Contrasting Evolutionary Rates in the Duplicate Chaperonin Genes of *Mycobacterium tuberculosis* and *M. leprae*

*Austin L. Hughes*

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University

A phylogenetic analysis of chaperonin (heat shock protein 60) sequences from prokaryotes and eukaryotes indicated that a single gene duplication event in the common ancestor of *Mycobacterium tuberculosis*, *M. leprae*, and *Streptomyces albus* gave rise to the duplicate chaperonin genes found in these species (designated *HSP65* and *GroEL* in the mycobacterial species). Comparison of rates of synonymous and nonsynonymous nucleotide substitution in different gene regions suggested that the 5' end of the *HSP65* gene was homogenized by an ancient recombination event between *M. tuberculosis* and *M. leprae*. In *S. albus*, the two duplicated chaperonin genes have evolved at essentially the same rate. In both *M. tuberculosis* and *M. leprae*, however, the *GroEL* gene has evolved considerably more rapidly at nonsynonymous nucleotide sites than has the *HSP65* gene. Because this difference is not seen at synonymous sites, it must be due to a difference in selective constraint on the proteins encoded by the two genes, rather than to a difference in mutation rate. The difference between *GroEL* and *HSP65* is striking in regions containing epitopes recognized by T cells of the vertebrate host; in certain cross-reactive epitopes conserved across all organisms, nonsynonymous sites in *GroEL* have evolved twice as fast as those in *HSP65*. It is suggested that these differences are correlated with differences in the way in which the duplicate chaperonins of *M. tuberculosis* and *M. leprae* interact with the host immune system.

Introduction

The chaperonin or heat-shock protein 60 (HSP60) family is a family of highly conserved chaperone proteins whose members are found both in prokaryotes and in eukaryotic organelles and function to mediate the folding of polypeptides (Ellis and van der Vies 1991; Welch 1991). Along with the heat-shock protein 70 (HSP70) and heat-shock protein 90 families, chaperonins are classified as stress proteins because they are typically expressed at high levels in response to environmental stress such as heat shock. In most bacterial species that have been studied so far, there is a single member of this gene family. However, there are two expressed chaperonin genes in *Mycobacterium tuberculosis* and in *M. leprae*, the causative organisms of tuberculosis and leprosy, respectively. The product of one of these two genes, generally called the “65-kDa HSP” or “HSP65,” is a major target for the immune response of the vertebrate host to these pathogens (Young et al. 1988; Thole and van der Zee 1990). In fact,
stress proteins of a wide variety of organisms parasitic on vertebrates have been found to stimulate an immune response by the host (Young and Elliott 1989; Young 1990), but the evolutionary consequences of this immune recognition by the host are so far poorly understood.

Certain proteins that serve as targets for the vertebrate immune system have evolved quite rapidly, apparently under selection to evade immune recognition. Indeed, some of these may even have no other function aside from stimulating an immune response (Enea and Arnot 1988; Sher 1988; Hughes 1991, 1992). The chaperonins, by contrast, are, in general, highly conserved across all organisms, presumably because they are subject to strong functional constraints (Thole and van der Zee 1990). Thus mycobacterial chaperonins may be subject to two opposing selective pressures—namely, that arising from functional constraints on the amino acid sequence and that arising from the need to evade host immune recognition (Hughes 1993b). To understand how mycobacterial chaperonins have evolved in response to these selective pressures, I analyze the evolutionary relationships of chaperonins from *Mycobacterium* and other organisms and compare the rates of evolution of these molecules at both the amino acid and nucleotide levels.

**Methods**

**DNA Sequences Analyzed**

In many bacteria, chaperonins have been found to be encoded in an operon (*GroES/L*) that includes two genes, called “*GroES*” and “*GroEL*.” The *GroES* gene encodes a polypeptide of ~10 kDa, while the *GroEL* gene encodes a polypeptide of ~60 kDa; but the *GroES* protein is similar to the N-terminal region of *GroEL*. The *GroEL* protein forms an oligomer of 14 subunits arranged as two rings of seven subunits each, while *GroES* forms a ring of seven subunits. These two proteins are believed to interact functionally (Welch 1991). In eukaryotes, there is evidence of a protein homologous to *GroES* (Laben et al. 1990), but so far sequence data are only available for eukaryotic homologues of *GroEL*.

Table 1 lists the sources for chaperonin DNA and amino acid sequences used in analyses. For 18 of the 22 bacterial chaperonin sequences, a *GroES* sequence from the same *GroES/L* operon was also analyzed. As mentioned previously, two chaperonins are known from *Mycobacterium tuberculosis* and *A. leprae*. One is the *GroEL* encoded in the *GroES/L* operon, while the other is designated “HSP65.” An HSP65 sequence from *M. hovis* (Thole et al. 1987) is identical, at the amino acid level, to that of *M. tuberculosis* and is not included in the analyses reported here. In *Streptomyces albus*, there are two chaperonin genes, designated “1” and “2”; an accompanying *GroES* sequence is available for the former only. By contrast, in *Rhizobium meliloti*, sequences are available for two separate *GroES/L* operons, designated “A” and “C.”

