Testing the Neutral Fixation of Hetero-Oligomerism in the Archaeal Chaperonin CCT

Valentin Ruano-Rubio and Mario A. Fares

Evolutionary Genetics and Bioinformatics Laboratory, Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland

The evolutionary transition from homo-oligomerism to hetero-oligomerism in multimeric proteins and its contribution to function innovation and organism complexity remain to be investigated. Here, we undertake the challenge of contributing to this theoretical ground by investigating the hetero-oligomerism in the molecular chaperonin cytosolic chaperonin containing tailless complex polypeptide 1 (CCT) from archaea. CCT is amenable to this study because, in contrast to eukaryotic CCTs where sub-functionalization after gene duplication has been taken to completion, archaeal CCTs present no evidence for subunit functional specialization. Our analyses yield additional information to previous reports on archael CCT paralogy by identifying new duplication events. Analyses of selective constraints show that amino acid sites from 1 subunit have fixed slightly deleterious mutations at inter-subunit interfaces after gene duplication. These mutations have been followed by compensatory mutations in nearby regions of the same subunit and in the interface contact regions of its paralogous subunit. The strong selective constraints in these regions after speciation support the evolutionary entrapment of CCTs as hetero-oligomers. In addition, our results unveil different evolutionary dynamics depending on the degree of hetero-oligomerism. Archaeal CCT protein complexes comprising 3 distinct classes of subunits present 2 evolutionary processes. First, slightly deleterious and compensatory mutations were fixed neutrally at inter-subunit regions. Second, sub-functionalization may have occurred at substrate-binding and adenosine triphosphate-binding regions after the 2nd gene duplication event took place. CCTs with 2 distinct types of subunits did not present evidence of sub-functionalization. Our results provide the 1st in silico evidence for the neutral fixation of hetero-oligomerism in archaeal CCTs and provide information on the evolution of hetero-oligomerism toward sub-functionalization in archaeal CCTs.

Introduction

Protein dimerization and oligomerization is a universal phenomenon in organisms and proteins with different degrees of oligomerization and is present in the most evolutionarily conserved protein complexes (Marianayagam et al. 2004). Dimerized and oligomerized proteins are present in many important pathways including functionally important proteins as previously reported (Simon and Goodenough 1998; Amontiazis et al. 2004; Marianayagam et al. 2004; Ronnstrand 2004; Lockhart et al. 2006).

Whether the transition from homo-oligomeric to hetero-oligomeric protein complexes is responsible for the emergence of new functions and metabolic pathways remains largely unanswered. Some authors have proposed a link between structural stability and self-interaction of oligomers (Dunbar et al. 2004). Others have demonstrated an important effect of the abundance of oligomers on the protein–protein interaction networks geometry (Ispolatov et al. 2005). Examples for homo-oligomeric and hetero-oligomeric protein complexes coexist in organisms irrespective of their complexity. For example, heat shock proteins (Hsps), also called molecular chaperones and chaperonins such as the homo-oligomeric GroEL and the hetero-oligomeric cytosolic chaperonin containing tailless complex polypeptide 1 (CCT), perform equally complex functions and both interact with proteins showing different levels of structure complexity. The fact that Hsps interact with a significant number of proteins in the cell and that the numbers of known protein clients for the different Hsps are qualitatively different make them ideal candidates to test the relationship between the degree of hetero-oligomerism and organisms complexity.

Chaperonins are double back-to-back oriented, ringed proteins that assist de novo protein folding in most cellular compartments (Hartl 1996; Ellis 1997; Bukau and Horwich 1998; Frydman 2001; Hartl and Hayer-Hartl 2002). Two groups of chaperonins have been characterized, Group I with GroEL being the best-studied protein and Group II represented by CCTs. Both groups of chaperonins are known to share a common 60-kDa protein ancestor that branched at the base of the tree of life. Group I evolved along the bacterial lineage and eukaryotic organelles (Bukau and Horwich 1998; Ellis and Hartl 1999), whereas Group II evolved along the archaeal and eukaryal lineages (Gutsche et al. 1999; Leroux and Hartl 2000). Both types of multisubunit chaperonins are highly similar in their structural as well as domain organization. Both proteins present complexes containing a central cavity that binds unfolded polypeptides, and each subunit comprises 3 structurally and functionally distinguishable domains, namely, the apical, equatorial, and intermediate domains (Ditzel et al. 1998; Llorca et al. 1998; Nitsch et al. 1998; Llorca, McCormack et al. 1999; Llorca, Smyth et al. 1999; Gutsche et al. 2000; Llorca et al. 2000; Schoehn, Hayes et al. 2000; Schoehn, Quaite-Randall et al. 2000).

Despite their structural similarities, important functional differences exist between both types of proteins. Unlike GroEL that utilizes the ring-shaped co-chaperonin GroES to discharge bound proteins to the cavity of GroEL (Hartl 1996; Buckle et al. 1997; Kadatel 1998; Grantcharova et al. 2001). CCTs use a helical protrusion at the apical domain that emulates the function of the co-chaperone GroES (Klump et al. 1997; Ditzel et al. 1998; Nitsch et al. 1998; Llorca, McCormack et al. 1999). CCT also functions in conjunction with the hetero-hexameric chaperone prefoldin (Geissler et al. 1998; Vainberg et al. 1998). In contrast to GroEL, CCTs do not release proteins into the cavity of
the complex, but proteins remain bound to the chaperonin (Llorca, Martin-Benito, Gomez-Puertas et al. 2001). Similarly to GroEL that binds a wide range of different proteins sharing a specific GroES-like motif (Stan et al. 2006), CCT also binds a wide range of distinct proteins including cytoskeletal proteins, such as actin and tubulin, and many non-cytoskeletal proteins, including luciferase (Frydman et al. 1994), G-α transducin (Farr et al. 1997), hepatitis B virus capsid protein (Lingappa et al. 1994), cyclin E (Yang et al. 2005), the Ebsterin-Barr nuclear antigen 1 viral protein (Kashuba et al. 1999), myosin (Srikakulam and Winkelmann 1999), and the Von Hippel-Lindau tumor suppressor VHL (Feldman et al. 1999). Studies based on proteomic approaches (Gavin et al. 2002; Ho et al. 2002) have extended this list.

