J Mol Evol (2006) 63:513–525
DOI: 10.1007/s00239-005-0289-7

JOURNAL OF MOLECULAR EVOLUTION

Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA)

Juan A. G. Ranea,[1] Antonio Sillero,[2] Janet M. Thornton,[3] Christine A. Orengo[1]

[1] Biomolecular Structure and Modelling Group, Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, UK
[2] Departamento de Bioquimica, Instituto de Investigaciones Biomedicas Alberto Sols UAM/CSIC, Facultad de Medicina, Arzobispo Morcillo 4, 28029 Madrid, Spain
[3] EMBL-EBI, Wellcome Trust Genome Campus, Hinxto, Cambridge CB10 1SD, UK

Received: 28 November 2005 / Accepted: 31 May 2006 [*Reviewing Editor:* Dr. Rafael Zardoya]

Abstract. By exploiting three-dimensional structure comparison, which is more sensitive than conventional sequence-based methods for detecting remote homology, we have identified a set of 140 ancestral protein domains using very restrictive criteria to minimize the potential error introduced by horizontal gene transfer. These domains are highly likely to have been present in the Last Universal Common Ancestor (LUCA) based on their universality in almost all of 114 completed prokaryotic (Bacteria and Archaea) and eukaryotic genomes. Functional analysis of these ancestral domains reveals a genetically complex LUCA with practically all the essential functional systems present in extant organisms, supporting the theory that life achieved its modern cellular status much before the main kingdom separation (Doolittle 2000). In addition, we have calculated different estimations of the genetic and functional versatility of all the superfamilies and functional groups in the prokaryote subsample. These estimations reveal that some ancestral superfamilies have been more versatile than others during evolution allowing more genetic and functional variation. Furthermore, the differences in genetic versatility between protein families are more attributable to their functional nature rather than the time that they have been evolving. These differences in tolerance to mutation suggest that some protein families have eroded their phylogenetic signal faster than others, hiding in many cases, their ancestral origin and suggesting that the calculation of 140 ancestral domains is probably an underestimate.

Key words: Last Universal Common Ancestor — Ancestral protein families — Genome comparison — Protein evolution — Evolutionary temperature — Protein structural domains — Ancestral protein functions

Introduction

Woese (1998, 2002) proposed that the Last Universal Common Ancestor (LUCA) was actually a community of organisms sharing genes. He also defined *evolutionary temperature* as the number of changes accepted by a given genetic subsystem and the number of variants that result. Originally, organisms were very simple and the evolutionary temperature was high. Subsystems progressively increased their complexity, reducing evolutionary temperature, as they became crystallized in a process described as *genetic annealing* by similarity to physical annealing.

It seems reasonable that as subsystems become more complex they would be less tolerant to change and thus, through time, stabilize at different stages of *cooling*. If sequence similarity between relatives is significant enough to validate the genes' homology among many distant species, it is possible to ensure their ancestral origin. Based on protein sequence

Correspondence to: J. A. G. Ranea; *email:* ranea@biochemistry.ucl.ac.uk

conservation and ubiquity through the tree of life, Woese concluded that the translation apparatus was probably the most primitive developed cellular system, crystallizing first, followed by transcription and then replication.

Also based on sequence conservation, previous comparative genome analyses of ~100 species to define the minimal gene set of the LUCA identified about 60 ubiquitous genes that could confidently be assigned to the LUCA (Koonin 2003). Most of these genes are translation-system components, and a few are basal components of the transcription system. However, it is difficult to imagine a primitive organism like the LUCA with a sophisticated translation apparatus but relatively few additional genes to perform metabolism, cellular processes, or generation of membrane envelopes. Ford Doolittle (2000) suggests that life achieved its modern cellular status much earlier than anything we can trace back and that complex metabolic and translational systems coexisted in primitive cells much as observed in modern organisms. So the fact that some ancestral protein families appear to have a modern origin might be due to the fact that they were more versatile, allowing more variation, and eroding the phylogenetic signal faster than other families (Doolittle 2000).

Therefore, evidence of sequence conservation may not be the most appropriate measure for exploring the theory of genetic annealing. Unfortunately without accurately knowing the complete set of remote homologues or the true frequencies of mutations in each gene family, parsimonious reconstruction of the LUCA's genetic pool is affected by error and subject to speculation and controversy (Koonin 2003; Mirkin et al. 2003; Whitfield 2004).

