FEBS Lett.* **379**: 127-131.
- Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223-230.
- Archibald, J. M., C. Blouin, and W. F. Doolittle. 2001. Gene duplication and the evolution of group II chaperonins: implications for structure and function. *J. Struct. Biol.* (in press).
- Archibald, J. M., T. Cavalier-Smith, U.-G. Maier, and S. Douglas. 2001. Molecular chaperones encoded by a reduced nucleus—the cryptomonad nucleomorph. *J. Mol. Evol.* (in press).
- Archibald, J. M., J. M. Logsdon, Jr., and W. F. Doolittle. 1999. Recurrent paralogy in the evolution of archaeal chaperonins. *Curr. Biol.* **9**: 1053-1056.
- Archibald, J. M., J. M. Logsdon, Jr., and W. F. Doolittle. 2000. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol. Biol. Evol.* **17**: 1456-1466.
- Baldauf, S. L. 1999. A search for the origins of animals and fungi: comparing and combining molecular data. *Am. Nat.* **154**(Suppl.): S178-S188.

- Baldauf, S. L., J. D. Palmer, and W. F. Doolittle. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA* **93**: 7749-7754.
- Barraclough, R., and R. J. Ellis. 1980. Protein synthesis in chloroplasts. IX. Assembly of newly-synthesized large subunits into ribulose biphosphate carboxylase in isolated intact pea chloroplasts. *Biochim. Biophys. Acta* **608**: 19-31.
- Baumeister, W., J. Walz, F. Zuhl, and E. Seemuller. 1998. The proteasome: paradigm of a self-compartmentalizing protease. *Cell* **92**: 367-380.
- Bohen, S. P., A. Kralli, and K. R. Yamamoto. 1996. Hold 'em and fold 'em: chaperones and signal transduction. *Science* **268**: 1303-1304.
- Bosch, G., W. Baumeister, and L. O. Essen. 2000. Crystal structure of the beta-apical domain of the thermosome reveals structural plasticity in the protrusion region. *J. Mol. Biol.* **301**: 19-25.
- Braig, K., Z. Otwinowski, R. Hegde, D. C. Boisvert, A. Joachimiak, A. L. Horwich, and P. B. Sigler. 1994. The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* **371**: 578-586.
- Bricheux, G., and G. Brugerolle. 1997. Molecular cloning of actin genes in *Trichomonas vaginalis* and phylogeny inferred from actin sequences. *FEMS Microbiol. Lett.* **153**: 205-213.
- Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**: 817-825.
- Brown, J. R., and W. F. Doolittle. 1995. Root of the universal tree of life based on ancient aminoacyl tRNA synthetase gene duplications. *Proc. Natl. Acad. Sci. USA* **92**: 2441-2445.
- Brown, J. R., and W. F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**: 456-502.
- Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle. 1994. Evolutionary relationships of bacterial and archaeal glutamate synthetase genes. *J. Mol. Evol.* **38**: 566-576.
- Brown, J. R., F. T. Robb, R. Weiss, and W. F. Doolittle. 1997. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. *J. Mol. Evol.* **45**: 9-16.
- Bukau, B., E. Deuerling, C. Pfund, and E. A. Craig. 2000. Getting newly synthesized proteins into shape. *Cell* **101**: 119-122.

- Bukau, B., and A. L. Horwich. 1998. The Hsp70 and Hsp60 chaperone machines. *Cell* **92**: 351-366.
- Burger, G., D. Saint-Louis, M. W. Gray, and B. F. Lang. 1999. Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. *Plant Cell* **11**: 1675-1694.
- Cavalier-Smith, T. 1983a. Endosymbiotic origin of the mitochondrial envelope. Pp. 265-279 in W. Schwemmler and H. E. A. Schenk, eds. *Endocytobiology II: intracellular space as an oligogenetic ecosystem*. De Gruyter, Berlin.
- Cavalier-Smith, T. 1983b. A 6-kingdom classification and unified phylogeny. Pp. 1027-1034 in W. Schwemmler and H. E. A. Schenk, eds. *Endocytobiology II: intracellular space as an oligogenetic ecosystem*. De Gruyter, Berlin.
- Cavalier-Smith, T. 1987. Eukaryotes with no mitochondria. *Nature* **326**: 332-333.
- Cavalier-Smith, T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Euk. Microbiol.* **46**: 347-366.
- Cavalier-Smith, T. 2000. Membrane heredity and early chloroplast evolution. *Trends Plant Sci.* **5**: 174-182.
- Cavalier-Smith, T., J. A. Couch, K. E. Thorsteinsen, P. Gilson, J. A. Deane, D. R. A. Hill, and G. I. McFadden. 1996. Cryptomonad nuclear and nucleomorph 18S rRNA phylogeny. *Eur. J. Phycol.* **31**: 315-328.
- Chandrasekhar, G. N., K. Tilly, C. Woolford, R. Hendrix, and C. Georgopoulos. 1986. Purification and properties of the groES morphogenetic protein of *Escherichia coli*. *J. Biol. Chem.* **261**: 12414-12419.
- Clark, C. G. 1992. DNA Purification from polysaccharide-rich cells. Pp. D3.1-D3.2 in J. J. Lee and A. T. Soldo, eds. *Protocols in Protozoology*. Allen, Lawrence, KS.
- Clark, C. G., and A. J. Roger. 1995. Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*. *Proc. Natl. Acad. Sci. USA* **92**: 6518-6521.
- Deane, J. A., M. Fraunholz, V. Su, U.-G. Maier, W. Martin, D. G. Durnford, and G. I. McFadden. 2000. Evidence for nucleomorph to host nucleus gene transfer: light-harvesting complex proteins from cryptomonads and chlorarachniophytes. *Protist* **151**: 239-252.
- Dingwall, C., and R. A. Laskey. 1991. Nuclear targeting signals-a consensus? *Trends Biochem. Sci.* **16**: 478-481.
- Ditzel, L., J. Lowe, D. Stock, K. O. Stetter, H. Huber, R. Huber, and S. Steinbacher. 1998. Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. *Cell* **93**: 125-138.

