Positive Selection and Subfunctionalization of Duplicated CCT Chaperonin Subunits

Mario A. Fares and Kenneth H. Wolfe

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin, Ireland

To reach a functional and energetically stable conformation, many proteins need molecular helpers called chaperonins. Among the group II chaperonins, CCT proteins provide crucial machinery for the stabilization and proper folding of several proteins in the cytosol of eukaryotic cells through interactions that are subunit-specific and geometry-dependent. CCT proteins are made up of eight different subunits, all with similar sequences, positioned in a precise arrangement. Each subunit has been proposed to have a specialized function during the binding and folding of the CCT protein substrate. Here, we demonstrate that functional divergence occurred after several CCT duplication events due to the fixation of amino acid substitutions by positive selection. Sites critical for ATP binding and substrate binding were found to have undergone positive selection and functional divergence predominantly in subunits that bind tubulin but not actin. Furthermore, we show clear functional divergence between CCT subunits that bind the C-terminal domains of actin and tubulin and those that bind the N-terminal domains. Phylogenetic analyses could not resolve the deep relationships between most subunits, except for the groups α/β/η and δ/ε, suggesting several almost simultaneous ancient duplication events. Together, the results support the idea that, in contrast to homo-oligomeric chaperonins such as GroEL, the high divergence level between CCT subunits is the result of positive selection after each duplication event to provide a specialized role for each CCT subunit in the different steps of protein folding.

Introduction

Protein folding does not usually occur spontaneously, but requires the assistance of chaperonins to overcome kinetic limitations that prevent them reaching the native conformation (Bukau and Horwich 1998). Two main chaperonin mechanisms for assisting protein folding have been characterized so far. Group I chaperonins include the GroEL protein, which is present in bacteria and organelles (Bukau and Horwich 1998; Ellis and Hartl 1999), and group II includes the CCT (chaperonin containing tailless complex polypeptide 1 [TCP-1]) proteins found in archaea and the cytosol of eukaryotes (Gutsche, Essen, and Baumeister 1999; Willison 1999). Both groups of chaperonins are ATPases of 60 kDa that form multimeric complexes organized in two back-to-back oriented rings. Furthermore, both groups of chaperonins have the ability to bind and fold nonnative proteins in an appropriate conformation, preventing any nonspecific aggregation. Like GroEL, CCT proteins comprise three main domains: apical, intermediate, and equatorial (Ditzel et al. 1998; Nitsch et al. 1998; Llorca et al. 1998, 1999*a*, 1999*b*, 2000; Gutsche et al. 2000; Schoehn et al. 2000*a*, 2000*b*). Structural differences between these two groups have however been reported, the lack of a GroES-like cochaperonin in eukaryotic cells being one of the most remarkable dissimilarities. This cofactor seems to be replaced by a helical protrusion located at the tip of the apical domain, which acts as a cap upon the binding of the unfolded protein in the chaperonin cavity (Klumpp, Baumeister, and Essen 1997; Ditzel et al. 1998; Nitsch et al. 1998; Llorca et al. 1999*a*). Another structural difference is that, in contrast to the seven identical subunits that build each ring of GroEL, eukaryotic CCT protein consists of eight subunits, with different degrees of sequence and functional divergence, precisely arranged in each ring (Liou and Willison 1997; Willison 1999; Llorca et al. 1999*a*; Grantham et al. 2000).

Unlike for GroEL, the number of substrates that bind to CCT is small (Kim, Willison, and Horwich 1994), the best characterized interactions being those of tubulins, actins, and a very small number of other proteins (Sternlicht et al. 1993; Kubota, Hynes, and Willison 1995; Hynes et al. 1996; Lewis et al. 1996; Thulasiraman, Yang, and Frydman 1999). Nonetheless, the list of known CCT-substrates continues to growth with the addition of proteins such as luciferase (Frydman et al. 1994), G-α-transducin (Farr et al. 1997), hepatitis B virus capsid protein (Lingappa et al. 1994), cyclin E (Won et al. 1998), the EBNA1 viral protein (Kashuba et al. 1999), myosin (Srikakulam and Winkelmann 1999), and the tumor-suppressor protein VHL (Feldman et al. 1999). Klumpp, Baumeister, and Essen (1997) suggested that the high flexibility of the CCT apical protrusion would accommodate a wide variety of different substrates.

The specificity of the interaction between CCT subunits and α-actins was previously demonstrated by electron microscopy (Llorca et al. 1999*a*). The description of the interaction between actin and CCT by Llorca et al. (1999*a*, 2000) and Hynes and Willison (2000) has been refined by McCormack, Rohman, and Willison (2001), who defined the main actin domains that bind to specific CCT subunits. Furthermore, the similarity of actin and tubulin domains that bind to CCT has been demonstrated in several biochemical studies (Hynes and Willison 2000; Llorca et al. 2000; Ritco-Vonsovici and Willison 2000). It has been demonstrated that the small N-terminal domain of actin binds to the CCTδ subunit, and the large C-terminal domain binds mainly CCTε but also CCTβ (Llorca et al. 1999*b*; Hynes and Willison 2000). Tubulin establishes two possible binding arrangements, but uses CCTβ and CCTε with the highest affinity. In contrast to GroEL, in which substrate-binding sites seem to accommodate many different amino acid sequences by means of hydrophobic

Key words: CCT, duplication, positive selection, functional divergence, convergent evolution, protein folding.

E-mail: faresm@tcd.ie.

Mol. Biol. Evol. 20(10):1588–1597. 2003
DOI: 10.1093/molbev/msg160
Molecular Biology and Evolution, Vol. 20, No. 10,

Table 1
CCT Subunits That Bind Either the N-Terminal Domains or the C-Terminal Domains of Actin and Tubulin

Subunit[a]	Actin[b]	Tubulin[b]
α	—	N
β	C	C
δ	N	N
ε	C	C
γ	—	C
η	—	N
θ	—	NC
ζ	—	C

[a] CCT subunits.