One of the most intriguing features of eukaryotic CCTs is their versatility in binding substrate proteins compared with GroEL (Leroux and Hartl 2000). For example, CCTs have the ability to establish hydrophobic interactions with partially folded proteins (Kluppm et al. 1997; Ditzel et al. 1998) and yet establish specialized non-hydrophobic-specific interactions with the cytoskeleton protein actin (Llorca, McCormack et al. 1999; Hynes and Willison 2000; Llorca et al. 2000; McCormack et al. 2001). This flexibility seems to be correlated with the hetero-oligomerism and sub-functionalization of eukaryotic CCT subunits (Llorca, Martin-Benito, Grantham et al. 2001; Fares and Wolfe 2003). For instance, unlike the homo-tetradecamer GroEL protein, CCTs may present up to 9 subunits per ring with different degrees of sequence divergence (Liou and Willison 1997; Sigler et al. 1998; Gutsche et al. 1999; Llorca et al. 1999; McCormack et al. 1999; Grantham et al. 2000; Llorca et al. 2000; Valpuesta et al. 2002).

The repetitive lineage-specific gene duplication and conversion in archaeal CCTs (Archibald and Roger 2002a, 2002b), with most CCTs presenting only 2–3 different subunits in each ring (Archibald et al. 1999) argue against duplication as a mean for sub-functionalization. In their insightful work, Archibald et al. (1999) proposed instead the neutral fixation of evolutionarily trapped CCT hetero-oligomers as a more plausible explanation for such a scenario. To certify the neutral fixation of hetero-oligomerism, 4 conditions have to meet our point of view: 1) accelerated fixation rates of amino acid replacements in both copies should have followed gene duplication as a result of relaxed selective constraints in at least one of the CCT subunits. Additionally, the average fixation rate of 1 subunit should have been greater than that for its paralogous subunit. 2) These substitutions should have cumulated in within-ring inter-subunit interfaces at amino acid sites that became constrained afterward leading to an evolutionary entrapment of CCT as hetero-oligomers. We distinguished 2 types of mutated sites depending on whether they are located in contact regions between subunits in the same ring (within-ring inter-subunit regions) or between subunits in different rings at their equatorial domains (between-ring inter-subunit regions). 3) Fixation of compensatory mutations at accelerated rates or by adaptive evolution in inter-subunit interfaces should have occurred to mitigate the effects of slightly deleterious mutations (e.g., mutations that were fixed at high rate despite their slight negative effect on the biological fitness of organisms) on the molecule structure or/and function. 4) Compensatory and slightly deleterious mutations should have coevolved to maintain the structural and functional characteristics of the ancestral homo-oligomer.

We tested all these evolutionary conditions in a data set including sequences from the archaeal groups Crenarchaeotes and Euryarchaeotes. We investigate this question by conducting a comprehensive in silico analysis of the selective constraints governing the evolution of the archaeal chaperonin CCT. The results obtained from this approach contribute to building the theoretical ground for the origin and evolution of hetero-oligomerism.

Materials and Methods
Sequence Alignments and Phylogenetic Analyses

We searched for all protein sequences Hsp60 homologues within the Archaea domain using the NCBI Blast service (Madden et al. 1996). We filtered out those sequences that appear to be partial (less than 500 residues long) or redundant (very similar to other sequences in fully annotated genomes). We also re-predicted the gamma subunit sequence for Methanococciides burtonii as the annotated sequence seems to be missing its amino-terminal region. The resulting subset included 100 sequences. We also excluded 4 eubacterial GroEL homologues probably acquired by lateral gene transfer (LGT). Therefore, we constructed an alignment based on a full data set of 96 sequences. Nonetheless, phylogenetic analyses revealed a lack of consistency in lineage resolution due to possible phylogenetic artifacts caused by long branches. Accordingly, we excluded those groups of sequences suspicious of presenting long-branch attraction (LBA), non-functionalization, or other artifacts that we did not consider vital for the intended downstream analyses. Consequently, the final reduced data set included just 72 sequences. For both data sets, we aligned protein sequences using the program Muscle v3.6 (Edgar 2004) with the default settings. We then aligned nucleotide sequences concatenating triplets of nucleotides according to the protein sequence alignment (alignments are available from the authors on request).

Regarding phylogenetic analyses, first we used ProtTest 1.3 (Abascal et al. 2005; Keane et al. 2006) and ModelGenerator v0.82 (Keane et al. 2006) to determine the best candidate substitution rate matrix for maximum likelihood (ML) inference. Both programs pinpointed RTREV+G+F as 1st option by many criteria despite the fact that this model was specifically devised for retro-transcriptase phylogenies (Dimmic et al. 2002). The second best model based on a different rate matrix was WAG+G (Whelan and Goldman 2001). Because both models produced similar ML topologies considering just well supported lineages, we decided to use the latter from that point on.

We ran the program Phylm v2.4.4 (Guindon and Gascuel 2003) upon the full and the reduced data sets to obtain a single ML tree estimate and 1,000 nonparametrical bootstrap topologies. Then, we used the program Consense from the PHYLIIP v3.6 package (Felsenstein 1989) to summarize those bootstrap replicates in a single fully resolved consensus tree using the extended majority rule
method. In either data set, the consensus tree seemed more reliable than the single ML estimate considering previous publications on the phylogeny of archaea (Brochier, Forterre, and Gribaldo 2005; Brochier, Gribaldo et al. 2005). Accordingly, we used consensus topologies for all downstream analyses and figures.