- Douglas, S. E., C. A. Murphy, D. F. Spencer, and M. W. Gray. 1991. Molecular evidence that cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* **350**: 148-151.
- Douglas, S. E., and S. L. Penny. 1999. The plastid genome from the cryptomonad alga, *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48**:236-244.
- Douglas, S. E., and S. Turner. 1991. Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. *J. Mol. Evol.* **33**: 267-273.
- Douglas, S. E., S. Zauner, M. Fraunholz, M. Beaton, S. Penny, L. Deng, X. Wu, M. Reith, T. Cavalier-Smith, and U.-G. Maier. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* (in press).
- Drouin, G., F. Prat, M. Ell, and G. D. Clarke. 1999. Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.* **16**: 1369-1390.
- Edgcomb, V. P., A. J. Roger, A. G. B. Simpson, D. T. Kysela, and M. L. Sogin. 2001. Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. *Mol. Biol. Evol.* **18**: 514-522.
- Edgell, D. R., S. B. Malik, and W. F. Doolittle. 1998. Evidence of independent gene duplications during the evolution of archaeal and eukaryotic family B DNA polymerases. *Mol. Biol. Evol.* **15**: 1207-1217.
- Ellis, J. 1987. Proteins as molecular chaperones. *Nature* **328**: 378-379.
- Ellis, M. J., S. Knapp, P. J. Koeck, Z. Fakoor-Biniyaz, R. Ladenstein, and H. Hebert. 1998. Two-dimensional crystallization of the chaperonin TF55 from the hyperthermophilic archaeon *Sulfolobus solfataricus*. *J. Struct. Biol.* **123**: 30-36.
- Ellis, R. J. 1990. Molecular chaperones: the plant connection. *Science* **250**: 954-959.
- Ellis, R. J. 1996. Chaperonins: introductory perspective. Pp. 1-25 in R. J. Ellis, ed. *The chaperonins*. Academic Press, San Diego.
- Ellis, R. J. 1997. Do molecular chaperones have to be proteins? *Biochem. Biophys. Res. Commun.* **238**: 687-692.
- Ellis, R. J., and S. M. Hemmingsen. 1989. Molecular chaperones: proteins essential for the biogenesis of some macromolecular structures. *Trends Biochem. Sci.* **14**: 339-342.
- Emmerhoff, O. J., H. P. Klenk, and N. K. Birkeland. 1998. Characterization and sequence comparison of temperature-regulated chaperonins from the hyperthermophilic archaeon *Archaeoglobus fulgidus*. *Gene* **215**: 431-438.

Farr, G. W., E. C. Scharl, R. J. Schumacher, S. Sondek, and A. L. Horwich. 1997. Chaperonin-mediated folding in the eukaryotic cytosol proceeds through rounds of release of native and nonnative forms. *Cell* **89**: 927-937.

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

Fast, N. M., J. M. Logsdon, Jr., and W. F. Doolittle. 1999. Phylogenetic analysis of the TATA box binding protein (TBP) gene from *Nosema locustae*: evidence for a microsporidia-fungi relationship and spliceosomal intron loss. *Mol. Biol. Evol.* **16**: 1415-1419.

Feldman, D. E., V. Thulasiraman, R. G. Ferreyra, and J. Frydman. 1999. Formation of the VHL-elongin BC tumor suppressor complex is mediated by the chaperonin TRiC. *Mol. Cell* **4**: 1051-1061.

Felsenstein, J. 1995. PHYLIP (phylogeny inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle.

Fenton, W. A., Y. Kashi, K. Furtak, and A. L. Horwich. 1994. Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* **371**: 614-619.

Fink, G. R. 1987. Pseudogenes in yeast? *Cell* **49**: 5-6.

Fitz-Gibbon, S., A. J. Choi, J. H. Miller, K. O. Stetter, M. I. Simon, R. Swanson, and U. J. Kim. 1997. A fosmid-based genomic map and identification of 474 genes of the hyperthermophilic archaeon *Pyrobaculum aerophilum*. *Extremophiles* **1**: 36-51.

Flavin, M., and T. A. Nerad. 1993. *Reclinomonas americana* n. g., n. sp., a new freshwater heterotrophic flagellate. *J. Euk. Microbiol.* **40**: 172-179.

Friedman, D. I., E. R. Olson, C. Georgopoulos, K. Tilly, I. Herskowitz, and F. Banuett. 1984. Interactions of bacteriophage and host macromolecules in the growth of bacteriophage lambda. *Microbiol. Rev.* **48**: 299-325.

Frydman, J., and J. Höhfeld. 1997. Chaperones get in touch: the Hip-Hop connection. *Trends Biochem. Sci.* **22**: 87-92.

Frydman, J., E. Nimmesgern, H. Erdjument-Bromage, J. S. Wall, P. Tempst, and F. U. Hartl. 1992. Function in protein folding of TRiC, a cytosolic ring complex containing TCP-1 and structurally related subunits. *EMBO J.* **11**: 4767-4778.

Furutani, M., T. Iida, T. Yoshida, and T. Maruyama. 1998. Group II chaperonin in a thermophilic methanogen, *Methanococcus thermolithotrophicus*. Chaperone activity and filament-forming ability. *J. Biol. Chem.* **273**: 28399-28407.