[b] Well-characterized protein substrates of CCT subunits. N and C indicate that the subunit binds the N-terminal or C-terminal domains, respectively, of actin or tubulin. NC indicates that subunit θ can bind both the N-terminal and C-terminal domains of tubulin. Binding data is from Llorca et al. (2001).

interactions (Fenton et al. 1994), there is biochemical evidence that specific subunits of CCT establish non-hydrophobic interactions with their protein substrates (Hynes and Willison 2000). Most interesting is that most of the amino acid differences among CCT subunits reside in the apical domain, which contains the substrate-binding sites (Kim, Willison, and Horwich 1994; Pappenberger et al. 2002), and that the different subunits are positioned in a specific arrangement (Liou and Willison 1997; Llorca et al. 1999*a*; Grantham et al. 2000). Although the CCT subunits are very divergent from one another, a detailed phylogenetic analysis revealed that each CCT subunit group of sequences can be distinguished by a conserved set of amino acid residue ''signatures'' located in the three protein domains (Archibald, Blouin, and Doolittle 2001). The existence of these amino acid signatures supports the hypothesis of functional divergence between the different CCT subunits, which was previously suggested by Kubota et al. (1994). Interestingly, some of these slowly evolving amino acid signatures are located in ATP-binding and hydrolysis motifs mapped onto the equatorial domain (Archibald, Logsdon, and Doolittle 2000; Archibald, Blouin, and Doolittle 2001).

Here we demonstrate that functional divergence occurred between the different CCT subunits after each gene duplication event during the early evolution of eukaryotes. This divergence is apparent at the protein level due to the fixation of amino acid replacements in sites involved in ATP binding and in protein binding. In this study, we also demonstrate that CCT subunits that bind actin and those that bind tubulin were subjected to different selective constraints. These evolutionary results integrate with biochemical and structural results to support the model proposed for the interaction of CCT chaperonins with actin and tubulin (Llorca et al. 2001).

Material and Methods

Sequence Data

The different CCT subunits used in this study and their interaction patterns with actin and tubulin proteins are shown in table 1. Accession numbers for the 50 full-length CCT subunit sequences used in the phylogenetic analyses are listed in table 2. In this study, we used CCT sequences with well-supported established phylogenetic relationships (Archibald, Blouin, and Doolittle 2001), since the phylogenetic tree was used as an initial tree in subsequent analyses. For maximum-likelihood estimation of the parameters of the models used to detect selection that require intensive computation, only a subset of 42 sequences was used. The sequences removed from the alignment were those that do not produce any change in the parameter estimates under maximum-likelihood models used in the phylogenetic reconstruction. The root of the phylogenetic tree was established using archaeal outgroups: *Sulfolobus solfataricus* α (NC_002754.1), *Sulfolobus solfataricus* β (AAK40620), *Aeropyrum pernix* α and β (NC_000854.1), *Methanococcus thermolithotrophicus* (O93624), *Haloferax volcanii* (AF298660), and *Thermoplasma acidophilum* (NC_002578.1).

Sequence Alignment and Phylogenetic Tree Reconstruction

Amino acid sequence alignments were performed with the ClustalX program version 1.81, available from http://inn-prot.weizmann.ac.il/software/ClustalX.html (Thompson, Higgins, and Gibson 1994), followed by a careful hand-correction of the alignments using the program GENEDOC (Nicholas and Nicholas 1997). To obtain accurate nucleotide alignments, we first aligned the amino acid sequences, and then nucleotide alignment was obtained by concatenating the triplets according to the amino acid sequence alignment. Highly divergent or ambiguous regions as well as gap regions were removed

Table 2
Genbank Accession Numbers for the Different CCT Subunits in Each Eukaryotic Species Used

Subunit	H.s.	M.m.	C.e.	S.c.	A.t.	G.i.	T.v.
α	X52882	P11984	NM_063321	M21160	NM_112896	AF226720	AF226714
β	NP_006422	P80314	U25632	X77675	NM_122097	AF226721	—
δ	NM_006430	NM_009837	NM_063349	Z33504	AB020749	—	AF226717
ε	D43950	NM_007637	U25698	L37350	O04450	AF226724	—
γ	NM_005998	NM_009836	NM_061817	U09480	NM_122537	—	—
η	NM_006429	NM_007638	NM_071121	NC_001142	NM_112016	—	AF226715
θ	NM_006585	NM_009840	NM_067634	P47079	AC011698	—	—
ζ_1	NM_001762	NM_009838	NM_065861	NC_001136	T51390	AF226726	AF226719
ζ_2	D78333	NM_009839					

NOTE.—Species abbreviations are *Homo sapiens* (H.s.), *Mus musculus* (M.m.), *Caenorhabditis elegans* (C.e.), *Saccharomyces cerevisiae* (S.c.), *Arabidopsis thaliana* (A.t.), *Giardia intestinalis* (G.i.), and *Trichomonas vaginalis* (T.v.). Underlined sequences are those that were not used in the maximum-likelihood parameter estimation.

from the alignment in all the analyses performed. Phylogenetic analysis of the aligned sequences was performed using the Neighbor-Joining (NJ) (Saitou and Nei 1987), maximum-likelihood (ML) methods (Felsenstein 1981), and maximum-parsimony (MP) methods (Fitch 1971). The MEGA program version 2.01 (Kumar, Tamura, and Nei 1993) was used to obtain NJ trees. MP and ML trees were reconstructed using DNAPARS and DNAML, respectively, from the PHYLIP package version 3.5 for Windows (Felsenstein 1993). Furthermore, a gamma distance–based phylogenetic tree (Feng and Doolittle 1997) was obtained for the amino acid sequences. The gamma shape parameter (α) was estimated using the program GAMMA (Gu and Zhang 1997). The support for each phylogenetic group was tested using 1,000 bootstrap pseudoreplicates for both nucleotide-based and amino acid–based phylogenetic trees.

Testing the Constancy of Substitution Rates in the Branches Leading to Each CCT Subunit Cluster

To know whether significant amino acid changes occurred in CCT genes after duplication events, the constancy of substitution rates among lineages was tested using the two-cluster test implemented in the LINTREE program (Takezaki, Rzhetski, and Nei 1995). The two-cluster test examines the equality of the average substitution rates for two clusters linked by a node on the phylogenetic tree, using one or several outgroup sequences. The root of the phylogenetic tree for the two-cluster test was established using a subunit sequences from *Methanococcus thermolithotrophicus*, *Thermoplasma acidophilum*, and *Sulfolobus sulfotaricus*.