However, although it is difficult to truly know which genes were absent in the LUCA, we can be confident that a gene family with representatives in all species and kingdoms was highly likely to be present in the LUCA (Doolittle 2000). Since our ability to identify the genetic content of LUCA depends on tracing the deepest phylogenetic relationships and since protein structures are more conserved through evolution than sequences, we have exploited protein three-dimensional (3D) data to detect these very remote homologies. The application of this more sensitive methodology has allowed us to clarify the genetic and functional complexity in the LUCA, so as to analyze some parameters associated with the evolutionary temperature of its functional components.

The increase in structural annotations for genomes sequences over the last few years (Lee et al. 2003; McGuffin et al. 2004) makes it possible to perform a sound statistical analysis and trace back a significant set of protein families with very weak or undetectable sequence homology. We have used the Gene3D database (Buchan et al. 2003; Lee et al. 2003) produced by assigning sequences from 114 completed genomes from all kingdoms of life to CATH domain superfamilies. CATH is a hierarchical classification of protein structural domains (Orengo et al. 1997). Superfamilies in CATH group together protein sequences with a high probability of sharing the same ancestor, based on structural similarity, sequence similarity, and/or functional similarity. This approach has successfully detected remote homologies between highly divergent sequences (Sillitoe et al. 2005).

To estimate the tolerance to genetic diversity, or what in this work is called the evolutionary temperature, of both the superfamilies and the different cellular functional groups that they represent, we have analyzed the number of homologous genes for each domain superfamily in each species (here defined as a domain's occurrence profile; see Supplementary Fig. 1). We have also identified a set of ancestral superfamilies with a broad distribution throughout species in all kingdoms of life and, therefore, with a high probability of being present in the LUCA.

Materials and Methods

Three-Dimensional Structure Comparison

Structural similarity between domains has been identified by using a structure comparison algorithm that exploits double dynamic programming approaches to improve the detection of remote homologues (SSAP; Taylor & Orengo 1989). To validate for homology, domains are also scanned against 3D templates specific to each superfamily (CORA; Orengo 1999) and domain similarities are manually inspected. In addition, profile-based approaches are used to detect sequence patterns between relatives and functional information is extracted from public resources (e.g., COGs, GO, KEGG) and the literature. The probability of homology, based on sequence, structural, and functional similarity, is then assessed by expert curators.

Genome Structural Annotation and Occurrence Profiles

Open reading frames (ORFs) from 114 complete genomes, made up of 100 prokaryotic species (85 Bacteria and 15 Archeobacteria species) and 14 complete eukaryotic genomes (Supplementary Table 1), were structurally annotated by scanning the protein sequences against representative 1D sequence profiles (HMM) from the CATH domain structure database (Lee et al. 2003). The structural annotation data are available from release 3 of the Gene3D database (Buchan et al. 2003).

Superfamily domain occurrence profiles (Supplementary Fig. 1) were constructed for the prokaryotic sample. Of the 1278 CATH superfamilies used for genome annotation, 940 were found to be present in at least 1 of the 100 prokaryotic species. The annotation coverage was, on average, about 50% (±5%) of the genes. A superfamily occurrence profile was derived, for each superfamily, from the number of domains observed in each of the 100 prokaryotes (Supplementary Fig. 1). These profiles report the number of relatives from a particular domain superfamily occurring in each of the 100 prokaryote genomes.

Ancestral Superfamily Set Selection

A protein superfamily domain was considered ancestral if it was present in at least 90% of species from all the kingdoms and, additionally, if it was also present in at least 70% of archaeal species and 70% of eukaryotic species. This additional requirement avoids selection of superfamilies overrepresented in Bacteria but poorly represented in the smaller groups of Archaea and Eukaryotes. These thresholds were chosen because (i) they give the necessary flexibility for considering the false-negative prediction error since the sequence-based methods used to structurally annotate the genomes are only able to recognize up to 70% of very remote homologues (Lee et al. 2003; Ranea et al. 2004, 2005; Sillitoe et al. 2005); (ii) they guarantee the selection of superfamilies with ancestral origin in the LUCA since any gene present in organisms of both sides of the deepest branching, Bacteria in one branch and Archaea and Eukaryotes in the other, is very likely to have been present in the universal ancestor (Doolittle 2000); and (iii) they are high enough to eliminate, for all practical purposes, the potential error introduced by horizontal gene transfer (HGT) (see Supplementary Fig. 3).