- Gao, Y., J. O. Thomas, R. L. Chow, G. H. Lee, and N. J. Cowan. 1992. A cytoplasmic chaperonin that catalyzes beta-actin folding. *Cell* **69**: 1043-1050.
- Gebauer, M., R. Melki, and U. Gehring. 1998. The chaperone cofactor Hop/p60 interacts with the cytosolic chaperonin-containing TCP-1 and affects its nucleotide exchange and protein folding activities. *J. Biol. Chem.* **273**: 29475-29480.
- Geissler, S., K. Siegers, and E. Schiebel. 1998. A novel protein complex promoting formation of functional alpha- and gamma-tubulin. *EMBO J.* **17**: 952-966.
- Georgopoulos, C. P., and B. Hohn. 1978. Identification of a host protein necessary for bacteriophage morphogenesis (the groE gene product). *Proc. Natl. Acad. Sci. USA* **75**: 131-135.
- Germot, A., and H. Philippe. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. *J. Euk. Microbiol.* **46**: 116-124.
- Germot, A., H. Philippe, and H. Le Guyader. 1996. Presence of a mitochondrial-type 70-kDa heat shock protein in *Trichomonas vaginalis* suggests a very early mitochondrial endosymbiosis in eukaryotes. *Proc. Natl. Acad. Sci. USA* **93**: 14614-14617.
- Germot, A., H. Philippe, and H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. *Mol. Biochem. Parasitol.* **87**: 159-168.
- Gibbs, S. P. 1979. The route of entry of cytoplasmically-synthesized proteins into chloroplasts of algae possessing chloroplast ER. *J. Cell Sci.* **35**: 253-166.
- Gibbs, S. P. 1981. The chloroplast endoplasmic reticulum: structure, function, and evolutionary significance. *Int. Rev. Cytol.* **72**: 49-99.
- Gillott, M. A., and S. P. Gibbs. 1980. The cryptomonad nucleomorph: its ultrastructure and evolutionary significance. *J. Phycol.* **16**: 558-568.
- Gilson, P. R., U. G. Maier, and G. I. McFadden. 1997. Size isn't everything: lessons in genetic miniaturisation from nucleomorphs. *Curr. Opin. Genet. Dev.* **7**: 800-806.
- Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida. 1989. Evolution of the vacuolar H<sup>+</sup>-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* **86**: 6661-6665.
- Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.* **49**: 652-670.
- Goloubinoff, P., J. T. Christeller, A. A. Gatenby, and G. H. Lorimer. 1989. Reconstitution of active dimeric ribulose bisphosphate carboxylase from an

- unfolded state depends on two chaperonin proteins and Mg-ATP. *Nature* **342**: 884-889.
- Gray, M. W., G. Burger, and B. F. Lang. 1999. Mitochondrial evolution. *Science* **283**: 1476-1481.
- Gray, M. W., B. F. Lang, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, N. Brossard, E. Delage, T. G. Littlejohn, I. Plante, P. Rioux, D. Saint-Louis, Y. Zhu, and G. Burger. 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.* **26**: 865-878.
- Gribaldo, S., and P. Cammarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J. Mol. Evol.* **47**: 508-516.
- Groll, M., L. Ditzel, J. Löwe, D. Stock, M. Bochtler, H. D. Bartunik, and R. Huber. 1997. Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* **386**: 463-471.
- Gupta, R. S. 1990. Sequence and structural homology between a mouse T-complex protein TCP- 1 and the 'chaperonin' family of bacterial (GroEL, 60-65 kDa heat shock antigen) and eukaryotic proteins. *Biochem. Int.* **20**: 833-841.
- Gupta, R. S. 1995. Phylogenetic analysis of the 90 kD heat shock family of protein sequences and an examination of the relationship among animals, plants, and fungi species. *Mol. Biol. Evol.* **12**: 1063-1073.
- Gutsche, I., L. O. Essen, and W. Baumeister. 1999. Group II chaperonins: new TRiC(k)s and turns of a protein folding machine. *J. Mol. Biol.* **293**: 295-312.
- Hartl, F. U. 1996. Molecular chaperones in cellular protein folding. *Nature* **381**: 571-579.
- Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Mol. Biol. Evol.* **11**: 65-71.
- Hemmingsen, S. M., C. Woolford, S. M. van der Vies, K. Tilly, D. T. Dennis, C. P. Georgopoulos, R. W. Hendrix, and R. J. Ellis. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* **333**: 330-334.
- Hendrick, J. P., and F. U. Hartl. 1993. Molecular chaperone functions of heat-shock proteins. *Ann. Rev. Biochem.* **62**: 349-384.
- Hendrix, R. W. 1979. Purification and properties of groE, a host protein involved in bacteriophage assembly. *J. Mol. Biol.* **129**: 375-392.

- Higgins, D. G., and P. M. Sharp. 1988. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**: 237-244.
- Hirt, R. P., B. Healy, C. R. Vossbrinck, E. U. Canning, and T. M. Embley. 1997. A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. *Curr. Biol.* **7**: 995-998.
- Hirt, R. P., J. M. Logsdon, Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc. Natl. Acad. Sci. USA* **96**: 580-585.
- Hofmann, C. J. B., S. A. Rensing, M. M. Hauber, W. F. Martin, S. B. Muller, J. Couch, G. I. McFadden, G. L. Igloi, and U.-G. Maier. 1994. Smallest eukaryotic genomes encode a protein gene: towards an understanding of nucleomorph functions. *Mol. Gen. Genet.* **243**: 600-604.
- Hohn, T., B. Hohn, A. Engel, M. Wurtz, and P. R. Smith. 1979. Isolation and characterization of the host protein groE involved in bacteriophage lambda assembly. *J. Mol. Biol.* **129**: 359-373.
- Horwich, A. L., and H. R. Saibil. 1998. The thermosome: chaperonin with a built-in lid. *Nat. Struct. Biol.* **5**: 333-336.
- Houry, W. A., D. Frishman, C. Eckerskorn, F. Lottspeich, and F. U. Hartl. 1999. Identification of *in vivo* substrates of the chaperonin GroEL. *Nature* **402**: 147-154.
- Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B Biol. Sci.* **256**: 119-124.
- Hughes, M. K., and A. L. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol. Biol. Evol.* **10**: 1360-1369.
- Hynes, G., H. Kubota, and K. R. Willison. 1995. Antibody characterisation of two distinct conformations of the chaperonin-containing TCP-1 from mouse testis. *FEBS Lett.* **358**: 129-132.
- Ishida, K., B. R. Green, and T. Cavalier-Smith. 2001. Endomembrane structure and the protein-targeting pathway to chloroplasts in the heterokont alga *Heterosigma akashiwo* (Raphidophyceae, Chromista). *J. Phycol.* (in press).
- Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. 1989. Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86**: 9355-9359.
- Jakob, U., and J. Buchner. 1994. Assisting spontaneity: the role of Hsp90 and small Hsps as molecular chaperones. *Trends Biochem. Sci.* **19**: 205-211.