Detection of Positive Selection in the Different Branches of the Eukaryotic CCT Phylogenetic Tree

To test the hypothesis of variable selective pressures among the different branches of the CCT phylogenetic tree, the free-ratio model was compared with the Goldman and Yang model by the likelihood-ratio test (LRT) (Huelsenbeck and Crandall 1997). This test is based on the fact that the log-likelihood values of two nested models can be compared since twice the log-likelihood difference among the nested models follows a χ^2 distribution, with the degrees of freedom being the difference between the numbers of free parameters between the models compared. These tests were implemented using the program CODEML from the PAML package version 3.0 (Yang 2000).

The Goldman and Yang (1994) model assumes a single ω value for the whole phylogenetic tree along the complete sequence alignment, whereas the free-ratio model (Yang 1998) estimates a log-likelihood value assuming a different ω value for each branch of the phylogenetic tree. These two models are nested, and, hence, their likelihood values can be compared using the LRT, the number of degrees of freedom being the difference in the number of ω values freely estimated between the two models compared. Therefore, for the comparison between the model of Goldman and Yang and the free-ratio model. the number of degrees of freedom is N–1, where N is the number of branches in the tree.

Examining Functional Divergence After CCT Gene Duplication

CCT sequence duplication events were tested for type I functional divergence (Wang and Gu 2001) using the method developed by Gu (1999). To implement this procedure, we used the program Diverge, version 1.04 (Gu and Vander Velden 2002; http://xgu1.zool.iastate.edu/cgi-bin/download.cgi). This method uses a maximum-likelihood procedure to estimate whether there has been a significant change in the rate of evolution after the emergence of two paralogs. This is done by calculating a coefficient of functional divergence (θ) and determining whether the value of the coefficient is large enough to reject the null hypothesis of no functional differentiation. If the null hypothesis is rejected, the program calculates a posterior probability for functional divergence for each position in the alignment. We established a cutoff value, according to the effect that the elimination of the sets of amino acids having a posterior probability higher than this cutoff value had on the θ value test (see Wang and Gu 2001). To detect the amino acids responsible for significant functional divergence after each gene duplication event, we compared subunits or well-supported groups of subunits to each other and determined the cutoff value above which the difference between the subunits becomes significant.

Results

Phylogenetic Relationship Between Paralogous Sequences and Multiple Timings for the Different CCT Duplication Events

Gamma distance–based phylogenetic inference was carried out on amino acid sequences (fig. 1) (the sequence alignment is available as Supplementary Material online), and the support for each group of sequences was tested by 1,000 bootstrap pseudoreplicates. In this phylogenetic tree, the relationships among orthologous sequences were identical to those obtained using nucleotide sequence alignment (data not shown). The phylogenetic tree is in agreement with that obtained by Archibald, Blouin, and Doolittle (2001) but shows lower bootstrap values for the well-established groups of CCT subunits. Interestingly, the bootstrap values for the deep branches of the tree are very low, whereas high bootstrap values give support to a common origin for the groups of subunits η-β (80%), η-β-α (83%), and δ-ε (98%). Little can be concluded about the temporal order of the earlier duplication events, because the use of archaeal sequences as outgroups did not resolve the deepest branches of the phylogenetic tree (fig. 1). A possible explanation for the pattern of CCT phylogenetic relationships is that four nearly simultaneous duplications occurred at the base of the eukaryotic phylogenetic tree giving rise to five groups: subunits θ, γ, ζ, the ancestor of δ-ε, and the ancestor of α-β-η. Three independent duplications subsequently produced the remaining subunits. Alternatively, all duplication events may have occurred at the same time at the base of the

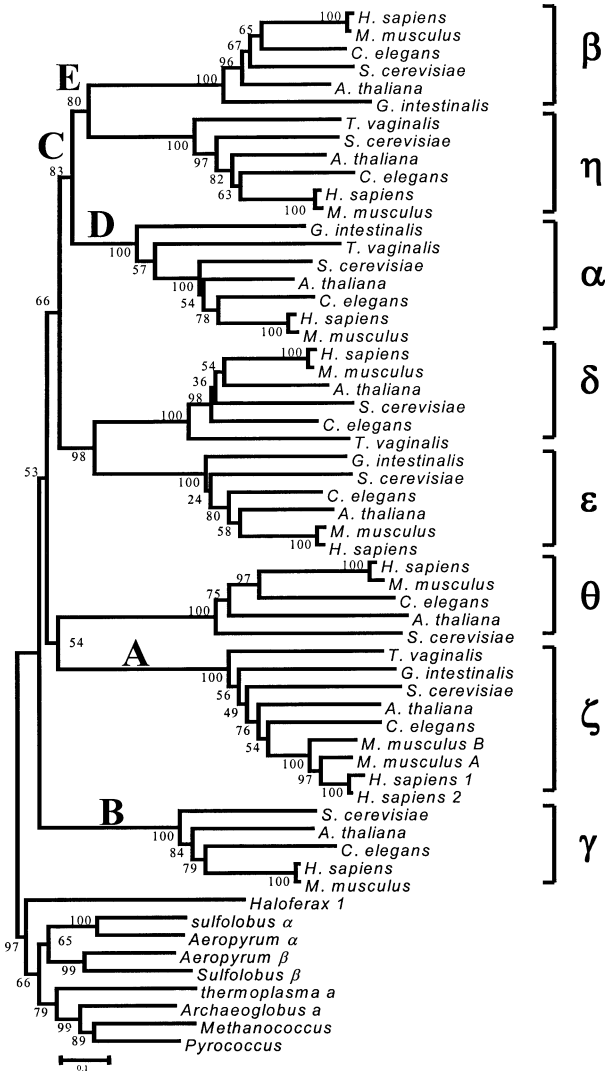

FIG. 1.—Phylogenetic tree for the CCT subunits α to ζ using gamma-corrected amino acid distances. Numbers in the nodes are the bootstrap values for 1,000 pseudoreplicates. Bold letters label branches of the tree where positive selection was detected. Branches *A–E* are discussed in the text.