Additionally, we resolved an incompatibility in the order of speciation events between subunits in *Halobacteriales* species by subunit alignment concatenation, ML reconstruction, and bootstrapped consensus tree.

Testing Non-Functionalization of Very Fast–Evolving Sequences

We used the program MEGA v3.1 (Kumar, Tamura and Nei 2004) to calculate synonymous (dS) and non-synonymous (dN) change distances between very fast and moderately evolving sequences in *Methanomicrobia* and *Halobacteria* clades. We calculated the transition/transversion ratio using the estimator (κ = 1.48) returned by the program Codeml in the PAML package v3.15 (Yang 1997) and applied the Nei and Gojobori modified method (Nei and Kumar 2000). If a sequence or group of sequences are pseudogenes, the ratio dS/dN observed between them or against any other sequence outside the group must approach 1. In any case, it must be clearly greater than the ratio obtained between putative functional genes specially when measured within orthologues where there is little room for functional divergence. We also used MEGA to calculate mean residue identity between accelerated, moderately evolving archaea sequences and eukaryotic CCT.

Testing the Constancy of Amino Acid Substitution Rates after Gene Duplications

To test whether changes in substitution rates occurred after the different duplication events in archaeal CCTs, we used the 2-cluster test implemented in the program Lintree (Takezaki et al. 1995). The 2-cluster test examines the equality of substitution rates for 2 clusters linked by a node on the phylogenetic tree. In order to determine what cluster is accelerated with respect to the other, we needed to use an outgroup cluster for the analyses. Because of the high divergence levels between CCTs from eukaryotes and those from archaea, we performed this test dividing the archaea data into 2 subsets, with the 1st set including sequences from Crenarchaeotes and the 2nd set including sequences from Eurarchaeotes (fig. 1 shows the phylogenetic tree with the different groups of sequences). We used representative sequences from each CCT subunit of Crenarchaeotes as outgroup sequences for the 2-cluster test analysis in the Eurarchaeotes data set and vice versa.

Identifying Lineages and Protein Regions under Selective Constraints

To detect accelerated rates of evolution simultaneously in specific lineages of the tree and regions of the sequence alignment, we used the sliding-window–based approach (Fares et al. 2002) implemented in the program SWAPSC v1.0 (Fares 2004). Briefly, the software slides a statistically optimum window size along the sequence alignment to detect selective constraints and estimates the probability of replacements per non-synonymous sites (dS) and substitutions per synonymous sites (dN). The window size is optimized by means of using a number of simulated data sets. The standard way to measure the intensity of selection when analyzing DNA variability is by comparing dN with dS (Kimura 1977; Sharp 1997; Posada and Crandall 1998; Akashi 1999; Thulasiraman et al. 1999). The ratio between the 2 rates (ω = dN/dS) helps to elucidate if the gene has been fixing amino acid replacements neutrally (ω = 1), replacements have been removed by purifying selection (ω < 1), or mutations have been fixed by adaptive evolution (ω > 1). It has been shown, however, that ω is a poor indicator of the action of adaptive evolution (Sharp 1997; Posada and Crandall 1998). The reason is that detection of episodic positive selection may be hampered by the strong purifying selection against the majority of amino acid mutations throughout most of the evolutionary time resulting in ω << 1. Thus, ω is a conservative detector of adaptive evolution.

We used 1,000 simulated data sets in our analysis obtained using the program Evolver from the PAML package (Yang 1997). To perform the simulations, we took as initial parameters the average ω value, transition-to-transversion rates ratio, and codon-usage table generated under the Goldman and Yang (1994) (G&Y) model, using the real sequence alignment as input. The program then slides the window along the real sequence alignment and estimates dS and dN by the method of Li (1993). The program determines significance of these estimates under a Poisson distribution of nucleotide substitutions along the alignment. Along with adaptive evolution, SWAPSC also tests for accelerated rates of amino acid substitutions without the restriction of ω > 1 (e.g., due to the fixation of slightly deleterious mutations by genetic drift), saturations of synonymous sites, and hot spots (where both dS and dN are significantly high but where ω < 1).

To figure out whether mutations fixed at inter-subunit interfaces have selectively trapped CCTs into hetero-oligomerism, we investigated the mutational dynamic of these sites in the lineages after-duplication-before-speciation (ADBS) compared with that in the lineage after-duplication-after-speciation (ADAS). If hetero-oligomers were selectively advantageous, we would expect relaxed constraints ADBS permitting the fixation of changes that increase the affinity between distinct monomers. We should then detect strong selective constraints ADAS at these same sites due to their importance to maintain hetero-oligomers.

Identifying Sites under Positive Darwinian Selection by ML in CCTs

We examined whether sites surrounding within-ring inter-subunits or between-ring contacts were under adaptive evolution as to compensate for slightly deleterious mutations fixed at these regions. The combination of both sets of mutations, however, would have no effect on the protein’s function or structure (neutral mutations).
Compensatory mutations could be also under accelerated rates of evolution, and hence, they did not need to be strictly under adaptive evolution. Selective constraints were hence tested using several codon substitution models implemented in the program Codeml in the PAML package v3.15 (Yang 1997).

As we wanted to find out whether post-duplication lineages were under specific selective constraints, we...
compared the fitness of the data with 3 evolutionary models in 2 steps. First, we compared the fitness of the data with the G&Y model (Goldman and Yang 1994) that assumes a single ω value for all lineages and sites against the occurrence of different categories of ω values per site shared across all lineages. Both models are implemented in the program Codeml from the PAML package v3.15 (Yang 1997) as M0 and M3, respectively. Then, we compared the outcome of M3 with models where each post-duplication branch was allowed to have some sites under a particular ω value potentially different to the ω considered for the rest of the tree (the so-called branch-site model B [BSB]) (Yang and Nielsen 2002; Zhang et al. 2005). M0 is a special case of M3 where all site categories have the same ω value. Similarly, BSB is an extension of M3 as long as just 2 site categories are considered (Yang et al. 2000). Therefore, both comparisons can be carried out using the likelihood ratio test.