- Johnson, J. L., and E. A. Craig. 1997. Protein folding *in vivo*: unraveling complex pathways. *Cell* **90**: 201-204.
- Kagawa, H. K., J. Osipiuk, N. Maltsev, R. Overbeek, E. Quait-Randall, A. Joachimiak, and J. D. Trent. 1995. The 60 kDa heat shock proteins in the hyperthermophilic archaeon *Sulfolobus shibatae*. *J. Mol. Biol.* **253**: 712-725.
- Kamaishi, T., T. Hashimoto, Y. Nakamura, F. Nakamura, S. Murata, N. Okada, K. Okamoto, M. Shimizu, and M. Hasegawa. 1996. Protein phylogeny of translation elongation factor EF-1 alpha suggests microsporidians are extremely ancient eukaryotes. *J. Mol. Evol.* **42**: 257-263.
- Kawarabayasi, Y., Y. Hino, H. Horikawa, S. Yamazaki, Y. Haikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, H. Nakazawa, M. Takamiya, S. Masuda, T. Funahashi, T. Tanaka *et al.* 1999. Complete genome sequence of anaerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6**: 83-101, 145-152.
- Kawarabayasi, Y., M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh *et al.* 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaeobacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5**: 55-76.
- Kawashima, T., N. Amano, H. Koike, S. Makino, S. Higuchi, Y. Kawashima-Ohya, K. Watanabe, M. Yamazaki, K. Kanehori, T. Kawamoto, T. Nunoshiba, Y. Yamamoto, H. Aramaki, K. Makino, and M. Suzuki. 2000. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. *Proc. Natl. Acad. Sci. USA* **97**: 14257-14262.
- Keeling, P. J., J. A. Deane, C. Hink-Schauer, S. E. Douglas, U.-G. Maier, and G. I. McFadden. 1999. The secondary endosymbiont of the cryptomonad *Guillardia theta* contains alpha-, beta-, and gamma-tubulin genes. *Mol. Biol. Evol.* **16**: 1308-1313.
- Keeling, P. J., and W. F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol. Biol. Evol.* **13**: 1297-1305.
- Keeling, P. J., M. A. Luker, and J. D. Palmer. 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol. Biol. Evol.* **17**: 23-31.
- Kim, S., K. R. Willison, and A. L. Horwich. 1994. Cytosolic chaperonin subunits have a conserved ATPase domain but diverged polypeptide-binding domains. *Trends Biochem. Sci.* **19**: 543-548.

- Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J. Mol. Evol.* **29**: 170-179.
- Klumpp, M., and W. Baumeister. 1998. The thermosome: archetype of group II chaperonins. *FEBS Lett.* **430**: 73-77.
- Klumpp, M., W. Baumeister, and L. O. Essen. 1997. Structure of the substrate binding domain of the thermosome, an archaeal group II chaperonin. *Cell* **91**: 263-270.
- Knapp, S., I. Schmidt-Krey, H. Hebert, T. Bergman, H. Jornvall, and R. Ladenstein. 1994. The molecular chaperonin TF55 from the thermophilic archaeon *Sulfolobus solfataricus*. A biochemical and structural characterization. *J. Mol. Biol.* **242**: 397-407.
- Kowalski, J. M., R. M. Kelly, J. Konisky, D. S. Clark, and K. D. Wittrup. 1998. Purification and functional characterization of a chaperone from *Methanococcus jannaschii*. *Syst. Appl. Microbiol.* **21**: 173-178.
- Kubota, H., G. Hynes, A. Carne, A. Ashworth, and K. Willison. 1994. Identification of six Tcp-1-related genes encoding divergent subunits of the TCP-1-containing chaperonin. *Curr. Biol.* **4**: 89-99.
- Kubota, H., G. Hynes, and K. Willison. 1995a. The chaperonin containing t-complex polypeptide 1 (TCP-1). Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. *Eur. J. Biochem.* **230**: 3-16.
- Kubota, H., G. Hynes, and K. Willison. 1995b. The eighth Cct gene, Cctq, encoding the theta subunit of the cytosolic chaperonin containing TCP-1. *Gene* **154**: 231-236.
- Kubota, H., G. M. Hynes, S. M. Kerr, and K. R. Willison. 1997. Tissue-specific subunit of the mouse cytosolic chaperonin-containing TCP-1. *FEBS Lett.* **402**: 53-56.
- Kwiatowski, J., M. Krawczyk, M. Kornacki, K. Bailey, and F. J. Ayala. 1995. Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene. *Proc. Natl. Acad. Sci. USA* **92**: 8503-8506.
- Lang, B. F., G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, and M. W. Gray. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**: 493-497.
- Lange, B. H. M., A. Bachi, M. Wilm, and C. Gonzalez. 2000. Hsp90 is a core centrosomal component and is required at different stages in the centrosome cycle in *Drosophila* and vertebrates. *EMBO J.* **19**: 1252-1262.

Laskey, R. A., B. M. Honda, A. D. Mills, and J. T. Finch. 1978. Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature* **275**: 416-420.

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

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

Leroux, M. R., and F. U. Hartl. 2000. Protein folding: versatility of the cytosolic chaperonin TriC/CCT. *Curr. Biol.* **10**: R260-R264.

Lewis, V. A., G. M. Hynes, D. Zheng, H. Saibil, and K. Willison. 1992. T-complex polypeptide-1 is a subunit of a heteromeric particle in the eukaryotic cytosol. *Nature* **358**: 249-252.