Mitochondrially expressed chaperonin sequences from human, mouse, yeast, and two plants were used in phylogenetic tree construction (table 1). The chaperonin homologue in chloroplasts is designated “rubisco-binding protein” (RBP), because of its role in binding the enzyme rubisco, which is involved in CO₂ fixation; there are two types of chains, designated “RBPa” and “RBPb” (Hemmingsen et al. 1988).

**Statistical Analyses**

Chaperonin amino acid sequences were aligned by the CLUSTAL V program (Higgins et al. 1992). The alignments (which are available on request) are essentially identical to those produced with more limited data sets by Thole and van der Zee.
Table 1
Chaperonin Sequences Used in Analyses

<table>
<thead>
<tr>
<th>Sequence</th>
<th>GenBank Accession No. (Reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>X07850 (Hemningsen et al. 1988)</td>
</tr>
<tr>
<td>Haemophilus ducreyi</td>
<td>M10130</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>M63957</td>
</tr>
<tr>
<td>Corynebacterium bontii</td>
<td>M26842 (Vodkin and Williams 1988)</td>
</tr>
<tr>
<td>Lepotella micaladi</td>
<td>X57520 (Hinderson et al. 1991)</td>
</tr>
<tr>
<td>Rhizobium meliloti A</td>
<td>M94192</td>
</tr>
<tr>
<td>Rhizobium meliloti C</td>
<td>M84190</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>M89930 (Lin et al. 1992)</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>M33301 (Johnson et al. 1989)</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Z11547 (Prasad and Stewart 1992)</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Z11546 (Prasad and Stewart 1992)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>M34664 (Venner et al. 1990)</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>X53584 (Venner and Gupta 1990)</td>
</tr>
<tr>
<td>Chlamydia trachomatis</td>
<td>M38027 (Cerrone et al. 1991)</td>
</tr>
<tr>
<td>Chlamydia pneumoniae</td>
<td>M69217 (Kikutani et al. 1991)</td>
</tr>
<tr>
<td>Chlamydia philaearum</td>
<td>X51404 M25101 (Morrison et al. 1989)</td>
</tr>
<tr>
<td>Borrelia burgdorferi</td>
<td>X56139</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>M84065 (Schmidt et al. 1992)</td>
</tr>
<tr>
<td>Coelomis perffinorum</td>
<td>X62914</td>
</tr>
<tr>
<td>Trichomon ascalium RB480</td>
<td>X07851 (Hemningsen et al. 1988)</td>
</tr>
<tr>
<td>Brassica napus RBPs</td>
<td>M35599 (Martel et al. 1990)</td>
</tr>
<tr>
<td>Brassica napus RB9</td>
<td>M35600 (Martel et al. 1990)</td>
</tr>
<tr>
<td>Sarcina lutea sp.</td>
<td>M27517 (Chirvas and Nelson 1991)</td>
</tr>
<tr>
<td>Syntrophus sp.</td>
<td>M58751 (Webb et al. 1990)</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis GroEL</td>
<td>X60305 (Shinnick 1987)</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis HSP65</td>
<td>M15467 (Shinnick 1987)</td>
</tr>
<tr>
<td>Mycobacterium leprae GroEL</td>
<td>Z11665 (Rinke de Wit et al. 1992)</td>
</tr>
<tr>
<td>Mycobacterium leprae HSP5</td>
<td>M14341 (Machra et al. 1988)</td>
</tr>
<tr>
<td>Streptococcus albus 1</td>
<td>M76567 (Mazodier et al. 1991)</td>
</tr>
<tr>
<td>Streptococcus albus 2</td>
<td>M76658 (Mazodier et al. 1991)</td>
</tr>
</tbody>
</table>

(1990). Cerrone et al. (1991), and others. In pairwise comparisons among the members of any set of amino acid or DNA sequences, any position at which the alignment postulated a gap in any sequence in the set was excluded from all comparisons. In pairwise comparisons of all amino acid sequences, 492 aligned residues were compared, and the number of amino acid replacements per site (d_{aa}) was estimated by the Poisson-correction formula (Nei 1987, p. 41).

This distance was used to estimate the minimum-evolution (ME) tree by Rzhetsky and Nei's (1992a) method. The ME tree is the tree with the minimum total sum of branch lengths; and both computer simulation (Saitou and Imanishi 1989) and analytical study (Rzhetsky and Nei 1992b) indicate that this method has a high probability of recovering the true phylogenetic relationships. This method outperforms most others when the rate of evolution differs in different lineages (Saitou and Imanishi 1989), as is often true in the case of multigene families. In searching for the ME tree, first a neighbor-joining (NJ) tree (Saitou and Nei 1987) was constructed; then all topologies with topological distance of 2 or 4 from the NJ tree were searched; then trees based on 1,000 bootstrap samples of the data were searched. In addition to ME trees based on d_{aa}, I also constructed (a) ME trees based on the proportion of amino acid difference
and (b) NJ trees based on the proportion of nonsynonymous nucleotide differences per site (Nei and Gojobori 1986) and on the number of nonsynonymous nucleotide substitutions per site \( (d_N) \), corrected by Jukes and Cantor's (1969) formula. Because all these methods produced essentially identical trees, in the following I only present ME trees based on \( d_{aa} \). Rzhetsky and Nei (1992a) also provide a method for estimating standard errors (SEs) for branch lengths within ME trees. These can be used to test the statistical significance of certain patterns within a phylogenetic tree. They can also be used to compare the rate of evolution in different branches of the tree.