Finally, to identify possible overestimated ω values due to saturation of synonymous sites we applied a sliding-windows analysis (Fares et al. 2002) using the program SWAPSC (Fares 2004).

Analysis of Intramolecular Coevolution

To test the hypothesis of coevolution between protein regions involved in CCT within-ring inter-subunits and between-ring contacts or between these regions and nearby amino acid sites, we used the nonparametric method based on the mutual information criterion (MIC) developed by Korber et al. (1993) (hereon called MICK) as well as a parametric method coevolution analysis using protein sequences (CAPS) developed by Fares and Travers (2006).

The MI is represented by the entropies that involve the joint probability distribution, $P(s_i,s_j)$, of occurrence of symbol $i$ at position $s$ and $j$ at position $s'$ of the multiple sequence alignment. The MI-generated values range between 0, indicating independent evolution, and a positive value whose magnitude depends on the amount of covariation. In contrast, CAPS compares the correlated variance of the evolutionary rates at 2 sites corrected by the time since divergence of the 2 sequences they belong to. This method compares the transition probability scores between 2 sequences at 2 particular sites, using the blocks substitution matrix (Henikoff S and Henikoff JG 1992). Variable positions included in the alignment for both types of analyses were those that are parsimony informative (i.e., they contain at least 2 types of amino acids and at least 2 of them occur with a minimum frequency of 2). We assessed the significance of the MI values and the CAPS correlation values by randomization of pairs of sites in the alignment, calculation of their MI or CAPS correlation values, and comparison of the real values with the distribution of 1 million randomly sampled values. To correct for multiple nonindependent tests, we implemented the step-down permutation procedure in both methods and corrected the probabilities accordingly (Westfall and Young 1993). MICK is implemented in the program PIMIC (available from the corresponding author on request), and CAPS is implemented in the program CAPS (Fares and McNally 2006). Coevolution analyses were performed in each data set containing the groups of paralogues highlighted in the phylogenetic tree of figure 1. Each sub-alignment contained at least 9 sequences to ensure a minimum acceptable sensitivity for the methods used (Fares and Travers 2006).

Results

Reinvestigating Paralogy in the Phylogeny of Archaeal CCTs

Prior to the analysis of selective constraints in the archaeal CCTs, we reinvestigated the phylogenetic distribution of gene duplications and compared our study with that by Archibald et al. (2001). We used all the sequences available in the NCBI database until 1 October 2006. After applying some quality filters, we reduced the number to 100 full-length sequences. Four of them turned out to be bacterial chaperonin (GroEL) homologues putatively gained by LGT events in Methanomicrobia. The remaining 96 genuine archaeal CCTs distribute unevenly among species, ranging from a single copy to up to 5 genes in *Methanosarcina acetivorans* that also count with a GroEL homologue. When all 96 sequences were considered, the tree resolution was poor at deep branches of the Euryarchaeotes group probably caused by artifacts due to fast-evolving and poorly sampled clades indicating the possible presence of LBA effects (supplementary fig. 1, Supplementary Material online). Gene gain and loss seem to be a recurrent phenomenon in the phylogeny of archaea, as pointed out previously (Archibald et al. 2001). We also noticed the occurrence of within-archaeal LGT as the most plausible explanation of the deep paralogy of both gene copies in *Methanosphaera stadtmannae* species.

There are 4 sequences remarkably accelerated: 2 gene copies in *M. acetivorans* species and another pair of orthologues in close related species *Natronomonas pharaonis* and *Haloarcula marismortui*. There is no evidence of nonfunctionalization in both cases because synonymous versus non-synonymous substitution ratio within accelerated pairs is much lower than 1 ($d_{NS}/d_S \approx 0.1$ and 0.5 for Methanomicrobia and Halobacterial, respectively). These values are also in the range of values obtained within related paralogue groups that are putatively functional. Sequence identity indicates that these 4 sequences are more divergent from the remaining “moderately” evolving archaeal sequences than are eukarotic homologues (26% identity within tree domain compared with 35% across tree domains). Additionally, these sequences belong to species with maximum number of CCT gene copies (4 or 5). These facts indicate that these genes may well constitute separate full-fledged chaperonin systems.

The final data set comprised 72 sequences that included those genes that belong to well-supported lineages including more than 2 taxa and that allow for a tree resolution consistent with recently published phylogenies for archaeal domain speciation (Brochier, Forterre, and Gribaldo 2005; Brochier, Gribaldo et al. 2005). Nonetheless, the support for the deep branches of the tree remained weak (fig. 1). The final tree included a total of 22 sequences for Crenarchaeotes group and comprised species with 2 and 3
Neutral Fixation of Hetero-Oligomerism in Archaeal CCT

Archibald et al. (1999). The paraphyletic nature of the CCT protein complexes into 9-member rings in this group of subunits. Species with 3 subunits belonged to the genera Methanomicrobia and Euryarchaeotes. Detection of Accelerated Rates of Amino Acid Substitution Using the 2-Cluster Test Method called c-subunit underwent a 2nd duplication event in the gene coding for the c-subunit of Crenarchaeotes indicates that a subset of species underwent gene duplication was almost always followed by accelerated rates of evolution in inter-subunit contact regions presented significantly greater ω values in the branches ADAS compared with those values in the branches ADAS. In fact, SWAPSC detected these regions as significantly accelerated in branches ADAS as well. All the within-ring inter-subunits contact regions presented significantly greater ω values in the branches ADAS compared with those values in the branches ADAS. In fact, SWAPSC detected these regions as significantly accelerated ADAS and under strong purifying selection ADAS (results not shown). Differences thus suggest that sites that fixed amino acid replacements immediately after duplication became highly constrained after speciation. Some of the accelerated regions were detected in substrate-binding sites, although they were always surrounded by compensatory mutations.