Functional Annotation

Automatic functional annotation was performed on the 940 structural superfamilies annotated in the sample of 100 prokaryotic species, using the COG database (as of May 2003) (Tatusov et al. 2001). Each superfamily was functionally classified according to its statistically most represented functional COG subcategory. From the 940 superfamilies found to be present in prokaryotes, 726 superfamilies were functionally annotated in COG. The minimal criterion to functionally annotate a superfamily domain was that the superfamily was represented in 5% or more of the species and comprised at least five genes functionally annotated in COG. In addition, for the ancestral superfamily set, more detailed functional annotation was performed using the Pfam domain database (version 9.0; May 2003) (Bateman et al. 2002) and the literature.

Universal Distribution Percentage of Superfamilies

Universal distribution percentages were calculated using the superfamily occurrence profiles derived from the prokaryotic sample (Archaea and Bacteria). When this distribution percentage is equal to 100% the superfamily is present (universal) in all the species. In contrast, when this parameter is close to 0% the superfamily has a highly specific distribution in just a few species.

Genome Size Correlation and the Coefficient of Interspecies Gene Variation (CIGV) of Superfamilies

To calculate these parameters, we used the domain occurrence profiles derived from the 100 prokaryotic sample. Correlation coefficients between superfamily occurrence profiles and genome size were obtained by Pearson's method. In order to assess the statistical trends away from random of different functional groups, the random distribution of genome size correlations was calculated for all the superfamilies. The occurrence profiles for each of the 726 functionally annotated superfamilies found in prokaryotes were randomly shuffled amongst the 100 prokaryotic species, and the correlation with genome size of the random occurrence profiles was recalculated (Supplementary Fig. 2c).

CIGV was calculated by dividing the standard deviation over all the values in the occurrence profile for a given superfamily by the mean of the same superfamily (Wayne 1995). To avoid statistical bias in the parameter calculation, null values were not considered, nor were superfamilies present in less than 5% prokaryotic species (Supplementary Fig. 2d).

Definition of the Superfamily Functional Groups

The 726 functionally annotated superfamilies are classified into six functional groups in this work. Each one of these functional groups contains the superfamilies, annotated with the following COG functional subcategories (Tatusov et al. 2001) indicated in parentheses: translation (translation, ribosomal structure, and biogenesis), replication (replication, recombination and repair), metabolism (energy production and conversion; carbohydrate transport and metabolism; amino acid transport and metabolism; nucleotide transport and metabolism; coenzyme transport and metabolism; lipid transport and metabolism; inorganic ion transport and metabolism; secondary metabolites biosynthesis, transport, and catabolism), cellular process (cell cycle control, cell division, chromosome partitioning; nuclear structure; defence mechanisms; signal transduction mechanisms; cell wall/membrane/envelope biogenesis; cell motility; cytoskeleton; extracellular structures; intracellular trafficking, secretion, and vesicular transport; posttranslational modification, protein turnover, chaperones), poorly characterized (general function prediction only; function unknown), and transcription (transcription). The two remaining COG's functional subcategories were not considered because they represent specific functions only present in eukaryotes, such as RNA processing and modification as well as chromatin structure and dynamics.

Statistical Analysis of Superfamily Distributions

The Kolmogorov-Smirnov two-sample test in the two-tailed version for large samples was used to compare all pairs of distribution samples between different functional groups (Supplementary Table 2 and Fig. 2). The null hypothesis (that the samples have come from the same distribution) was rejected at the level of significance $p = 0.01$ (Siegel and Castellan 1988).

Results and Discussion

In this work, we first analyze the current set of completed genomes from all kingdoms of life in order to identify a set of ancestral domains, in both eukaryotes and prokaryotes. This has necessitated very detailed functional analyses based on annotations from Pfam and the literature as well as COG (Clusters of Orthologous Genes database) functional annotations (Tatusov et al. 2001).

We subsequently consider various approaches for estimating the evolutionary temperature, or genetic and functional diversity, of domain superfamilies and functional groups. These calculations were restricted to the prokaryotic sample to ensure maximum accuracy in gene identification. This is discussed in more detail below.

Superfamily Functional Distribution in the Ancestral Domain Set

Since nobody knows the average Horizontal Gene Transfer (HGT) rate, we have tried to avoid speculation on HGT estimations for superfamilies reconstruction in the LUCA by using a very conservative selection protocol (see Materials and Methods). Even though this choice risks underestimating the complexity of the LUCA by rejecting many ancestral superfamilies with species distribution percentages below the threshold.