Several experimental studies have identified T cell epitopes (regions containing peptides presented by major-histocompatibility-complex molecules to T cells) in the HSP65 of mycobacterial species (Lamb et al. 1987; Anderson et al. 1988; Thole et al. 1988; Van Schooten et al. 1989). These were located in four regions of the protein, which in this paper are designated “epitope regions 1-4” (ER1-ER4) (fig. 1). In addition, Munk et al. (1989) found that T cells from healthy humans respond to epitopes from the human chaperonin that are highly conserved in most chaperonins, including mycobacterial HSP65. These regions, designated in this paper as “conserved epitopes,” are illustrated in figure 1. To study the evolutionary effects of immune recognition of these epitopes, I computed both the number of synonymous nucleotide substitutions per site \( (d_S) \) and \( d_N \) in comparisons among mycobacterial and selected other chaperonin genes.

In estimating \( d_S \) and \( d_N \), I first estimated, by Nei and Gojobori's (1986) method, the proportions of synonymous and nonsynonymous nucleotide differences per site and then corrected for multiple hits by two methods: Jukes and Cantor's (1969) and Bulmer et al.'s (1991). In applying the latter method, I computed proportions of the four nucleotides separately for synonymous and nonsynonymous sites, counting fractional sites as in the case of Nei and Gojobori's method. Jukes and Cantor's method assumes equal use of the four nucleotides, whereas Bulmer et al.'s method does not. The assumption of equal use of the four nucleotides appeared to be approximately valid in the case of nonsynonymous sites, and thus only \( d_N \) values corrected by Jukes and Cantor's method are reported. Streptomyces albus I and S. albus 2 were found to have an extraordinarily high percentage (95.7% and 94.3%, respectively), of G+C at synonymous sites. In this case, the Jukes-Cantor correction would be expected to underestimate the rate of synonymous substitution per site; therefore, in comparisons involving these genes, Bulmer et al.'s correction was used.

Results

Gene Phylogenies

The ME tree for chaperonins is shown in figure 2. The chaperonins form two major subfamilies (designated “I” and “II” in fig. 2) separated by a statistically significant internal branch. The subfamily I sequences include GroEL from Escherichia coli and related purple bacteria, GroEL of Chlamydia and Borrelia, and mitochondrial chaperonins of eukaryotes. The subfamily II sequences include those of Bacillus, Clostridium, the cyanobacteria, Mycobacterium, Streptomyces, and RBPα and RBPβ from plant chloroplasts. This phylogeny is consistent with the hypothesis that chloroplasts were derived from endosymbiotic Cyanobacteria whereas mitochondria were derived from endosymbiotic purple bacteria (Woese 1991).

The chaperonins of Mycobacterium and of S. albus form a clade of sequences that is separated from other members of this family by a significant internal branch (fig. 2). The topology of the tree within this cluster supports the hypothesis that a
FIG. 1.—Amino acid sequences of duplicated chaperoning genes of *Mycobacterium tuberculosis* and *M. leprae*. A dash indicates an amino acid identical to that in the top row; an asterisk indicates a gap. ER1–ER4 are overlined. Highly conserved, cross-reactive epitopes (Munk et al. 1989) are indicated by a plus sign.
FIG. 2.—ME tree based on $d_{aa}$ among prokaryotic and eukaryotic organelle chaperonins. Sequences are grouped into two subfamilies, I and II. Results of tests of the hypothesis that internal branch length is equal to zero are as follows: * = $P < 0.05$; ** = $P < 0.01$; and *** = $P < 0.001$. 
single event of gene duplication, occurring in the common ancestor of these three
species, gave rise to duplicate chaperonin genes. This hypothesis is supported by the
statistically significant internal branch grouping the mycobacterial HSP65 with chap-
eronin 2 of *S. albus* (fig. 2).

To obtain further evidence regarding the relationships of chaperonin genes of
*Mycobacterium* and *S. albus*, I estimated $d_S$ and $d_N$ in pairwise comparisons among
these genes (table 2). The pattern of nucleotide substitution at nonsynonymous sites
supports essentially the same groupings of genes as is shown by the tree based on
amino acid replacements per site (fig. 2). At nonsynonymous sites, *S. albus GroEL 2*
is significantly more similar to mycobacterial *HSP65* than it is to *S. albus GroEL 1*
(table 2).

At synonymous sites, the pattern was somewhat different. The values of $d_S$ between
*S. albus 2* and mycobacterial *HSP65* were lower than those between *S. albus 2* and
mycobacterial *GroEL*; but in these comparisons synonymous sites were nearly
saturated, and these differences were not statistically significant (table 2). By contrast, $d_S$
between *S. albus 2* and *S. albus 1* was significantly lower than that between *S. albus 2*
and any of the mycobacterial genes (table 2). This pattern is not expected if the
phylogeny of figure 2 is correct. Explanations for low $d_S$ between the two *S. albus*
genes include the following: (1) Bulmer et al.'s (1991) correction formula may not
provide an adequate correction in the case of genes with as extreme a G + C content
bias at synonymous sites and/or as great a number of nucleotide differences per site
as was seen in this case. (2) The high G + C content in *S. albus* may reflect a biased
pattern of mutation in this species, leading to a reduced substitution rate (Wolfe et
al. 1989). (3) The rate of synonymous substitution may not reflect the mutation rate
in *S. albus* because purifying selection on codon usage (Sharp and Li 1986) eliminates
most mutations to A or T at synonymous sites. In the present case, it is not possible
to decide among these explanations.