Gene Duplication Followed by Accelerated Rates of Evolution in the Archaeal CCT

The Euryarchaeotes group included 50 sequences and comprised 2 distinct cases of single duplication events (2 subunits) in Thermoplasmata and Thermococci clades and 2 groups of species with 3 subunits, Methanomicrobia and Halobacteriaceae, and comprised 2 distinct cases of single duplication events (2 subunits) in Thermoplasmata and Thermococci clades (table 2; taking as reference the sequence of Thermoplasma acidophilum, accession number NP_394733, for which the 3-dimensional structure is available). Sites having undergone accelerated rates of evolution not directly involved in inter-subunit interfaces were located significantly close (less than 8 Å distance) to within-ring inter-subunits contact regions that present evidence of accelerated rates of evolution (table 2 and green space–fill structured sites in fig. 1). In figure 2a, we rotated the 3-dimensional structure to make possible the visualization of sites under selective constraints. We also verified whether CCTs became evolutionarily trapped in hetero-oligomerism by testing whether these regions were accelerated in branches ADAS as well. All the within-ring inter-subunits contact regions presented significantly accelerated rates of evolution not directly involved in inter-subunit interfaces were located significantly close (less than 8 Å distance) to within-ring inter-subunits contact regions that present evidence of accelerated rates of evolution (table 2 and green space–fill structured sites in fig. 1). In figure 2a, we rotated the 3-dimensional structure to make possible the visualization of sites under selective constraints. We also verified whether CCTs became evolutionarily trapped in hetero-oligomerism by testing whether these regions were accelerated in branches ADAS as well. All the within-ring inter-subunits contact regions presented significantly greater ω values in the branches ADAS compared with those values in the branches ADAS. In fact, SWAPSC detected these regions as significantly accelerated ADAS and under strong purifying selection ADAS (results not shown). Differences thus suggest that sites that fixed amino acid replacements immediately after duplication became highly constrained after speciation. Some of the accelerated regions were detected in substrate-binding sites, although they were always surrounded by compensatory mutations.

Accelerated Evolution of CCT Inter-Subunit Interfaces after Gene Duplication

To test the hypothesis of the origin of hetero-oligomerism as a result of accelerated rates of evolution in inter-subunit contacts after gene duplication, we applied the program SWAPSC v1.0 to detect selective constraints. The advantage of using this program over others is that, together with the analysis of $d_A$ and $d_S$, it allows testing saturation of synonymous sites and thus removing them from subsequent analyses. Our first approach was to analyze species presenting CCTs formed by 2 divergent subunits (namely, α and β; note that the nomenclature of the subunit is totally arbitrary). In order to distinguish species-specific selective constraints from those constraints imposed after gene duplication, we focused on duplicates present in more than 1 species (ancestral paralogy). Selective constraints analyses in CCTs present 2 different evolutionary scenarios in species comprising 2 distinct types of subunits compared with those with 3 different types of subunits (table 1).

Table 1
Detection of Accelerated Rates of Amino Acid Substitution Using the 2-Cluster Test Method

<table>
<thead>
<tr>
<th>Archaea group</th>
<th>Species group</th>
<th>Subunits</th>
<th>$\delta \pm SE$</th>
<th>Z^4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euryarchaeotes</td>
<td>Methanomicrobia</td>
<td>Sub 1 &lt; Sub 2 &lt; Sub 3</td>
<td>0.174 ± 0.030</td>
<td>5.788 **</td>
</tr>
<tr>
<td>Methanomicrobia</td>
<td>Sub 1 &lt; Sub 2</td>
<td>0.041 ± 0.017</td>
<td>2.343 *</td>
<td></td>
</tr>
<tr>
<td>Halobacteriaceae</td>
<td>Subα &lt; Subβ</td>
<td>0.108 ± 0.028</td>
<td>6.976 **</td>
<td></td>
</tr>
<tr>
<td>Halobacteriaceae</td>
<td>Subβ-α &lt; Suby</td>
<td>0.157 ± 0.019</td>
<td>8.223 **</td>
<td></td>
</tr>
<tr>
<td>Thermococci</td>
<td>Subα &gt; Subβ</td>
<td>0.032 ± 0.010</td>
<td>3.039 **</td>
<td></td>
</tr>
<tr>
<td>Crenarchaeotes</td>
<td>All species included</td>
<td>Subβ-α &lt; Suby</td>
<td>0.105 ± 0.027</td>
<td>3.834 **</td>
</tr>
<tr>
<td>Crenarchaeotes</td>
<td>All species included</td>
<td>Subα-β &lt; Suby</td>
<td>0.217 ± 0.026</td>
<td>8.422 **</td>
</tr>
</tbody>
</table>

Note: Only clades with significant results are shown.

a Species included in the comparison between paralogues.
b Subunits compared. Names of grouped subunits are separated by a hyphen. Subunits names are arbitrary as used in literature.
c $\delta$ is the absolute value of the difference between the mean branch lengths of the subunit groups compared.
d The normal distribution value for $\delta$. Significant values at $\alpha = 0.01$ are indicated by **.

Species with 3 subunits belonged to the genera Sulfolobus, and previous reports pinpoint the evolution of CCT protein complexes into 9-member rings in this group (Archibald et al. 1999). The paraphyletic nature of the α subunit of Crenarchaeotes indicates that a subset of species underwent a 2nd duplication event in the gene coding for subunit α, generating a subunit we name α’ and a third subunit called γ (fig. 1).