Lin, P., T. S. Cardillo, L. M. Richard, G. B. Segel, and F. Sherman. 1997. Analysis of mutationally altered forms of the Cct6 subunit of the chaperonin from *Saccharomyces cerevisiae*. *Genetics* **147**: 1609-1633.

Lin, P., and F. Sherman. 1997. The unique hetero-oligomeric nature of the subunits in the catalytic cooperativity of the yeast Cct chaperonin complex. *Proc. Natl. Acad. Sci. USA* **94**: 10780-10785.

Liou, A. K., E. A. McCormack, and K. R. Willison. 1998. The chaperonin containing TCP-1 (CCT) displays a single-ring mediated disassembly and reassembly cycle. *Biol. Chem.* **379**: 311-319.

Liou, A. K., and K. R. Willison. 1997. Elucidation of the subunit orientation in CCT (chaperonin containing TCP1) from the subunit composition of CCT micro-complexes. *EMBO J.* **16**: 4311-4316.

Llorca, O., J. Martin-Benito, M. Ritco-Vonsovici, J. Grantham, G. M. Hynes, K. R. Willison, J. L. Carrascosa, and J. M. Valpuesta. 2000. Eukaryotic chaperonin CCT stabilizes actin and tubulin folding intermediates in open quasi-native conformations. *EMBO J.* **19**: 5971-5979.

Llorca, O., E. A. McCormack, G. Hynes, J. Grantham, J. Cordell, J. L. Carrascosa, K. R. Willison, J. J. Fernandez, and J. M. Valpuesta. 1999a. Eukaryotic type II chaperonin CCT interacts with actin through specific subunits. *Nature* **402**: 693-696.

Llorca, O., M. G. Smyth, J. L. Carrascosa, K. R. Willison, M. Radermacher, S. Steinbacher, and J. M. Valpuesta. 1999b. 3D reconstruction of the ATP-bound

- form of CCT reveals the asymmetric folding conformation of a type II chaperonin. *Nat. Struct. Biol.* **6**: 639-642.
- Logsdon, J. M., Jr. 1998. The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **8**: 637-648.
- Logsdon, J. M., Jr., M. G. Tyshenko, C. Dixon, J. D.-Jafari, V. K. Walker, and J. D. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. *Proc. Natl. Acad. Sci. USA* **92**: 8507-8511.
- Lorimer, G. 2001. A personal account of chaperonin history. *Plant Physiol.* **125**: 38-41.
- Marco, S., D. Urena, J. L. Carrascosa, T. Waldmann, J. Peters, R. Hegerl, G. Pfeifer, H. Sack-Kongehl, and W. Baumeister. 1994. The molecular chaperone TF55. Assessment of symmetry. *FEBS Lett.* **341**: 152-155.
- Martin, J., M. Mayhew, and F.-U. Hartl. 1996. Role of prokaryotic chaperonins in protein folding. Pp. 213-244 in R. J. Ellis, ed. *The chaperonins*. Academic Press, San Diego.
- McFadden, G. I., P. R. Gilson, S. E. Douglas, T. Cavalier-Smith, C. J. Hofmann, and U.-G. Maier. 1997. Bonsai genomics: sequencing the smallest eukaryotic genomes. *Trends Genet.* **13**: 46-49.
- McKerracher, L., and S. P. Gibbs. 1982. Cell and nucleomorph division in the alga *Cryptomonas*. *Can. J. Bot.* **60**: 2440-2452.
- Melki, R., G. Batelier, S. Soulie, and R. C. Williams, Jr. 1997. Cytoplasmic chaperonin containing TCP-1: structural and functional characterization. *Biochemistry* **36**: 5817-5826.
- Meyer, S. 1987. The taxonomic implications of the ultrastructure and cell division of a stigma-containing *Chroomonas* sp. (Cryptophyceae) from Rocky Bay, Natal, South Africa. *S. Afr. J. Bot.* **53**: 129-139.
- Miklos, D., S. Caplan, D. Mertens, G. Hynes, Z. Pitluk, Y. Kashi, K. Harrison-Lavoie, S. Stevenson, C. Brown, B. Barrell, A. L. Horwich, and K. Willison. 1994. Primary structure and function of a second essential member of the heterooligomeric TCP1 chaperonin complex of yeast, TCP1 beta. *Proc. Natl. Acad. Sci. USA* **91**: 2743-2747.
- Minuth, T., G. Frey, P. Lindner, R. Rachel, K. O. Stetter, and R. Jaenicke. 1998. Recombinant homo- and hetero-oligomers of an ultrastable chaperonin from the archaeon *Pyrodictium occultum* show chaperone activity *in vitro*. *Eur. J. Biochem.* **258**: 837-845.