Rates of Amino Acid Evolution

Terminal branch lengths within the phylogenetic tree of chaperonins (fig. 2) vary
considerably, suggesting that different members of this family may have evolved at
different rates at the amino acid level. The hypothesis of a constant rate was tested
separately for different clusters within the tree by comparing within-tree distances
from certain nodes representing the common ancestors of groups of genes. Although
the tree itself is unrooted, clusters within the tree are rooted by the remainder
of the tree.

Node 1 in figure 2 represents the common ancestor of the eukaryotic mitochondri-
al chaperonins and of the most closely related bacterial genes, i.e., those of *E. coli*
and related purple bacteria. Distances from node 1 to GroEL of *E. coli* and related
bacteria are markedly less than those to the human mitochondrial chaperonin, whereas
the distances to eukaryotic mitochondrial chaperonins do not differ greatly from one
another. For example, the distance from node 1 to *E. coli* GroEL is 0.215 ± 0.023,
whereas that to human chaperonin is 0.407 ± 0.033; these two distances are significantly
different at the 0.1% level.

Node 2 in figure 2 represents the common ancestor of cyanobacterial GroEL and
of plant chloroplast chaperonins. The distances from node 2 to plant chloroplast RBPα
and RBPβ are consistently greater than are those to cyanobacterial GroEL. For ex-
ample, the distance from node 2 to wheat RBPα is 0.360 ± 0.035, whereas that to *Synechococcus*
GroEL is 0.236 ± 0.024; these two distances are significantly different
Table 2

d_s ± SE (above the Diagonal) and d_N ± SE (below the Diagonal), per 100 Sites, in Pairwise Comparisons of Chaperonin Genes from *Mycobacterium* and *Streptomyces albus*.

<table>
<thead>
<tr>
<th></th>
<th><em>M. tuberculosis</em> GroEL</th>
<th><em>M. leprae</em> GroEL</th>
<th><em>S. albus</em> 1 HSP65</th>
<th><em>M. tuberculosis</em> HSP65</th>
<th><em>M. leprae</em> HSP65</th>
<th><em>S. albus</em> 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>M. tuberculosis</em> GroEL</td>
<td>10.1 ± 1.0</td>
<td>83.5 ± 6.6</td>
<td>115.9 ± 9.2</td>
<td>113.3 ± 11.3</td>
<td>157.6 ± 41.4</td>
<td>125.5 ± 10.6*</td>
</tr>
<tr>
<td><em>M. leprae</em> GroEL</td>
<td>31.8 ± 2.0</td>
<td>32.6 ± 2.0</td>
<td>120.8 ± 10.9</td>
<td>114.3 ± 12.8</td>
<td>151.9 ± 42.9</td>
<td>131.5 ± 12.9*</td>
</tr>
<tr>
<td><em>S. albus</em> 1</td>
<td>33.6 ± 2.0</td>
<td>33.4 ± 2.0</td>
<td>214 ± 1.8</td>
<td>85.2 ± 5.3</td>
<td>64.5 ± 6.6</td>
<td>93.2 ± 6.0*</td>
</tr>
<tr>
<td><em>M. tuberculosis</em> HSP65</td>
<td>34.6 ± 2.1</td>
<td>33.5 ± 2.0</td>
<td>27.2 ± 1.8</td>
<td>3.5 ± 0.6</td>
<td>113.1 ± 11.3</td>
<td>113.1 ± 11.3*</td>
</tr>
<tr>
<td><em>M. leprae</em> HSP65</td>
<td>35.3 ± 2.1*</td>
<td>34.5 ± 2.1*</td>
<td>23.8 ± 1.6</td>
<td>11.8 ± 1.1*</td>
<td>11.7 ± 1.1*</td>
<td></td>
</tr>
</tbody>
</table>

*NOTE.*—Data are based on 528 aligned codons. Bulmer et al.'s (1991) correction formula was used in estimating *d_s*. Jukes and Cantor's (1969) correction formula was used in estimating *d_N*.

*a* Tests of the hypothesis that *d_s* between *S. albus* 2 and another gene is different from *d_s* between *S. albus* 2 and *S. albus* 1: *P* < 0.001.

*b* Tests of the hypothesis that *d_N* between *S. albus* 2 and another gene is different from *d_N* between *S. albus* 2 and *S. albus* 1: *P* < 0.001.
at the 1% level. The evidence that chaperonins of the eukaryotic organelles have evolved more rapidly at the amino acid level than have related bacterial chaperonins is consistent with the hypothesis that there has been a change of evolutionary constraint on chaperonins of organelles, in comparison with that on related bacterial chaperonins.