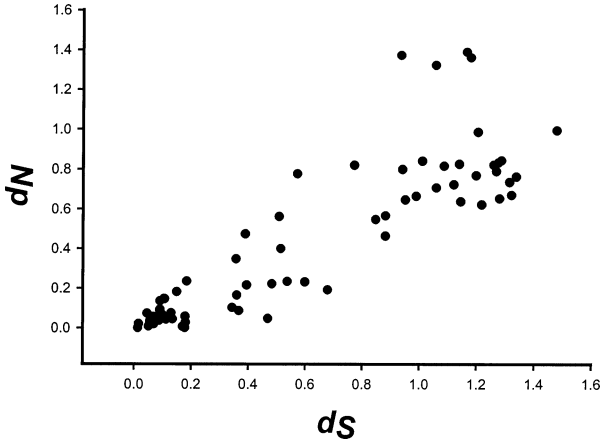

FIG. 2.—Plot of the nonsynonymous (d_N) against synonymous (d_S) nucleotide distances estimated by the modified method of Nei and Gojobori. Each point represents a branch of the phylogenetic tree, with the sequences at interior nodes having been inferred by maximum likelihood.

eukaryotic tree, followed by several independent co-evolutionary processes between subunits α, β and η, and subunits δ and ε.

Examining Saturation of Synonymous Sites in CCT Subunits

Saturation of synonymous sites is an important problem in the analysis of selective constraints when nucleotide sequences are compared. Several methods that estimate synonymous (d_S) and nonsynonymous (d_N) nucleotide distances can correct for multiple hits up to a threshold saturation level of nucleotide divergence. To ensure that CCT nucleotide alignments are still useful to estimate selective constraints, we made an initial analysis of the correlation between d_S and d_N. To do so, we used the protein-based phylogenetic tree to infer nucleotide sequences at the interior nodes of the phylogenetic tree by maximum-likelihood using the program CODEML from the PAML package version 3.0 (Yang 2000). d_S and d_N were then estimated by the modified method of Nei and Gojobori (Zhang, Rosenberg, and Nei 1998) for each branch of the phylogenetic tree. We plotted d_N versus d_S (fig. 2) and tested the correlation between these two distances. If synonymous sites are saturated, we might expect a logarithmic or quadratic model to be a better fit to the data than a linear model, whereas the opposite is true if synonymous sites are not saturated. The regression coefficient for the linear model was $r = 0.88$, whereas for the quadratic and logarithmic models were $r = 0.88$ and $r = 0.81$, respectively. The application of the quadratic and the logarithmic regression models did not significantly improve the sum of squares explained by the linear regression model ($t_1 = 1.38$, $P = 0.199$ and $t_1 = 32.22$, $P = 0.009$ for the comparison with the quadratic and logarithmic models, respectively), indicating that the linear model is good enough to explain the correlation between the two types of substitutions and that no saturation exists in either synonymous or nonsynonymous codon sites. We therefore conclude that synonymous sites are not saturated in this analysis and consequently that d_S and d_N can be compared to infer selective constraints in specific branches of the phylogenetic tree.

Accelerated Substitution Rates After CCT Duplication Events

The phylogenetic relationships between the groups of paralogous sequences from subunits α-β-η, δ-ε, γ, ζ, and θ cannot be resolved due to the low bootstrap values for the deepest branches of the tree. Therefore, we considered an initial polytomous phylogenetic tree and tested the constancy of amino acid substitution rates between well-supported groups of subunit sequences, as well as between pairs of CCT subunits within the α-β-η and δ-ε clusters. A note of caution is required about the interpretation of the results obtained using this test because failure to observe

Table 3
Two-Cluster Test for the Constancy of Amino Acid Substitution Rates Among the Different Well-Supported Groups of CCT Subunits

Group A	Group B	b_A[a]	b_B[a]	d[b]	Z	*P*
ζ	γ	0.628	0.459	0.168	2.464	**0.014**
ζ	θ	0.831	0.792	0.039	0.254	0.803
ζ	α-β-η	0.751	0.593	0.159	3.129	**0.002**
ζ	δ-ε	0.764	0.589	0.175	1.970	**0.024**
α-β-η	δ-ε	0.616	0.567	0.049	0.806	0.424
α-β-η	γ	0.628	0.614	0.014	0.192	0.849
α-β-ε	θ	0.639	0.813	0.173	1.323	0.187
β-η	α	0.626	0.469	0.157	3.282	**0.001**
β	η	0.552	0.557	0.005	0.118	0.924
θ	γ	0.802	0.624	0.178	1.257	0.211
θ	δ-ε	0.802	0.578	0.224	1.686	0.093
γ	δ-ε	0.586	0.604	0.017	0.215	0.834
δ	ε	0.396	0.475	0.078	1.401	0.162

[a] Distances from the root of the tree to groups A and B are represented by b_A and b_B.

[b] d estimates the difference between the distances of cluster A (b_A) and cluster B (b_B) to the root of the tree, and this difference is tested against the molecular clock assumption using a normal (Gaussian distribution) test. Significant probability values are bold.

significantly accelerated rates of amino acid substitution does not necessarily imply no variation in selective constraints, but rather just means that there was no difference in the evolutionary rate of the two groups of sequences that were compared. Two examples of accelerated rates for the fixation of amino acid replacement were seen in lineages leading to the different CCT subunits (table 3). The first one occurred in the branch leading to the β-η group CCT subunits ($P = 0.001$). Significantly accelerated substitution rates are also obtained when subunit ζ was compared with every subunit group except for subunit θ (table 3). Thus, at least two main changes in selective constraints have occurred during the evolution of the different CCT subunits.

Positive Selection Governing the Evolution of CCT Subunits

To highlight changes in selective constraints not revealed by the two-cluster test, we examined the nonsynonymous-to-synonymous rate ratio ($\omega = d_N/d_S$). This ratio is a good indicator of the intensity of selective pressure (Sharp 1997; Akashi 1999; Crandall et al. 1999), values of $\omega = 1$, $\omega < 1$, and $\omega > 1$ indicating neutrality, purifying selection, and positive selection, respectively. We compared the goodness-of-fit of the CCT data to the model of Goldman and Yang (1994), which assumes a single ω value for the complete tree and alignment, and the free-ratio model of Yang (1998), which allows free variation of ω values between the branches of the tree. The comparison between the models shows that the free-ratio model significantly improves the log-likelihood value obtained under the Goldman and Yang model ($2D\ell = 452.902$, 1-tailed $P = 0$).