- Moreira, D., H. Le Guyader, and H. Phillippe. 2000. The origin of red algae and the evolution of chloroplasts. *Nature* **405**: 69-72.
- Morrall, S., and A. D. Greenwood. 1982. Ultrastructure of nucleomorph division in species of Cryptophyceae and its evolutionary implications. *J. Cell Sci.* **54**: 311-328.
- Morrison, H. G., A. J. Roger, T. G. Nystul, F. D. Gillin, and M. L. Sogin. 2001. *Giardia lamblia* expresses a proteobacterial-like DnaK homolog. *Mol. Biol. Evol.* **18**: 530-541.
- Nakamura, N., H. Taguchi, N. Ishii, M. Yoshida, M. Suzuki, I. Endo, K. Miura, and M. Yohda. 1997. Purification and molecular cloning of the group II chaperonin from the acidothermophilic archaeon, *Sulfolobus* sp. strain 7. *Biochem. Biophys. Res. Commun.* **236**: 727-732.
- Ng, W. V., S. P. Kennedy, G. G. Mahairas, B. Berquist, M. Pan, H. D. Shukla, S. R. Lasky, N. S. Baliga, V. Thorsson, J. Sbrogna, S. Swartzell, D. Weir, J. Hall, T. A. Dahl, R. Welti, Y. A. Goo, B. Leithauser, K. Keller, R. Cruz, M. J. Danson *et al.* 2000. Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. USA* **97**: 12176-12181.
- Nikaido, I., E. Asamizu, M. Nakajima, Y. Nakamura, N. Saga, and S. Tabata. 2000. Generation of 10,154 expressed sequence tags from a leafy gametophyte of a marine red alga, *Porphyra yezoensis*. *DNA Res.* **7**: 223-227.
- Nitsch, M., M. Klumpp, A. Lupas, and W. Baumeister. 1997. The thermosome: alternating alpha and beta-subunits within the chaperonin of the archaeon *Thermoplasma acidophilum*. *J. Mol. Biol.* **267**: 142-149.
- O'Kelly, C. J. 1993. The jakobid flagellates: structural features of *Jakoba*, *Reclinomonas* and *Histiona* and implications for the early diversification of eukaryotes. *J. Euk. Microbiol.* **40**: 627-636.
- O'Kelly, C. J. 1997. Ultrastructure of trophozoites, zoospores and cysts of *Reclinomonas americana* Flavin & Nerad, 1993 (Protista *incertae sedis*: Histonidae). *Eur. J. Protistol.* **33**: 337-348.
- O'Kelly, C. J., and T. A. Nerad. 1999. *Malawimonas jakobiformis* n. gen., n. sp. (Malawimonadidae n. fam.): a *Jakoba*-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. *J. Euk. Microbiol.* **46**: 522-531.
- Ohno, S. 1973. Ancient linkage groups and frozen accidents. *Nature* **244**: 259-262.
- Palmer, J. D. 1997. Genome evolution. The mitochondrion that time forgot. *Nature* **387**: 454-455.
- Palmer, J. D. 2000. A single birth of all plastids? *Nature* **405**: 32-33.



- Palmer, J. D., and J. M. Logsdon, Jr. 1991. The recent origins of introns. *Curr. Opin. Genet. Dev.* 1: 470-477.
- Patterson, D. J. 1990. *Jakoba libera* (Ruinen, 1938), a heterotrophic flagellate from deep oceanic sediments. *J. Mar. Biol. Ass. U.K.* 70: 381-393.
- Patterson, D. J., A. G. B. Simpson, and N. Weerakoon. 1999. Free-living flagellates from anoxic habitats and the assembly of the eukaryotic cell. *Biol. Bull.* 196: 381-384.
- Peyretailade, E., V. Broussolle, P. Peyret, G. Metenier, M. Gouy, and C. P. Vivares. 1998. Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein of mitochondrial evolutionary origin. *Mol. Biol. Evol.* 15: 683-689.
- Philippe, H., and A. Adoutte. 1998. The molecular phylogeny of Eukaryota: solid facts and uncertainties. Pp. 25-56 in G. H. Coombs, K. Vickerman, M. A. Sleigh and A. Warren, eds. *Evolutionary relationships among protozoa*. Chapman and Hall, London.
- Philippe, H., and P. Forterre. 1999. The rooting of the universal tree of life is not reliable. *J. Mol. Evol.* 49: 509-523.
- Philippe, H., and A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol. Biol. Evol.* 17: 830-834.
- Phipps, B. M., A. Hoffmann, K. O. Stetter, and W. Baumeister. 1991. A novel ATPase complex selectively accumulated upon heat shock is a major cellular component of thermophilic archaeobacteria. *EMBO J.* 10: 1711-1722.
- Phipps, B. M., D. Typke, R. Hegerl, S. Volker, A. Hoffmann, K. O. Stetter, and W. Baumeister. 1993. Structure of a molecular chaperone from a thermophilic archaeobacterium. *Nature* 361: 475-477.
- Piatigorsky, J., and G. Wistow. 1991. The recruitment of crystallins: new functions precede gene duplication. *Science* 252: 1078-1079.
- Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817-818.
- Ranson, N. A., H. E. White, and H. R. Saibil. 1998. Chaperonins. *Biochem. J.* 333: 233-242.
- Rensing, S. A., and U.-G. Maier. 1994. Phylogenetic analysis of the stress-70 protein family. *J. Mol. Evol.* 39: 80-86.
- Rensing, S. A., P. Obrdlik, N. Rober-Kleber, S. B. Müller, C. J. B. Hofmann, Y. van de Peer, and U.-G. Maier. 1997. Molecular phylogeny of the stress-70 protein family with reference to algal relationships. *Eur. J. Phycol.* 32: 279-285.

Ritco-Vonsovici, M., and K. R. Willison. 2000. Defining the eukaryotic cytosolic chaperonin-binding sites in human tubulins. *J. Mol. Biol.* **304**: 81-98.

Robbins, J., S. M. Dilworth, R. A. Laskey, and C. Dingwall. 1991. Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. *Cell* **64**: 615-623.

Rodriguez, F., J. F. Oliver, A. Marin, and J. R. Medina. 1990. The general stochastic model of nucleotide substitutions. *J. Theor. Biol.* **142**: 485-501.

Roger, A. J. 1999. Reconstructing early events in eukaryotic evolution. *The Am. Nat.* **154**(Suppl): S146-S163.

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

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

Rommelaere, H., M. Van Troys, Y. Gao, R. Melki, N. J. Cowan, J. Vandekerckhove, and C. Ampe. 1993. Eukaryotic cytosolic chaperonin contains t-complex polypeptide 1 and seven related subunits. *Proc. Natl. Acad. Sci. USA* **90**: 11975-11979.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**: 526-538.

Scheufler, C., A. Brinker, G. Bourenkov, S. Pegoraro, L. Moroder, H. Bartunik, F. U. Hartl, and I. Moarefi. 2000. Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell* **101**: 199-210.

Schornstein, K. L., and J. Scott. 1982. Ultrastructure and cell division in the unicellular red alga *Porphyridium purpureum*. *Can. J. Bot.* **60**: 85-97.