Node 3 in figure 2 represents the common ancestor of the chaperonins of Mycobacterium and of Streptomyces albus. The distances from node 3 to GroEL of M. tuberculosis and M. leprae are greater than that to mycobacterial HSP65 or to either Streptomyces albus chaperonin, but the distances from node 3 to both GroEL of Mycobacterium are approximately the same. From node 3 to M. leprae GroEL the distance is 0.298 ± 0.028, whereas to M. leprae HSP65 it is only 0.182 ± 0.018; these two distances are significantly different at the 0.1% level. Thus, mycobacterial GroEL have evolved more rapidly at the amino acid level than have both the presumably orthologous Streptomyces albus chaperonin 1 and the paralogous mycobacterial HSP65 and Streptomyces albus 2.

Rates of Nucleotide Substitution

To understand the factors responsible for the rapid rate of amino acid evolution of mycobacterial GroEL, I estimated \( d_S \) and \( d_N \) between the two mycobacterial GroEL genes and between the two mycobacterial HSP65 genes, separately for the four epitope regions (ER1-ER4) and for the remainder of the gene (table 3). For both orthologous pairs of genes, \( d_S \) was significantly greater than \( d_N \), in all regions analyzed (table 3), indicating that these genes are subject to purifying selection, which eliminates most nonsynonymous mutations. However, there was evidence that, with respect to the strength of purifying selection, there are differences between the two orthologous pairs of genes. In ER2, in ER3, and in the remainder of the gene (i.e., outside the epitope regions), \( d_N \) between the two HSP65 genes was significantly lower than that between the two GroEL genes (table 3), indicating that purifying selection is acting more strongly on the former genes in these regions. These analyses suggest that the rapid rate of amino acid level evolution in mycobacterial GroEL (fig. 2) is due to reduced

<table>
<thead>
<tr>
<th>GroEL</th>
<th>HSP65</th>
</tr>
</thead>
<tbody>
<tr>
<td>dS</td>
<td>dN</td>
</tr>
<tr>
<td>ER1</td>
<td>69.9 ± 13.7</td>
</tr>
<tr>
<td>ER2</td>
<td>129.5 ± 48.2</td>
</tr>
<tr>
<td>ER3</td>
<td>87.7 ± 23.6</td>
</tr>
<tr>
<td>ER4</td>
<td>74.6 ± 21.0</td>
</tr>
<tr>
<td>Remainder</td>
<td>67.3 ± 8.4</td>
</tr>
</tbody>
</table>

Note.—Jukes and Cantor's (1969) correction formula was used to estimate \( d_S \) and \( d_N \).

a Test of hypothesis that \( d_S = d_N \): \( P < 0.001 \).
b Test of hypothesis that \( d_S \) or \( d_N \) equals that in remainder: \( P < 0.001 \).
c Test of hypothesis that \( d_S \) or \( d_N \) equals that in GroEL: \( P < 0.05 \).
d Test of hypothesis that \( d_S \) or \( d_N \) equals that in the remainder: \( P < 0.01 \).
e Test of hypothesis that \( d_S = d_N \): \( P < 0.05 \).
f Test of hypothesis that \( d_S = d_N \): \( P < 0.01 \).
g Test of hypothesis that \( d_S \) or \( d_N \) equals that in GroEL: \( P < 0.01 \).
h Test of hypothesis that \( d_S \) or \( d_N \) equals that in GroEL: \( P < 0.001 \).
constraint at the amino acid level in certain regions; furthermore, since, in all regions but ER4, comparisons between \textit{M. tuberculosis} and \textit{M. leprae} show a higher rate of nonsynonymous substitution in \textit{GroEL} than in \textit{HSP65} (table 3), these results indicate that the more rapid evolution of \textit{GroEL} has continued after divergence of these two species.

There is also evidence that, within each pair of orthologous genes, there are marked differences among regions with respect to the degree of constraint at the amino acid level. In the case of \textit{GroEL} genes, \(d_N\) in ER1 was significantly lower than that in the remainder of the gene; but in the three other epitope regions, \(d_N\) was not significantly different from that in the remainder of the gene (table 3). By contrast, in the case of \textit{HSP65}, \(d_N\) in the remainder of the gene was not significantly different from that in the epitope regions, except for ER4, in which \(d_N\) was significantly higher than in the remainder of the gene (table 3). When the duplicate chaperonin genes of \textit{M. tuberculosis}, \textit{M. leprae}, and \textit{Streptomyces albus} were compared with each other (table 4), ER1 and ER2 were found to be the most conserved regions, as indicated by the lowest values of \(d_N\). However, between the two \textit{Mycobacterium} species and \textit{Streptomyces albus}, there was a difference in the extent of this conservation. In ER1 and ER2, \(d_N\) between the two \textit{Streptomyces albus} genes was significantly lower than that between the duplicate genes of either \textit{Mycobacterium} species (table 4). \textit{Mycobacterium tuberculosis} differed from both other species in that, in this species, ER3 and ER4 were also highly conserved between the two duplicated genes (table 4).

In all regions analyzed, except ER1, \(d_S\) between the two mycobacterial \textit{GroEL} genes was not significantly different from that between the two \textit{HSP65} genes. This was seen whether Jukes and Cantor’s (table 3) or Bulmer et al.‘s (data not shown) correction formula was used. This supports the hypothesis that the reason for the difference in rate of amino acid evolution between these two genes is a difference in the extent of functional constraint rather than a difference in rate of mutation, because a difference in mutation rate should affect synonymous as well as nonsynonymous sites.