The estimation of ω values for six branches of the phylogenetic tree revealed multiple episodes of positive selection distributed among the different branches leading to several subunits (table 4 and fig. 1). Values of ω significantly greater than 1 were obtained for the branch leading to ζ, γ, α, the β-η group, and the α-η-β group (table 4).

Notably, differences were observed between the selective constraints on CCT subunits that bind actin and those that bind tubulin. In fact, three subunits (α, γ, and ζ) that bind tubulin but not actin (Llorca et al. 2001) are under positive selection, whereas two subunits (δ and ε) that bind actin as well as tubulin are under strong purifying selection.

Table 4
Amino Acid Changes in Those Positions Involved in CCT Substrate Binding or ATP Binding

Branch[a]	Subunits	ω[b]	ATP Binding[c]	Protein Binding[d]
A	ζ	3.553	32–35,48–50,51,54	193,213–214,241,298,320,221,222,358,415
B	γ	1.2	35,48–49,53	242,293,297,318,320,321,323,359
C	α-β-η	5.42	32	219,282,284,286,289
D	α	10.73	30	219,238,240,242,289,313,317
E	β-η	2.211	35	313,317

[a] Branches are labeled by letters as shown in figure 1.

[b] Nonsynonymous-to-synonymous rate ratio.

[c] Amino acid positions of CCT involved in ATP binding where amino acid substitutions were fixed by positive selection. When ranges are listed, every residue in the range showed positive selection. Coordinates for amino acid positions follow Pappenberg et al. (2002).

[d] Amino acid positions of CCT involved in protein binding where positive selection of amino acid substitutions was detected.

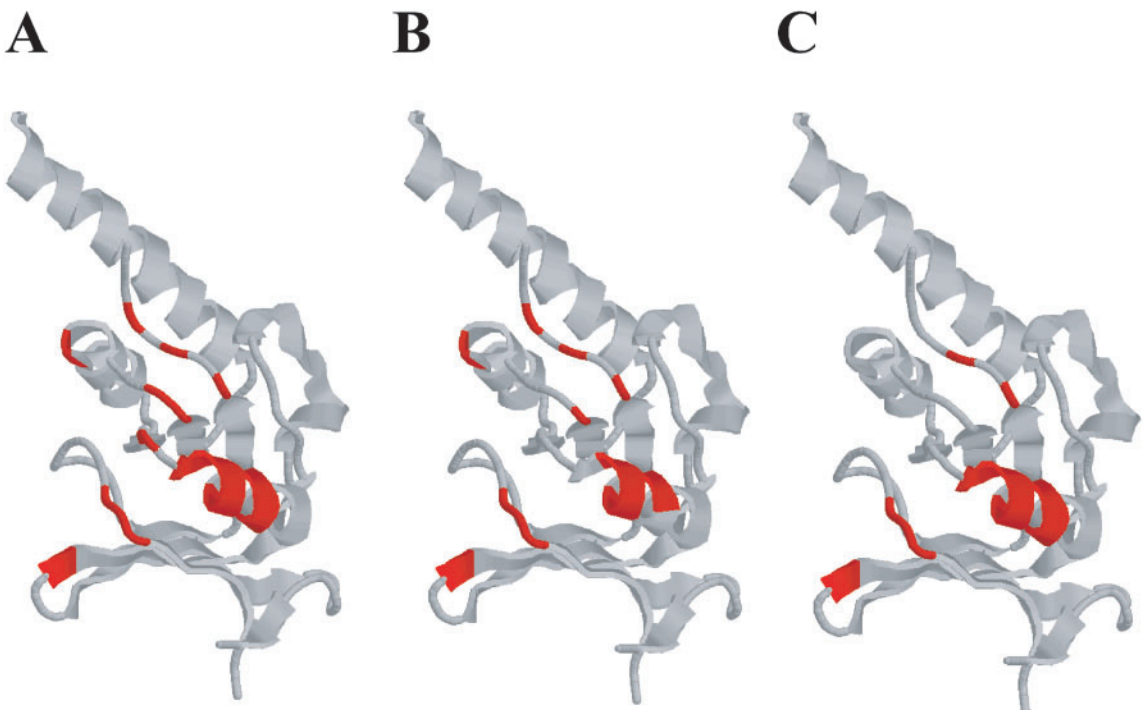

FIG. 3.—Three-dimensional structure of the mouse CCTγ apical domain. Regions where positively selected amino acid sites were detected in the different CCT subunits were mapped onto the mouse CCTγ subunit. (*A*) The three-dimensional structure showing substrate-binding sites in CCTγ (conserved motifs) following the coordinates of Pappenberger et al. (2002). (*B*) Amino acid regions where positive selection was detected. (*C*) amino acid regions identified as causing functional divergence between CCT subunits.

Posterior probabilities were used to identify the most likely codon changes in the branches detected under positive selection. Figure 3*A* shows the location in the three-dimensional structure of the mouse CCTγ subunit of amino acid sites that are important for ATP binding and hydrolysis as well as for protein substrate binding. Interestingly, among the amino acid substitutions detected with high posterior probabilities in those branches subject to positive selection, several are positioned in ATP-binding sites and substrate-binding sites (fig. 3*B*). Some of the substitutions are in sites previously proposed as ATP-binding sites (amino acid positions 31 to 35, 45 to 53, 89 to 91, and 114 to 117 [Kim, Willison, and Horwich 1994]), and others are located in conserved motifs proposed to be involved in protein binding (table 4).

Most interesting is that amino acid sites detected under positive selection and that are involved in ATP-binding sites or substrate-binding sites are hydrophilic residues (an average of 35% among the positively selected amino acid sites). This includes the amino acids K48, K50, D221, H222, D282, D284, Q286, E193, and E317, whereas the remaining amino acids (65%) did not imply any variation in amino acid hydrophobicity. These results indicate that positive selection has affected regions involved in substrate protein binding located in the apical domain, perhaps to optimize the binding and folding cycles, or, alternatively, that these changes are due to selection for functional divergence of each CCT subunit towards a specialized substrate-domain binding.