One hundred forty superfamilies were found in practically all organisms of the three main kingdoms (Bacteria, Archaea, and Eukaryotes), and therefore considered to be ancestral (the "Ancestral superfamilies" row in Table 1). This ancestral set represents 15% of all superfamilies, 55% of all domains found in bacterial genes (last row in Table 1), and 18% of all domains in eukaryotes (data not shown), indicating that an important proportion of the currently annotated genes in Gene3D come from a relatively few phylogenetic lineages that originated before the separation of the major kingdoms.

The ancestral domains have representatives in all six functional groups from the COGs database. The translation and the metabolic functional groups comprise the majority of ancestral domains, with very similar numbers of ancestral superfamilies (48 and 46, respectively; Table 1). Metabolism has undergone a higher expansion than translation, with more new variants appearing during evolution (385 versus 90 superfamilies and 106,294 versus 14,748 domains; Table 1).

Analysis of the Cellular Functions of Ancestral CATH Superfamilies in the LUCA

We have assumed that when a domain superfamily is ubiquitous in the majority of species, it is highly probable that this structural domain was present in the LUCA, since the alternative evolutionary scenario, based on independent HGT events, is very unlikely. Additionally, domains that share the same genetic lineage are also highly likely to share similar molecular mechanisms and function. However, even though there may be conservation of very general functional or molecular mechanisms (see Todd et al. 2001), some ancestral superfamily domains show such high functional diversification that it is difficult to define a concrete function for them. For example, the ATP-loop superfamily has representatives in 230 different orthologue clusters divided into 19 different COG functional subcategories. Although the ATP-loop domain is mainly represented in metabolic pathways, this domain is also involved in disparate functional roles.

Another example is the NADH binding domain, which, broadly speaking, provides reducing energy when combined with other gene domains involved in practically all cellular functions. Knowing that these two ancestral domains are involved in ATP hydrolysis or NADH binding is not saying much about their putative functional roles in the LUCA. Therefore, although it is clear that these two domains were

Table 1. Superfamily functional annotation statistics

Functional group	Translation (J)	Replication (L)	Metabolism (M)	Cellular processes (C)	Transcription (K)	Poorly characterized (P)	Functionally annotated in COG	Not annotated in COG	Total
All superfamilies	90	56	385	107	33	55	726	214	940
Ancestral Superfamiles	48	12	46	21	6	7	140	0	140
Percentages	53%	21%	12%	20%	18%	13%	19%	0%	15%
All domains	14,748	9,392	106,294	29,297	16,563	12,244	188,538	2,089	190,627
Ancestral domains	11,695	4,384	59,195	12,863	8,131	7,819	104,087	0	104,087
Percentages	79%	47%	56%	44%	49%	64%	55%	0%	55%

Note. Number and percentages of superfamilies and gene domains (rows) distributed in different functional groups (columns). Column heads give the six functional groups (see Materials and Methods), with the one-letter code in parentheses. "All superfamilies" is the distribution in functional classes of the 940 superfamilies found in bacteria. "Ancestral superfamilies" is the distribution of the 140-ancestral superfamily set. "Percentages" (third row) are the percentage ratios of ancestral superfamilies with respect to all superfamilies in each vertical subdivision. "All domains" is the number of domains in each column subdivision. "Ancestral domains" is the number of domains in the ancestral set. "Percentages" (sixth row) are the percentage ratios of ancestral domains with respect to all domains in each column subdivision.