Sensen, C. W., R. L. Charlebois, C. Chow, I. G. Clausen, B. Curtis, W. F. Doolittle, M. Duguet, G. Erauso, T. Gaasterland, R. A. Garrett, P. Gordon, I. H. de Jong, A. C. Jeffries, C. Kozera, N. Medina, A. De Moors, J. van der Oost, H. Phan, M. A. Ragan, M. E. Schenk *et al.* 1998. Completing the sequence of the *Sulfolobus solfataricus* P2 genome. *Extremophiles* **2**: 305-312.

Shi, Y., D. D. Mosser, and R. I. Morimoto. 1998. Molecular chaperones as HSF1-specific transcriptional repressors. *Genes Dev.* **12**: 654-666.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**: 1114-1116.

Siegers, K., T. Waldmann, M. R. Leroux, K. Grein, A. Shevchenko, E. Schiebel, and F. U. Hartl. 1999. Compartmentation of protein folding *in vivo*: sequestration of non-native polypeptide by the chaperonin-GimC system. *EMBO J.* **18**: 75-84.

Siegert, R., M. R. Leroux, C. Scheufler, F. U. Hartl, and I. Moarefi. 2000. Structure of the molecular chaperone prefoldin: unique interaction of multiple coiled coil tentacles with unfolded proteins. *Cell* **103**: 621-632.

Sigler, P. B., Z. Xu, H. S. Rye, S. G. Burston, W. A. Fenton and A. L. Horwich. 1998. Structure and function in GroEL-mediated protein folding. *Annu. Rev. Biochem.* **67**: 581-608.

Silver, L. M., K. Artzt, and D. Bennett. 1979. A major testicular cell protein specified by a mouse T/t complex gene. *Cell* **17**: 275-284.

Simpson, A. G. B. 1999. The ultrastructure of *Carpodiemonas membranifera* (Eukaryota) with reference to the "excavate hypothesis". *Eur. J. Protistol.* **35**: 353-370.

Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski *et al.* 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. *J. Bacteriol.* **179**: 7135-7155.

Smith, M. W., S. B. Aley, M. Sogin, F. D. Gillin, and G. A. Evans. 1998. Sequence survey of the *Giardia lamblia* genome. *Mol. Biochem. Parasitol.* **95**: 267-280.

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

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

Srikakulam, R., and D. A. Winkelmann. 1999. Myosin II folding is mediated by a molecular chaperonin. *J. Biol. Chem.* **274**: 27265-27273.

Stanier, R. A. 1970. Some aspects of the biology of cells and their possible evolutionary significance. Pp. 1-38 in H. P. Charles and B. C. J. G. Knight, eds. *Organization and control of prokaryotic and eukaryotic cells: 20th symposium of the society for general microbiology*. Cambridge University Press, Cambridge.

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

- Stiller, J. W., E. C. Duffield, and B. D. Hall. 1998. Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. *Proc. Natl. Acad. Sci. USA* **95**: 11769-11774.
- Stiller, J. W., and B. D. Hall. 1997. The origin of red algae: implications for plastid evolution. *Proc. Natl. Acad. Sci. USA* **94**: 4520-4525.
- Stiller, J. W., and B. D. Hall. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol. Biol. Evol.* **16**: 1270-1279.
- Stoldt, V., F. Rademacher, V. Kehren, J. F. Ernst, D. A. Pearce, and F. Sherman. 1996. The Cct eukaryotic chaperonin subunits of *Saccharomyces cerevisiae* and other yeasts. *Yeast* **12**: 523-529.
- Strimmer, K., and A. von Haeseler. 1997. PUZZLE. Zoologisches Institut, Universitat Muenchen, Germany.
- Swofford, D. L. 1998. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods). Sinauer Associates, Sunderland, Massachusetts.
- Takai, K., and K. Horikoshi. 1999. Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* **152**: 1285-1297.
- Takai, K., and Y. Sako. 1999. A molecular view of archaeal diversity in marine and terrestrial hot water environments. *FEMS Microbiol. Ecol.* **28**: 177-188.
- Thompson, D. K., and C. J. Daniels. 1998. Heat shock inducibility of an archaeal TATA-like promoter is controlled by adjacent sequence elements. *Mol. Microbiol.* **27**: 541-551.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673-4680.
- Thulasiraman, V., C. F. Yang, and J. Frydman. 1999. *In vivo* newly translated polypeptides are sequestered in a protected folding environment. *EMBO J.* **18**: 85-95.
- Trent, J. D., E. Nimmesgern, J. S. Wall, F. U. Hartl, and A. L. Horwich. 1991. A molecular chaperone from a thermophilic archaeobacterium is related to the eukaryotic protein t-complex polypeptide-1. *Nature* **354**: 490-493.
- Trent, J. D., J. Osipiuk, and T. Pinkau. 1990. Acquired thermotolerance and heat shock in the extremely thermophilic archaeobacterium *Sulfolobus* sp. strain B12. *J. Bacteriol.* **172**: 1478-1484.