To test the significance of the distribution of amino acid substitutions, we conducted a χ^2 test comparing the observed number of substitutions in ATP-binding sites and substrate-binding sites with the expected number for each branch where positive selection was detected. The Bonferroni corrected probabilities of the observed amino acid substitutions in these sites gave values of 7.11×10^{-27}, 1.24×10^{-15}, and 1.52×10^{-23}, for the branches leading to CCT subunits ζ, γ, and α-η-β, respectively. Here, we only considered those amino acid substitutions with posterior probabilities higher than 0.90. This result indicates that subunits ζ, γ, and α-η-β, have acquired new functions or specialized its role by the fixation of amino acid substitutions in regions involved in substrate binding as well as ATP binding. On the other hand, we might expect that if saturation of synonymous sites had occurred, it would be randomly distributed in the gene sequence. This could result in the artifactual discovery of sites that appear to be under positive selection, but these sites should be at random locations. The amino acid positions where positive selection was detected were, however, significantly concentrated in sites involved in ATP and substrate binding.

Detection of Critical Amino Acids for Altered Functional Constraints on CCT Subunits

Functional divergence between genes results from changes in the functional constraints on one of the genes, resulting in high sequence conservation of one paralog in different species, while the other paralog evolves more freely. We analyzed functional divergence using the method of Gu (1999). The maximum-likelihood estimation of the coefficient of functional divergence (θ), its standard error and its significance for CCT subunit pairwise comparisons are shown in table 5. Significant θ values were

Table 5
Test of Functional Divergence Between Subunits or Group of Subunits Defined by the Phylogenetic Tree

Comparison	θ[a]	SE (θ)	LRT[b] (θ)	*P*[c]	Probability Cutoff[d]
α-β-η versus δ-ε	0.257	0.039	42.255	< 0.0001	0.50
α-β-η versus γ	0.592	0.064	83.203	< 0.0001	0.75
α-β-η versus θ	0.257	0.065	29.308	< 0.0001	0.50
α-β-η versus ζ	0.293	0.052	31.324	< 0.0001	0.60
β-η versus α	0.235	0.046	25.787	< 0.001	0.50
β versus η	0.382	0.066	33.009	< 0.0001	0.51
δ-ε versus γ	0.496	0.072	46.472	< 0.0001	0.60
δ-ε versus θ	0.464	0.086	30.084	< 0.0001	0.55
δ-ε versus ζ	0.399	0.066	36.348	< 0.0001	0.50
δ versus ε	0.378	0.094	16.357	< 0.01	0.50
γ versus ζ	0.465	0.061	57.581	< 0.0001	0.80
γ versus θ	0.646	0.080	64.282	< 0.0001	0.80
ζ versus θ	0.334	0.080	17.581	< 0.0001	0.50

[a] θ is the parameter of functional divergence (Gu 1999).
[b] LRT (θ) is a likelihood ratio test comparing the likelihood values with and without the assumption of functional divergence.
[c] Probability of the LRT value corrected by Bonferroni.
[d] Probability cutoff is the minimum posterior probability for amino acids causing functional divergence in each case.

obtained for the pairwise subunit comparisons as well as for comparisons among different well-supported groups of subunits (table 5), implying that functional divergence between the different subunits is significantly supported by the data. Notably, functional divergence was detected after the duplications that gave CCT subunits β and ε that bind the C-terminal domains of actin and tubulin with the highest affinity compared with the remaining CCT subunits. Critical amino acids for functional divergence were detected by estimating the posterior probability of belonging to an S_1 class (see Wang and Gu 2001) for each amino acid in the alignment in each comparison. To detect critical amino acids for type I functional divergence, different posterior-probability cutoff values were used, depending on the subunits or group of subunits compared (table 5). When critical amino acids (amino acids with probabilities higher than the cutoff value) were removed from the alignment, the θ value was not significantly different from 0, ranging between 0 and 3.730 ($P = 1$ and $P = 0.053$, respectively).

The amino acid positions causing functional divergence were found to coincide in many cases among the different comparisons performed and in many cases with those identified as contributing to high ω values in the previous sections. Notably, some of the amino acids critical for functional divergence are also critical for substrate binding (amino acid positions 221 to 223, 242, 244, 316 to 319, 321, 322, and 358) as well as close to other regions involved in ATP hydrolysis (sites 31, 33 to 35, 91, 415, and 416, following the coordinates of Kim, Willison, and Horwich [1994] by comparison with GroEL for the ATP-binding amino acid positions [fig. 3*C*]).

Because the scheme of binding C-terminal domains by CCT subunits is more complex than that regarding the binding of the N-terminal domains, we expect functional divergence also between CCT subunits binding different protein domains. To test this hypothesis, we analyzed the functional divergence between the two types of subunits (for example, we compared CCT subunit β with subunit η). Among the critical amino acids causing functional divergence between actin/tubulin-C-terminal–binding CCT subunits and actin/tubulin-N-terminal–binding CCT subunits, many (amino acid positions 31, 33 to 35, 48, 415, and 516) are identified as involved in ATP binding or in substrate binding (amino acid positions 222, 223, 244, 298, 316 to 318, and 358) (fig. 3*C*), also detected by the likelihood-based free-ratio model (fig. 3*B*).

The pairwise comparison of CCT subunits as well as the comparison between subunits able to bind either the N-terminal domains or the C-terminal domains of actin and tubulin show that functional divergence has occurred at amino acid sites important for both substrate binding and ATP binding.

Discussion

The high frequency of gene duplication (Lynch and Conery 2000; Li et al. 2001) indicates that this phenomenon is a good candidate to explain the emergence of complexity in eukaryotic organisms. The relationship between gene duplication and the emergence of complexity is clearly exemplified by the CCT proteins. Detailed phylogenetic analyses suggested that the CCT family probably originated by duplication and divergence at the base of the eukaryotic tree (Kubota et al. 1994). Our phylogenetic analysis gives similar results to those of Archibald et al. (2001). The methods used failed to unambiguously resolve the deep branching patterns, although some of the subunits appear to share more recent ancestry than others. Our analyses suggest that four duplication events occurred in quick succession with a slight divergence between the duplicated gene copies to give the ancestors of subunits θ, γ, ζ, δ-ε, and α-β-η. Later, three independent duplication events occurred, giving rise to the remaining subunits.