The Euryarchaeotes group included 50 sequences and comprised 2 distinct cases of single duplication events (2 subunits) in Thermoplasmata and Thermococci clades and 2 groups of species with 3 subunits, Methanomicrobia and Halobacteriaceae clades (fig. 1).

Gene Duplication Followed by Accelerated Rates of Evolution in the Archaeal CCT

One of the first questions we aimed at answering was whether gene duplication was followed by accelerated fixation rates of amino acid replacements indicating changes on selective constraints. Application of the 2-cluster test using the gamma-corrected amino acid distances between sequences supports that gene duplication was almost always followed by accelerated rates of evolution (table 1). The only exception was the case of the Thermoplasmata lineage where, despite the greater fixation rate of mutations in the β subunit compared with the α subunit, the difference was not significant. Our results, however, indicate that one of the subunit clusters has undergone greater acceleration of amino acid replacement rates than the other and therefore duplication altered selective constraints.

Accelerated Evolution of CCT Inter-Subunit Interfaces after Gene Duplication

To test the hypothesis of the origin of hetero-oligomerism as a result of accelerated rates of evolution in inter-subunit contacts after gene duplication, we applied the program SWAPSC v1.0 to detect selective constraints. The advantage of using this program over others is that, together with the analysis of $d_A$ and $d_S$, it allows testing saturation of synonymous sites and thus removing them from subsequent analyses. Our first approach was to analyze species presenting CCTs formed by 2 divergent subunits (namely, α and β; note that the nomenclature of the subunit is totally arbitrary). In order to distinguish species-specific selective constraints from those constraints imposed after gene duplication, we focused on duplicates present in more than 1 species (ancestral paralogy). Selective constraints analyses in CCTs present 2 different evolutionary scenarios in species comprising 2 distinct types of subunits compared with those with 3 different types of subunits (table 1).

Analysis of the duplication events that took place on the ancestor of the Thermoplasmata clade after gene duplication uncovers events of accelerated rates of evolution in the branch leading to subunits α and β at sites directly involved in inter-subunits contacts (table 2; taking as reference the sequence of Thermoplasma acidophilum, accession number NP_394733, for which the 3-dimensional structure is available). Sites having undergone accelerated rates of evolution not directly involved in inter-subunit interfaces were located significantly close (less than 8 Å distance) to within-ring inter-subunits contact regions that present evidence of accelerated rates of evolution (table 2 and green space–fill structured sites in fig. 1). In figure 2a, we rotated the 3-dimensional structure to make possible the visualization of sites under selective constraints. We also verified whether CCTs became evolutionarily trapped in hetero-oligomerism by testing whether these regions were accelerated in branches ADAS as well. All the within-ring inter-subunits contact regions presented significantly greater ω values in the branches ADBS compared with those values in the branches ADAS. In fact, SWAPSC detected these regions as significantly accelerated ADBS and under strong purifying selection ADAS (results not shown). Differences thus suggest that sites that fixed amino acid replacements immediately after duplication became highly constrained after speciation. Some of the accelerated regions were detected in substrate-binding sites, although they were always surrounded by compensatory mutations (fig. 2a).

In the Thermococci clade, subunit α was much more constrained than subunit β. The latter presented several regions affected by accelerated rates of fixation of amino acid replacements in the lineage ADBS (table 2). These sites had significantly lower ω values in the branch ADAS (data not shown). Evolutionarily accelerated sites not directly involved in within-ring inter-subunits contacts were
neighboring regions that are involved in inter-subunit contacts (fig. 2b).

Unlike clades of species having 2 different CCT subunits, groups of species presenting 3 different CCT subunits in Euryarchaeotes and Crenarchaeotes showed a more complex evolutionary dynamic. In these groups, regions affected by accelerated rates of fixation in branches ADBS included sites involved in within-ring inter-subunit and between-ring interfaces, sites responsible for substrate binding, and sites involved in ATP binding (fig. 1 and table 2). We also detected accelerated regions close to sites involved in substrate binding, within-ring and between-ring contacts (fig. 2c and e). Other accelerated regions always surrounded most of the accelerated sites in substrate-binding regions. However, some other regions involved in substrate binding also showed adaptive Darwinian selection and were not compensated by neighboring accelerated regions (see next section). The results on CCT presenting 3 classes of subunits suggest that sub-functionalization may have occurred in these cases in sites involved in ATP binding and substrate binding.

Episodic Darwinian Selection of Compensatory Mutations in CCTs

The next step to detect compensatory mutations was to identify mutations fixed by adaptive evolution at regions close to accelerated amino acid sites. We calculated the likelihood of the data under G&Y, M3, and the BSB codon models (supplementary table 1, Supplementary Material online). In all the comparisons, the M3 improved significantly the log-likelihood value with respect to G&Y. This indicates that CCT subunits have fixed changes under heterogeneous evolutionary constraints across codons in general.

To test for the presence of heterogeneous constraints and adaptive evolution in ADBS lineages, we compared the BSB model for each of the lineages with the M3 model. In all cases, the BSB significantly improved the log-likelihood value of the M3 model even after correcting by Bonferroni (supplementary table 1, Supplementary Material online). Positive selection was detected in almost all lineages examined especially in sites neighboring within-ring inter-subunit contacts (between 4 and 8 Å distance) that underwent accelerated fixation rates of amino acid substitutions (fig. 2a–e and table 3). Subunit α from the Thermoplasma clade showed evidence for adaptive evolution in 1 single amino acid site located in the substrate-binding region, although this single site was close to a site under adaptive evolution that is located outside the substrate-binding region, indicating compensatory effects. Apart from this example, adaptive evolution in CCTs comprising 2 distinct classes of subunits was always associated to sites neighboring inter-subunit contact regions (fig. 2a and b) with a single exception. In contrast, adaptive evolution was massively detected in substrate-binding sites in CCTs showing 3 different types of subunits in addition to inter-subunit regions (fig. 2c–e and table 3).