In ER1, by contrast, \(d_S\) between the two \textit{HSP65} genes is only approximately half that between the two \textit{GroEL} genes; and the difference between these two \(d_S\) values is statistically significant (table 3). One possible explanation for the surprisingly low \(d_S\)

<table>
<thead>
<tr>
<th>Table 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d_N \pm \text{SE}, \text{per 100 Sites, in Comparisons between Duplicate Chaperonin Genes in T-Cell ER1–ER4 and Remainder of \textit{Mycobacterium tuberculosis}, \textit{M. leprae}, and \textit{Streptomyces albus}})</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>\textit{M. tuberculosis}</th>
<th>\textit{M. leprae}</th>
<th>\textit{S. albus}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER1</td>
<td>27.4 ± 4.0a</td>
<td>22.4 ± 4.5b</td>
</tr>
<tr>
<td>ER2</td>
<td>12.6 ± 4.1b</td>
<td>18.1 ± 5.0b</td>
</tr>
<tr>
<td>ER3</td>
<td>17.6 ± 4.1b</td>
<td>41.9 ± 7.2c</td>
</tr>
<tr>
<td>ER4</td>
<td>8.3 ± 2.7b</td>
<td>39.5 ± 7.2c</td>
</tr>
<tr>
<td>Remainder</td>
<td>37.5 ± 3.1</td>
<td>38.0 ± 3.1</td>
</tr>
</tbody>
</table>

*a Test of hypothesis that \(d_N\) equals that in remainder: \(P < 0.05\).
*b Test of hypothesis that \(d_N\) equals that in remainder: \(P < 0.001\).
*c Test of hypothesis that \(d_N\) equals that in \textit{M. tuberculosis}: \(P < 0.01\).
*d Test of hypothesis that \(d_N\) equals that in \textit{M. tuberculosis}: \(P < 0.05\).
*e Test of hypothesis that \(d_N\) equals that in \textit{M. tuberculosis}: \(P < 0.001\).
value between the ER1 regions of the two \textit{HSP65} genes is that this region was homogenized in the past by a recombination event between the ancestors of the two species. An alternative explanation for the low $d_{s}$ estimate between the two genes in this region might be that a shared G+C-content bias has made the Jukes-Cantor correction unreliable in this case. However, there is no evidence of an unusual pattern of G+C content in ER1 of the \textit{HSP65} genes, either in comparison with other regions of the same genes or in comparison with the two \textit{GroEL} genes (data not shown). Furthermore, when Bulmer et al.'s correction was used, $d_{s}$ in ER1 between the two \textit{HSP65} genes ($0.363 \pm 0.087$) remained approximately half as high as that between the two \textit{GroEL} genes ($0.799 \pm 0.186$).

To understand the rate of nonsynonymous evolution in the conserved epitopes identified by Munk et al. (1989), I estimated $d_{N}$ in these regions and in the remainder of the gene, between the duplicate chaperonin genes of \textit{M. tuberculosis}, \textit{M. leprae}, and \textit{Streptomyces albus} and \textit{Streptomyces albus} and two reference sequences: the human mitochondrial chaperonin and the \textit{GroEL} gene of \textit{Bacillus subtilis}. These reference sequences were used to perform relative-rate tests (Wu and Li 1985), in order to test the hypothesis that nonsynonymous nucleotide sites in the duplicate chaperonin genes have evolved at the same rate (table 5). In the case of \textit{Streptomyces albus}, the results showed that nonsynonymous sites both in the conserved epitopes and in the remainder of the gene have evolved at essentially the same rates in the two duplicate genes (table 5). In the case of \textit{M. tuberculosis} and \textit{M. leprae}, the \textit{GroEL} genes evolved more rapidly than did the \textit{HSP65} genes (table 5). What was particularly striking was the extent of this difference in rate in the conserved epitopes. In these epitopes, the \textit{GroEL} genes had diverged from the reference species approximately twice as rapidly as had the \textit{HSP65} genes (table 5). In other regions, by contrast, the \textit{GroEL} genes had diverged only \textasciitilde1.1-1.4 times as fast as in the conserved epitopes (table 5). Nonetheless, these conserved epitopes were still highly conserved in \textit{GroEL} of \textit{Mycobacterium}, with $d_{N}$