The obvious question that arises here is whether these duplication events meant the acquisition of new functions (neofunctionalization) by each CCT subunit (for example to provide a broader specificity of substrate recognition) or whether a coevolutionary process occurred between several CCT subunits to accommodate a unique function

in substrate binding while each of the others subunits maintained its old function (subfunctionalization). In the former case, we might expect to see functional divergence between the different subunits and no correlation in the amino acid sites at which positive selection is detected. In the latter case, however, although functional divergence might also be detected between CCT subunits, a correlation might be expected in the selective constraints on subunits involved in the same function, as was seen here. It is straightforward to hypothesize that, given the close physical contact between the different CCT subunits in each ring, all the subunits may have to coevolve to optimize substrate binding.

The significant acceleration in the rates of amino acid replacements detected by the two-cluster test in the lineages leading to the α-β-η group and in the lineage leading to subunit ζ suggests functional divergence after the duplications that gave rise to these subunits. The detection of positive selection in branches leading to different CCT subunits after many gene duplication events also suggests functional divergence.

Our detection of positive selection in different lineages agrees with CCT subunit-substrate interaction model suggested by Llorca et al. (2001). In fact, we have shown that positive selection has occurred in the branches leading to CCT subunits involved exclusively in tubulin binding (α, ζ, and γ) but not in those also involved in actin binding (δ and ε). This suggests that the initial capability of CCT to bind proteins was restricted to a very simple model and that the duplication events gave rise to CCT subunits able to bind tubulin in a more complex and specialized way. In addition, functional divergence was detected between CCT subunits proposed to bind the N-terminal domain of actin and tubulin and those that bind the C-terminal domains of the proteins. These results agree with biochemical studies that corroborated the similarity of the domains of actin and tubulin involved in binding to specific CCT subunits (Hynes and Willison 2000; Llorca et al. 2000; Ritco-Vonsovici and Willison 2000).

Two main evolutionary scenarios can be drawn from the results obtained in this work. The first result implies a specialization of CCT subunits in binding a specific protein by the fixation of amino acid substitutions by positive selection in the CCT subunit binding tubulin. On the other hand, the functional divergence between subunits involved in binding differentially N-terminal and C-terminal domain indicates a subfunctionalization of the different CCT subunits.

When amino acid substitutions were examined in those branches found to be under positive selection, we found that amino acid positions under positive selection coincided between the branches leading to CCT subunits involved in binding tubulin (fig. 3*A* and *B*), suggesting functional convergence in the amino acid composition of these subunit regions. Interestingly, amino acid residues involved in substrate binding or ATP hydrolysis are highly hydrophilic (nine amino acid changes [see *Results*]) in all the branches in which positive selection was detected, except in the branch (branch D) leading to the CCTα subunit, which showed a higher fixation of hydrophobic amino acids in different regions of the apical domain without significant concentration in substrate-binding sites or ATP-binding sites. These results suggest that the interaction between CCT subunits and their substrates is through hydrophilic amino acids, which is in agreement with structural analysis of the apo-CCT α-actin complexes, which raised the possibility that CCT may not be using amino acids located in the hydrophobic groove of the apical domain to bind to its substrates (Llorca et al. 1999*a*). Archibald et al. (2001) also highlighted the fact that many conserved amino acid motifs in the apical domain are mainly composed of charged residues. Their conclusions, together with our results, put forward the conclusion that CCTs interact with their protein substrates through a different mechanism than does GroEL chaperonin in prokaryotic cells. Interestingly, many of the amino acid residues detected under positive selection are located in or close to regions involved in ATP binding or hydrolysis. As noted by Archibald, Blouin, and Doolittle (2001), ATP hydrolysis or binding motifs include amino acid signatures that distinguish CCT subunit groups from one another. Therefore, these amino acid sites are very likely involved in the functional divergence among the different CCT subunits.

The analysis of functional divergence between the different subunits demonstrate that amino acids that are critical for functional divergence between subunits are also involved in ATP as well as substrate binding (fig. 3*C*). Additionally, when subunit pairwise comparisons were carried out between subunits involved in binding the N-terminal and those involved in binding the C-terminal domains of actin and tubulin, several but not all of the amino acids involved in substrate binding were also detected to be critical for the functional divergence between these CCT subunits. We would expect that strong functional constraints on the amino acid sites involved in ATP binding and substrate binding occurred as noted by Archibald, Blouin, and Doolittle (2001). Therefore, amino acid replacements in these sites might imply an optimization of the function, a neofunctionalization or, alternatively, a subfunctionalization. The high convergence in the selective constraints between CCT subunits involved in binding tubulin and the C-terminal domains gives support to the hypothesis of subfunctionalization after the CCT duplication events and agrees with the model proposed by Llorca et al. (2001) for substrate binding by CCT chaperonins.

Acknowledgments

We would like to acknowledge Andrew Lloyd and Avril Coghlan for careful reading and helpful suggestions on the manuscript. We are also grateful to two anonymous reviewers for very helpful suggestions to improve the manuscript. This study was supported by Science Foundation Ireland.

Literature Cited

Akashi, H. 1999. Within- and between-species DNA sequence variation and the ''footprint'' of natural selection. Gene **238**:39–51.