- Ursic, D., and M. R. Culbertson. 1991. The yeast homolog to mouse Tcp-1 affects microtubule-mediated processes. *Mol. Cell. Biol.* **11**: 2629-2640.
- Ursic, D., and B. Ganetzky. 1988. A *Drosophila melanogaster* gene encodes a protein homologous to the mouse t complex polypeptide 1. *Gene* **68**: 267-274.
- Ursic, D., J. C. Sedbrook, K. L. Himmel, and M. R. Culbertson. 1994. The essential yeast Tcp1 protein affects actin and microtubules. *Mol. Biol. Cell* **5**: 1065-1080.
- Vainberg, I. E., S. A. Lewis, H. Rommelaere, C. Ampe, J. Vandekerckhove, H. L. Klein, and N. J. Cowan. 1998. Prefoldin, a chaperone that delivers unfolded proteins to cytosolic chaperonin. *Cell* **93**: 863-873.
- Voight, M. 1901. Mitteilungen aus der Biolog. Station Plön, Holstein-Uber einige bisher unbekannte Süßwasserorganismen. *Zool. Az.* **24**: 191-195.
- Vossbrinck, C. R., J. V. Maddox, S. Friedman, B. A. Debrunner-Vossbrinck, and C. R. Woese. 1987. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature* **326**: 411-414.
- Vossbrinck, C. R., and C. R. Woese. 1986. Eukaryotic ribosomes that lack a 5.8S RNA. *Nature* **320**: 287-288.
- Waldmann, T., A. Lupas, J. Kellermann, J. Peters, and W. Baumeister. 1995a. Primary structure of the thermosome from *Thermoplasma acidophilum*. *Biol. Chem. Hoppe Seyler* **376**: 119-126.
- Waldmann, T., E. Nimmesgern, M. Nitsch, J. Peters, G. Pfeifer, S. Muller, J. Kellermann, A. Engel, F. U. Hartl, and W. Baumeister. 1995b. The thermosome of *Thermoplasma acidophilum* and its relationship to the eukaryotic chaperonin TRiC. *Eur. J. Biochem.* **227**: 848-856.
- Wang, S. L., and X.-Q. Liu. 1991. The plastid genome of *Cryptomonas*  $\Phi$  encodes an hsp70-like protein, a histone-like protein, and an acyl carrier protein. *Proc. Natl. Acad. Sci. USA* **88**: 10783-10787.
- Wastl, J., M. Fraunholz, S. Zauner, S. Douglas, and U.-G. Maier. 1999. Ancient gene duplication and differential gene flow in plastid lineages: the GroEL/Cpn60 example. *J. Mol. Evol.* **48**: 112-117.
- Willison, K., A. Kelly, K. Dudley, P. Goodfellow, N. Spurr, V. Groves, P. Gorman, D. Sheer, and J. Trowsdale. 1987. The human homologue of the mouse t-complex gene, TCP1, is located on chromosome 6 but is not near the HLA region. *EMBO J.* **6**: 1967-1974.
- Willison, K. R. 1999. Composition and function of the eukaryotic cytosolic chaperonin-containing TCP-1. Pp. 555-571 in B. Bukau, ed. *Molecular chaperones and folding catalysts*. Harwood, Amsterdam.

- Willison, K. R., K. Dudley, and J. Potter. 1986. Molecular cloning and sequence analysis of a haploid expressed gene encoding t complex polypeptide 1. *Cell* **44**: 727-738.
- Willison, K. R., and J. Grantham. 2001. The roles of the cytosolic chaperonin, CCT, in normal eukaryotic cell growth. Pp. 90-118 in P. Lund, ed. *Molecular chaperones: frontiers in molecular biology*. Oxford University Press, Oxford, U.K.
- Willison, K. R., and A. L. Horwich. 1996. Structure and function of chaperonins in archaeobacteria and eukaryotic cytosol. Pp. 107-136 in R. J. Ellis, ed. *The chaperonins*. Academic Press, San Diego.
- Willison, K. R., and H. Kubota. 1994. The structure, function, and genetics of the chaperonin containing TCP-1 (CCT) in eukaryotic cytosol. Pp. 299-312 in R. I. Morimoto, A. Tissieres, and C. Georgopoulos, eds. *The biology of heat shock proteins and molecular chaperones*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- Woese, C. R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**: 221-271.
- Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain. The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**: 5088-5090.
- Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a natural system of organisms, proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**: 4576-4579.
- Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**: 708-713.
- Won, K. A., R. J. Schumacher, G. W. Farr, A. L. Horwich, and S. I. Reed. 1998. Maturation of human cyclin E requires the function of eukaryotic chaperonin CCT. *Mol. Cell. Biol.* **18**: 7584-7589.
- Xu, Z., A. L. Horwich, and P. B. Sigler. 1997. The crystal structure of the asymmetric GroEL-GroES-(ADP)<sub>7</sub> chaperonin complex. *Nature* **388**: 741-750.
- Yaffe, M. B., G. W. Farr, D. Miklos, A. L. Horwich, M. L. Sternlicht, and H. Sternlicht. 1992. TCP1 complex is a molecular chaperone in tubulin biogenesis. *Nature* **358**: 245-248.
- Yoshida, T., M. Yohda, T. Iida, T. Maruyama, H. Taguchi, K. Yazaki, T. Ohta, M. Odaka, I. Endo, and Y. Kagawa. 1997. Structural and functional characterization of homo-oligomeric complexes of alpha and beta chaperonin subunits from the hyperthermophilic archaeum *Thermococcus* strain KS-1. *J. Mol. Biol.* **273**: 635-645.

- Yoshida, T., M. Yohda, M. Suzuki, K. Yazaki, K. Miura, and I. Endo. 1998. Characterization of homo-oligomeric complexes of alpha and beta chaperonin subunits from the acidothermophilic archaeon, *Sulfolobus* sp. strain 7. *Biochem. Biophys. Res. Commun.* **242**: 640-647.
- Zarzov, P., H. Boucherie, and C. Mann. 1997. A yeast heat shock transcription factor (Hsf1) mutant is defective in both Hsc82/Hsp82 synthesis and spindle pole body duplication. *J. Cell Sci.* **110**: 1879-1891.
- Zauner, S., M. Fraunholz, J. Wastl, S. Penny, M. Beaton, T. Cavalier-Smith, U.-G. Maier, and S. Douglas. 2000. Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. *Proc. Natl. Acad. Sci. USA* **97**: 200-205.
- Zuo, J., R. Baler, G. Dahl, and R. Voellmy. 1994. Activation of the DNA-binding ability of human heat shock transcription factor 1 may involve the transition from an intramolecular to an intermolecular triple-stranded coiled-coil structure. *Mol. Cell. Biol.* **14**: 7557-7568.
- Zuo, J., Y. Guo, T. Guettouche, D. F. Smith, and R. Voellmy. 1998. Repression of heat shock transcription factor HSF1 activation by HSP90 (HSP90 complex) that forms a stress-sensitive complex with HSF1. *Cell* **94**: 471-480.