\begin{table}[h]
\centering
\begin{tabular}{llllll}
\hline
 & \multicolumn{2}{c}{CONSERVED EPITOPES} & \multicolumn{2}{c}{REMAINDER} \\
 & \textit{Human} & \textit{B. subtilis} & \textit{Human} & \textit{B. subtilis} \\
\hline
\textit{M. tuberculosis}: & & & & \\
\textit{GroEL} (D1) & 12.6 $\pm$ 3.5 & 8.9 $\pm$ 2.9 & 67.8 $\pm$ 3.8 & 48.3 $\pm$ 2.8 \\
\textit{HSP65} (D2) & 6.1 $\pm$ 2.5 & 3.0 $\pm$ 1.6 & 56.8 $\pm$ 3.2 & 33.7 $\pm$ 2.2 \\
D1-D2$^{a}$ & 6.5 $\pm$ 2.8$^{b}$ & 5.9 $\pm$ 2.6$^{b}$ & 11.0 $\pm$ 3.3$^{c}$ & 14.6 $\pm$ 2.6$^{c}$ \\
\textit{M. leprae}: & & & & \\
\textit{GroEL} (D1) & 15.0 $\pm$ 3.8 & 11.3 $\pm$ 3.3 & 64.3 $\pm$ 3.6 & 45.9 $\pm$ 2.7 \\
\textit{HSP65} (D2) & 7.8 $\pm$ 2.7 & 4.8 $\pm$ 2.1 & 57.3 $\pm$ 3.3 & 34.9 $\pm$ 2.2 \\
D1-D2$^{a}$ & 7.2 $\pm$ 3.1$^{b}$ & 6.5 $\pm$ 3.0$^{b}$ & 7.0 $\pm$ 3.2$^{c}$ & 11.0 $\pm$ 2.6$^{c}$ \\
\textit{S. albus}: & & & & \\
1 (D1) & 6.6 $\pm$ 2.4 & 2.1 $\pm$ 1.4 & 60.6 $\pm$ 3.4 & 36.2 $\pm$ 2.3 \\
2 (D2) & 7.4 $\pm$ 2.6 & 3.0 $\pm$ 1.6 & 56.6 $\pm$ 3.2 & 33.7 $\pm$ 2.3 \\
D1-D2$^{a}$ & -0.8 $\pm$ 0.9 & -0.9 $\pm$ 0.9 & 4.0 $\pm$ 2.7 & 2.5 $\pm$ 2.2 \\
\hline
\end{tabular}
\caption{\textit{d}_{N} \pm SE, per 100 Sites, in Conserved Epitopes and Remainder of Gene, between Chaperonins of \textit{Mycobacterium} and \textit{Streptomyces albus} and Two Reference Species (Human and \textit{Bacillus subtilis}), with Relative-Rate Test Comparing Duplicate Chaperonin Genes in Each Species}
\end{table}

\(^{a}\) Relative-rate test based on D1-D2, with SE of difference estimated by the method Wu and Li (1985).

\(^{b}\) Significance of relative-rate test: $P < 0.05$.

\(^{c}\) Significance of relative-rate test: $P < 0.001$. 

in this region between the mycobacterial GroEL and the reference sequence being less than one quarter that in the remainder of the gene (table 5).

Discussion

Evolutionary Rates

Analysis of both amino acid (fig. 2) and DNA (tables 3–5) sequences provides evidence of a substantial difference in the rate of nonsynonymous evolution between the two duplicate chaperonin genes of Mycobacterium tuberculosis and M. leprae. Since there are no consistent differences in the rate of evolution at synonymous sites (table 2), it seems likely that the difference is not due to a difference in rate of mutation but, rather, to a difference in the nature of the selective constraint acting on the proteins encoded by these genes. The reason for this difference in selective constraint is not known at present. Here I will review some evidence regarding possible causes for such a difference.

First, an acceleration of the rate of amino acid evolution is not unique to mycobacterial GroEL but was found in two other groups of chaperonin genes: the eukaryotic mitochondrial chaperonins and the plant chloroplast chaperonins (fig. 2). Both of these groups of chaperonins of eukaryotic organelles have in common that they are expressed in a biochemical environment quite different from that of bacterial chaperonins. It is probable, for example, that chaperonins in eukaryotic cells must interact with a set of proteins somewhat different from those encountered in prokaryotic cells. This in turn might give rise to differences in functional constraints on the protein. For example, bacterial and eukaryotic members of the HSP70 family have diverged markedly from one another in certain domains, particularly domains involved in interaction with other proteins (Hughes 1993a).

Similarly, a functional difference may be associated with the difference in evolutionary constraint acting on the duplicate mycobacterial chaperonins. One hypothesis that would account for such a difference is that, because of the gene duplication, the mycobacterial GroEL gene is redundant and that therefore its product is not subject to evolutionary constraint (Ohno 1970, p. 72). This would be consistent with the failure of Rinke de Wit et al. (1992) to detect GroEL from M. leprae in response to heat shock in an Escherichia coli system. However, these authors note that the M. leprae GroEL gene possesses regulatory sequences required for expression and thus does not appear to be a pseudogene. Their failure to detect GroEL may rather be due to the fact that the antibodies and T-cell probes that Rinke de Wit et al. used were actually mainly directed against HSP65, and these proteins may differ immunologically (Rinke de Wit et al. 1992).

Furthermore, it seems unlikely that the mycobacterial GroEL genes are entirely freed from constraint. In all regions analyzed, \( d_s \) between the GroEL genes of M. tuberculosis and M. leprae exceeds \( d_N \), whereas \( d_s \) and \( d_N \) are expected to be equal in a truly functionless gene. Also, \( d_N \) differs from one region to another of mycobacterial GroEL (table 5). Such differences would not be expected in the absence of functional constraint. Moreover, it is difficult to explain why one of the duplicate chaperonins should be freed from constraint in Mycobacterium but not in Streptomyces albus, especially since these duplicate genes seem to have arisen from the same duplication event (fig. 2). Yet, there is no evidence of an accelerated rate of amino acid evolution in either of the two chaperonin genes of S. albus (fig. 2 and table 5).