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. **135**:157–169.
Archibald, J. M., J. M. Logsdon, 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.
Bukau, B., and A. L. Horwich. 1998. The hsp70 and hsp60 chaperone machines. Cell **92**:351–366.
Crandall, K. A., C. R. Kelsey, H. Imanichi, H. C. Lane, and N. P. Salzman. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Mol. Biol. Evol. **16**:372–382.
Ditzel, L., J. Lowe, D. Stock, K. O. Stetter, H. Huber, and R. Huber. 1998. Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. Cell **93**: 125–138.
Ellis, R. J., and F. U. Hartl. 1999. Principles of protein folding in the cellular environment. Curr. Opin. Struct. Biol. **9**:102–110.
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.
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. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.
———. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.
Feng, D. F., and R. F. Doolittle. 1997. Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. J. Mol. Evol. **44**:361–370.
Fenton, W. A., Y. Kashi, K. Frutak, and A. L. Horwich. 1994. Residues in chaperonin GroEL required for polypeptide binding and release. Nature **371**:614–619.
Fitch, W. M. 1971. Towards defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.
Frydman, J., E. Nimmesgern, K. Ohtsuka, and F. U. Hartl. 1994. Folding of nascent polypeptide chains in a high molecular mass assembly with molecular chaperones. Nature **370**: 111–117.
Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.
Grantham, J., O. Llorca, J. M. Valpuesta, and K. R. Willison. 2000. Partial occlusion of both cavities of the eukaryotic chaperonin with antibody has no effect upon the rates of beta-actin or alpha-tubulin folding. J. Biol. Chem. **275**:4587–4591.
Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664–1674.
Gu, X., and K. Vander Velden. 2002. Diverge: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics **18**:500–501.
Gu, X., and J. Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.
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.
Gutsche, I., O. Mihalache, R. Hegerl, D. Typke, and W. Baumeister. 2000. ATPase cycle controls the conformation of an archaeal chaperonin as visualized by cryo-electron microscopy. FEBS Lett. **477**:278–282.
Huelsenbeck, J. P., and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. **28**:437–466.
Hynes, G., J. E. Celis, V. A. Lewis, A. U. S. Carne, J. B. Lauridsen, and K. R. Willison. 1996. Analysis of chaperonin-containing TCP-1 subunits in the human keratinocyte two-dimensional protein database: further characterisation of antibodies to individual subunits. Electrophoresis **17**:1720–1727.
Hynes, G. M., and K. R. Willison. 2000. Individual subunits of the eukaryotic cytosolic chaperonin mediate interactions with binding sites located on subdomains of β-actin. J. Biol. Chem. **275**:18985–18994.
Kashuba, E., K. Pokrovskaja, G. Klein, and L. Szekely. 1999. Epstein-Barr virus-encoded nuclear protein EBNA-3 interacts with the epsilon-subunit of the T-complex protein 1 chaperonin complex. J. Hum. Virol. **2**:33–37.
Kim, S., K. R. Willison, and A. L. Horwich. 1994. Cystosolic chaperonin subunits have a conserved ATPase domain but diverged polypeptide-binding domains. Trends Biochem. Sci. **19**:543–548.
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.
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. 1995. 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.
Kumar, S., K. Tamura, and M. Nei. 1993. MEGA (molecular evolutionary genetics analysis). Version 1.01. Distributed by the authors, The Pennsylvania State University, University Park.
Lewis, S. A., G. Tian, I. E. Vainberg, and N. J. Cowan. 1996. Chaperonin-mediated folding of actin and tubulin. J. Cell Biol. **132**:1–4.
Li, W.-H., Z. Gu, H. Wang, and A. Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature **409**: 847–849.
Lingappa, J. R., R. L. Martin, M. L. Wong, D. Ganem, W. J. Welch, and V. R. Lingappa. 1994. A eukaryotic cytosolic chaperonin is associated with a high molecular weight intermediate in the assembly of hepatitis B virus capsid, a multimeric particle. J. Cell Biol. **125**:99–111.
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, J. Grantham, M. Ritco-Vonsovici, K. R. Willison, J. L. Carrascosa, and J. M. Valpuesta. 2001. The 'sequential allosteric ring' mechanism in the eukaryotic chaperonin-assisted folding of actin and tubulin. EMBO J. **20**:4065–4075.
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. 1999*a*. 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. 1999*b*. 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.

Llorca, O., M. G. Smyth, S. Marco, J. L. Carrascosa, K. R. Willison, and J. M. Valpuesta. 1998. ATP binding induces large conformational changes in the apical and equatorial domains of the eukaryotic chaperonin containing TCP-1 complex. J. Biol. Chem. **273**:10091–10094.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

McCormack, E. A., M. J. Rohman, and K. R. Willison. 2001. Mutational screen identifies critical amino acid residues of beta-actin mediating interaction between its folding intermediates and eukaryotic cytosolic chaperonin CCT. J. Struct. Biol. **135**:185–197.

Nicholas, K. B., and H. B. Nicholas (1997) GENEDOC. Distributed by the author (www.cris.com/~ketchup/genedoc.shtml).

Nitsch, M., J. Walz, D. Typke, M. Klumpp, L. O. Essen, and W. Baumeister. 1998. Group II chaperonin in an open conformation examined by electron tomography. Nat. Struct. Biol. **5**:855–857.

Pappenberger G., J. A. Wilsher, S. M. Roe, D. J. Counsell, K. R. Willison, and L. H. Pearl. 2002. Crystal structure of the CCTγ apical domain: implications for substrate binding to the eukaryotic cytosolic chaperonin. J. Mol. Biol. **318**:1367–1379.

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

Saitou, N., and M. Nei. 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Schoehn, G., M. Hynes, M. Cliff, A. R. Clark, and H. R. Saibil. 2000*a*. Domain rotations between open, closed and bullet-shaped forms of the thermosome, an archaeal chaperonin. J. Mol. Biol. **301**:323–332.

Schoehn, G., E. Quaite-Randall, J. L. Jimenez, A. Joachimiak, and H. R. Saibil. 2000*b*. Three conformations of an archaeal chaperonin, TF55 from *Sulfolobus shibatae*. J. Mol. Biol. **296**:813–819.

Sharp, P. M. 1997. In search of molecular Darwinism. Nature **385**:111–112.

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

Sternlicht, H., G. W. Farr, M. L. Sternlicht, J. K. Driscoll, and K. Willison. 1993. The t-complex polypeptide 1 complex is a chaperonin for tubulin and actin in vivo. Proc. Natl. Acad. Sci. USA **90**:9422–9426.

Takezaki, N., A. Rzhetsky, and M. Nei. 1995. Phylogenetic test of the molecular clock and linearized tree. Mol. Biol. Evol. **12**:823–833.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-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.

Wang, Y., and X. Gu. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics **158**:1311–1320.

Willison, K. R. 1999. Molecular chaperones and folding catalysts: composition and function of the eukaryotic cytosolic chaperonin containing TCP1. Pp. 555–571 *in* B. Bukau, ed. Regulation, cellular functions and mechanisms. Harwood Academic Publishers, Amsterdam.

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.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568–573.

———. 2000. Phylogenetic analysis by maximum likelihood (PAML). Version 3. University College London. London.

Zhang, J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Claudia Schmidt-Dannert, Associate Editor

Accepted April 30, 2003