Table 2
Amino Acid Sites Detected As Having Undergone Accelerated Fixation Rates of Evolution in the Archaeal Chaperonin CCT

<table>
<thead>
<tr>
<th>Clade</th>
<th>Subunits</th>
<th>Amino acid sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermoplasma</td>
<td>α</td>
<td>I208, E270, D297, M298, S505, T507, I511, M512</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>S89, F195, A399, G403, A404</td>
</tr>
<tr>
<td>Thermococci</td>
<td>α</td>
<td>I142, K163, S164, D170, G403</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>T142, K143, K203, T333</td>
</tr>
<tr>
<td>Methanomicrobia</td>
<td>α</td>
<td>S141, L128, A133, K134, V201, I217, M208, Q300, I334, S339, F360, V361, A450</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>T142, K163, S164, D170, G403</td>
</tr>
<tr>
<td>Halobacteriales</td>
<td>α</td>
<td>K24, S123, A133, V167, K202, T333</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>P43, R44, S130, V136, S141, I208, E336, I337, K24, E25, R275–M277</td>
</tr>
<tr>
<td>Crenarchaeotes</td>
<td>α-α’</td>
<td>I511, M512</td>
</tr>
<tr>
<td></td>
<td>α’</td>
<td>L54, G55, E270, A299, Q300, P446, R447</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>E132, T142, A146, N272, M277, E377, T378</td>
</tr>
</tbody>
</table>

NOTE.—Amino acids’ 1-letter code is followed by a number indicating the position of sites taking the amino acid sequence of Thermoplasma acidophilum as reference (accession number NP_394733). This table includes only subunits with significant results within each group of the tree of figure 1. Here, we show only sites showing accelerated rates of evolution detected by SWAPSC in regions neighboring or located in within-ring inter-subunit interfaces (black letters), between-ring inter-subunit interfaces (underlined letters), and substrate-binding sites (white letters in a black background).
FIG. 2.—Three-dimensional localization of amino acid regions having undergone accelerated fixation rates of amino acid substitutions and positive Darwinian selection. We use the structure of *Thermoplasma acidophilum* (PDB file: 1A6D) that illustrates 2 adjacent subunits. Because of the homogeneous distribution of selective constraints in CCTs with 3 different types of subunits, only examples for sites under constraints in 2 (**a** and **b**) or combination of 2 subunits (e.g., the cluster **a**-**b**) were plotted in the structure. For example, where 3 subunits were present in a given organism, we only plotted the results of coevolution corresponding to 2 clusters. In the structure we emphasize important regions only. Regions under constraints are shown as solid sphere structures for Thermoplasmata (**a**), Thermococci (**b**), Methanomicrobia (**c**), Halobacteriales (**d**), and Crenarchaeotes (**e**). Regions under accelerated rates of evolution involved in within-ring inter-subunit interfaces, between-ring interfaces, substrate binding, and proximal amino acids to any of these regions are labeled in yellow, pink, blue, and green, respectively. Regions under positive Darwinian selection for the same functional categories are labeled in brown, red, black, and green.
percentage of compensatory mutations (fig. 3c–e). In addition to the compensatory relationship between mutations, we also detected coevolution among sites belonging to inter-subunit contacts in all the groups examined (fig. 3).

Analysis of compensatory mutations in Crenarchaeotes presented very similar results to those of the CCT with 3 different types of subunits of Halobacteriales species group. Crenarchaeotes CCT subunits have undergone fixation of coevolving amino acid replacements at regions neighboring within-ring inter-subunit contacts (fig. 3e). It is worth noticing that accelerated sites were always detected to coevolve with positively selected sites (fig. 3).

When the restriction of parsimony was removed, most of the sites detected in previous analyses as accelerated and under adaptive evolution at inter-subunit contacts were reported as coevolving. In summary, coevolution has been occurring mainly between accelerated sites and sites under adaptive evolution supporting the compensatory role hypothesis for positively selected amino acid substitutions in CCTs presenting 2 distinct classes of subunits. In contrast, CCTs with 3 distinct classes of subunits present clear evidence of hetero-oligomerisation only in inter-subunit interfaces.

### Discussion

In this study, we propose a role for gene duplication and selection in protein functional innovation. Our results studying the hetero-oligomeric molecular chaperin CCT in archaea yield interesting information about the evolution of protein complexity in terms of hetero-oligomerism versus homo-oligomerism. Initially, there is no structural reason as to why hetero-oligomerism should be more advantageous than homo-oligomerism unless this hetero-oligomerism is linked to some functional role. Example of such role is the previously shown correspondence between hetero-oligomerism and sub-functionalisation in the eukaryotic CCT (Fares and Wolfe 2003). This is strongly supported by the fact that all 8-gene copies have been kept because the rapid successive gene duplications of CCTs took place in the most recent ancestor of all eukaryotes. As ours and others’ results indicate, archaeal CCT show a strikingly contrasting evolutionary scenario compared with eukaryotic CCTs: recurrent paralogy followed by gene non-functionalization and disintegration has been a recurrent phenomenon (Archibald and Roger 2002a, 2002b). These observations combined with evidence for gene conversion in archaeal CCTs yield information supporting the role of neutral evolution in the emergence of protein complexity.