An alternative explanation for the apparent difference in functional constraint between the duplicate mycobacterial chaperonins is that it relates to differential in-
teraction between these two proteins and the vertebrate immune system. Mycobacterial HSP65, but not GroEL, is known to be a major antigen in the immune response of the mammalian host (Thole and van der Zee 1990). In the case of S. albus, although opportunistic infection of vertebrates occurs, this does not appear to be a major aspect of the species' niche (Mishra and Gordon 1986). By contrast, M. tuberculosis and M. leprae belong to a genus including numerous intracellular parasites of vertebrates (Wayne 1984). The fact that the duplicate chaperonins of these two species are subject to different constraints but those of the largely free-living S. albus are not is consistent with the hypothesis that the host immune system is responsible for these different constraints. Furthermore, the difference in evolutionary rates between mycobacterial GroEL and HSP65 is most striking in regions known to play a role in immune recognition, particularly the conserved, cross-reactive epitopes identified by Munk et al. (1989) (table 5).

The chaperonins and other stress proteins are expressed in heat shock and other environmental stresses, apparently functioning to prevent damage to cellular proteins under such conditions (Ellis and van der Vies 1991; Welch 1991). The difference in level of expression between normal and stress conditions can be great, at least in prokaryotes. For example, in E. coli, GroEL accounts for 1.6% of total protein under normal conditions but for 15% after heat shock (Herendeen et al. 1979). The hostile environment encountered on invasion of a vertebrate host may induce increased levels of expression of bacterial chaperonins (Young and Elliott 1989). Given the ubiquity and strong conservation at the amino acid level of bacterial chaperonins, it is not surprising that there is some degree of cross-reactivity in the immune response to chaperonins of different species. In fact, the immunogenic protein shared by most bacteria and designated "common antigen" has been found to be chaperonin (Shinnick et al. 1988).

In mammals, T cells bearing γδ T-cell receptors have been implicated in recognition of mycobacterial chaperonins, and it has been suggested that these T-cell receptors (which have a much lower potential diversity than do αβ T-cell receptors) are specialized for recognition of conserved epitopes such as those of mycobacterial chaperonins (Young and Elliott 1989; Kabelitz 1992). In the case of Mycobacterium, the γδ T-cell response may be a relatively ineffective defense, at least in certain individuals. For example, in leprosy, these T cells contribute to granuloma formation (Modlin et al. 1989).

Given the ability of the vertebrate host to recognize conserved antigens on mycobacterial HSP65, it may seem paradoxical that this protein has not evolved rapidly under selection favoring evasion of such recognition. One hypothesis that would explain the strong conservation of HSP65 is that the functional constraints on this molecule are too great to permit many changes at the amino acid level. Alternatively, it is possible that HSP65 has maintained a conserved amino acid sequence precisely because this sequence attracts a relatively ineffective γδ T-cell–mediated immune response. This hypothesis would explain why mycobacterial GroEL has evolved more rapidly than HSP65 and why this rapid divergence is particularly marked in conserved, cross-reactive epitopes (table 5). Under this hypothesis, GroEL would serve as a functional chaperonin but would have diverged sufficiently at the amino acid level to avoid recognition by T cells specific for conserved chaperonin epitopes. HSP65, on the other hand, would be a kind of molecular decoy whose role is to stimulate such an ineffective immune response and thereby protect the other chaperonin from attack by the immune system.
Recombination

This study has provided evidence of a past recombination event that homogenized the 5′ end of the HSP65 gene between M. tuberculosis and M. leprae. This event occurred at a time sufficiently distant in the past that it would probably be difficult to determine the exact boundary of the recombination region. However, most if not all of ER1 seems to have been affected (table 3). Because the event served to homogenize the two species in this region, it also cannot be determined which species was the donor.

The extent to which, in nature, recombination occurs among different genotypes or closely related species of bacteria is controversial (Maynard Smith et al. 1991). Multilocus enzyme electrophoresis has revealed a clonal population structure, which implies little recombination, in natural populations of several bacterial species (Whittem et al. 1983; Caugant et al. 1985; Musser et al. 1985). On the other hand, certain DNA sequences from bacterial genes show a “mosaic” structure, indicating recombination between different strains or species (Maynard Smith et al. 1991). One possible resolution of the apparent contradiction between these findings is the hypothesis that rates of recombination are generally low in bacteria but that recombinants are more likely to be seen in cases where a recombinant gene has some selective advantage (Nelson and Selander 1992). This hypothesis is supported by the strong evidence of past recombination in the case of penicillin-binding proteins (Maynard Smith et al. 1991) and flagellin (Smith et al. 1990).

It is not known whether recombination at the HSP65 locus was selectively favored, as might be expected under this hypothesis. However, the region involved is highly conserved and includes known T-cell epitopes (tables 3 and 4). If HSP65 in M. tuberculosis and M. leprae functions to elicit a cross-reactive T-cell response on the part of the host, homogenization in this region may have been advantageous to these two species of Mycobacterium.

Acknowledgment

This research was supported by a grant from the National Institutes of Health.

LITERATURE CITED


MORRISON, P. R., R. L. BELLAND, K. LYNG, and H. D. CALDWELL. 1989. Chlamydial disease


WALTER M. FITCH, reviewing editor

Received January 22, 1993; revision received July 27, 1993

Accepted July 27, 1993