Previous work presented the coevolved interdependence between subunits as the most appealing possibility to explain the patterns of recurrent paralogy in archaeal CCTs (Archibald et al. 1999). These authors proposed that ancestral archaeal CCTs were homo-oligomers but that gene duplication was followed at some evolutionary stage by the neutral fixation of slightly deleterious mutations in one of the CCT subunits. These mutations may have been responsible for a decrease in the fitness (stability) of homo-oligomers therefore favoring fixation of hetero-oligomers. Fixation of mutations in the other parologue may have increased the stability of the hetero-oligomer outcompeting homo-oligomeric CCTs (Archibald et al. 1999). These authors proposed that ancestral archaeal CCTs became evolutionarily trapped as hetero-oligomers. A likely evolutionary explanation to such an entrapment is that the 1st set of mutations fixed in the 1st paralogue was slightly deleterious or nearly neutral, whereas the 2nd set of mutations in the 2nd paralogue was fixed by adaptive evolution conditional to the fixation of slightly deleterious mutations (apparently non-neutral mutations). Both sets of mutations had no selective value on the function of the protein when combined and were hence neutrally fixed.
FIG. 3.—Three-dimensional localization of regions under coevolution in the archaeal CCTs based on the Thermoplasma acidophilum structure (PDB file: 1A6D). Here, we highlighted coevolving sites under accelerated rates of evolution as solid spheres including sites in within-ring inter-subunits contact interfaces (yellow), substrate-binding regions (blue), between-ring contact interfaces (pink), and proximal regions (green). To visualize the sites clearly, we have rotated conveniently 3-dimensional structures. The different structures are therefore differently orientated. When sites evolved under adaptive evolution, we labeled them in brown, black, red, and green instead. Dashed-line circles surround sites belonging to the same coevolving group. Here, we illustrate a total of 5 examples in Thermoplasmata (a), Thermococci (b), Methanomicrobia (c) Halobacteriales (d), and Crenarchaeotes (e). We only present cases of coevolution involving 2 subunits and arbitrarily assigning each subunit to 1 shown in this figure.
Several lines of evidence in our study support the neutral fixation of archaeal CCT hetero-oligomerism and argue against the role of selection through sub-functionalization or neo-functionalization. First, we found that most of the mutations fixed in archaeal CCTs with 2 different classes of subunits are located within-ring inter-subunit or between-ring interfaces. Second, these regions show a significantly accelerated fixation rate of mutations. Third, accelerated fixation rates of mutations and mutations fixed by adaptive evolution have also affected peptide regions in the structure neighboring those accelerated interfaces. Finally, these 2 sets of mutations present evidence of coevolution, indicating a compensatory relationship between them.

Interestingly, we found contrasting evolutionary scenarios when comparing archaeal CCTs with 2 different classes of subunits to archaeal CCTs showing 3 different classes of subunits. Several mutations in genes coding for CCTs with 3 distinct subunits have affected substrate-binding regions and ATP-binding sites. These mutations were fixed by adaptive evolution in one of the subunits, supporting their possible involvement in sub-functionalization, and present no evidence of secondary compensatory changes. These differences between both groups of CCTs provide evidence in support of a more complex hypothesis to explain the evolution of hetero-oligomerism. A plausible evolutionary explanation for these conclusions is that once hetero-oligomerism became neutrally fixed, distinct CCT subunits started accumulating specializing mutations leading to sub-functionalization. Our data suggest that sub-functionalization may depend on the number of distinct subunits available in the complex. In this context, eukaryotic CCTs constitute an excellent example of such evolutionary process taken to completion. On the other hand, dispensability of several subunits in the hyperthermophilic archaeon suggests functional redundancy of some of the subunits and argues against sub-functionalization by formation of CCTs comprising 3 distinct subunits with a fixed geometry (Kagawa et al. 2003). Analysis of the dispensability of the different CCT genes (cct1, cct2, and cct3) in Halobacterium volcanii demonstrates dispensability of at least 2 out of the 3 genes (Kapatai et al. 2006). These authors also show that the CCT protein complex in this archaeon is a double ring with 8-fold symmetry but that the rings are mixed complexes of different subunits (Kapatai et al. 2006). Although, this study does not support a fixed geometry for a CCT complex formed from 3 distinct subunits, the different use of combinations of 2 subunits suggests that at least one of the subunits may have undergone sub-functionalization after gene duplication, something also supported by the fact that CCT3 (called CCT γ subunit in this work) cannot support growth on its own (Kapatai et al. 2006). The variability in the dispensability of the different CCT subunits in different organisms and our limited knowledge about the flexibility of the CCT complex geometry preclude stating definitive conclusions and demonstrate that there is still much to be learned about CCT complexes’ structural and evolutionarily properties in different archaeal lineages.

Coevolution between CCTs and client proteins in eukaryotic cells may have fuelled a sub-functionalization process in eukaryotic CCTs compared with archaeal CCTs. For instance, there are several examples of coevolution between eukaryotic CCTs and hetero-oligomeric proteins, also generated at the origin of the eukaryotic cell, such as actin and tubulin (Horwich and Willison 1993; Llorca, McCormack et al. 1999; Llorca et al. 2000). Tubulin binds to 5 CCT subunits in 2 different arrangements, utilizing hence the 8 CCT subunits (Llorca et al. 2000). Llorca et al. (Llorca, McCormack et al. 1999; Llorca et al. 2000) proposed sub-functionalization leading to protein-binding specialization in CCT to explain the binding and folding model with actin and tubulin. Greater number of distinct subunits would permit greater possible arrangements in which protein clients could bind CCTs and greater protein client’s versatility. This is only possible through the coevolution between these protein clients and CCTs. Then, it follows that CCTs hetero-oligomerism and protein client’s versatility are 2 tightly linked phenomena. In conclusion, hetero-oligomerism may well be related to the gain of cell protein complexity. Establishing a ground theory for the origin and evolution of hetero-oligomerism may unearth breakthrough information on the origin and evolutionary factors responsible for the emergence of cell complexity in eukaryotes.

Supplementary Material

Supplementary table 1 and figure 1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work has been supported by Science Foundation Ireland. We would also like to thank the editor and the reviewers for their helpful suggestions.

Literature Cited