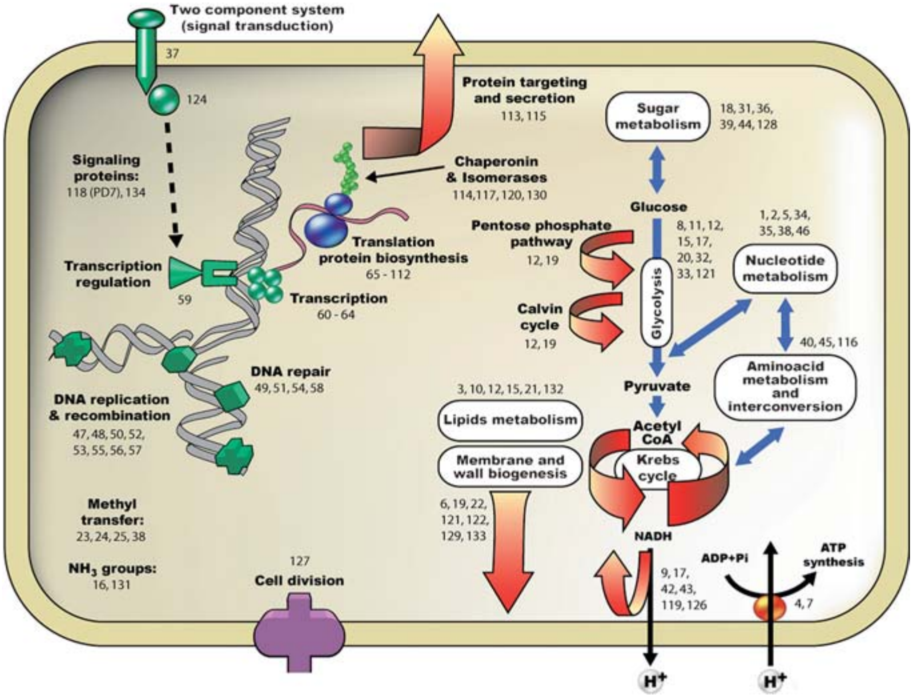

Fig. 1. Schematic representation of cellular functions represented by the ancestral set of superfamilies. The cellular and/or functional locations of the superfamilies' domains are represented by numbers. CATH identifications and functional description of all ancestral superfamilies are given in Supplementary Table 3 following the same numbering code.

present in the LUCA, their specific roles are more difficult to define. Thus, there are two issues to consider in defining ancestry: the first refers to the analysis of the domains ubiquity through all species; the second refers to the probable functions that such domains could have performed in the LUCA.

The functional analysis for all ancestral superfamilies is shown online (Supplementary Table 3) and their functions are mapped onto a scheme of the cellular processes and components in Fig. 1, with the following correlation between them. In each case, the superfamily domains are grouped by functional groups and then given consecutive numbers from 1 to 140 in Supplementary Table 3. These numbers are used to illustrate the location of the domain in Fig. 1 and are shown in brackets below. A relationship between ancestral structural domains and specific ancestral functions could not be established for all the superfamilies, therefore some superfamilies from Supplementary Table 3 are not indicated in Fig. 1. In analyzing the functions identified in the LUCA, a number of sources were consulted (Nelson & Cox 2000; Metzler 2002; Voet & Voet 2004).

There are universal CATH domains related to interconversion of sugars and synthesis of polysaccharides (18, 31, 36, 39, 44, and 128). The glycolytic or Embden-Meyerhof pathway (glucose → pyruvate) is very well represented in practically all of its steps (12, 15, 20, 32, 33, 121, 17, 8, 11). Two enzymes in this pathway (glyceraldehyde 3-phosphate dehydrogenate [20] and glycerol 3-phosphate kinase [32 and 33]) are particularly important. They catalyze the following two reversible reactions, respectively.

$$\begin{aligned}&\text{Glyceraldehyde 3-phosphate} + \text{P}_\text{i} + \text{NAD}^+ \leftrightarrow\\&\quad 1,3\text{-bisphosphoglycerate} + \text{NADH} + \text{H}^+\end{aligned} \qquad (1)$$

$$\begin{aligned}&1,3\text{-Bisphosphoglycerate} + \text{ADP} \leftrightarrow\\&\quad 3\text{ phosphoglycerate} + \text{ATP}\end{aligned} \qquad (2)$$

These reactions could be particularly important in the LUCA, as, through them, inorganic phosphate is incorporated into a high-energy bond, the adequate equilibrium NAD/NADH is partially maintained, and the net synthesis of ATP at substrate-level phosphorylation is produced. These two enzymes greatly endow the LUCA with the ability of developing in an anaerobic environment, as oxygen was not required for the synthesis of ATP (through the Krebs cycle) and NAD$^+$ could be recovered by some fermentative process.

The occurrence of transaldolase (12) and transketolase (29) also deserves a special mention, as these enzymes catalyze two reversible steps in the Calvin cycle and in the pentose-phosphate pathway (Sillero et al. 2006). The Calvin cycle would allow the LUCA to assimilate CO_2 and generate hexoses according to the equation:

$$\begin{aligned}&6\text{CO}_2 + 18\text{ATP} + 12\text{NADPH} + 12\text{H}^+ \rightarrow\\&\quad \text{hexose} + 18\text{ADP} + 18\text{P} + 12\text{NADP}^+\end{aligned} \qquad (3)$$

Related to this cycle, an additional system to capture light energy and to produce NADPH and ATP would be needed.

The pentose phosphate pathway has strong metabolic similarities to the Calvin cycle. Although the argument in favor of both pathways relates to the occurrence of these two enzymes (transketolase and transaldolase), we may suggest that a cycle with these characteristics would be very convenient for the LUCA as a way of producing sugars of different chain lengths (with three, four, five, six, or seven carbons atoms) and of allowing an alternative route for glucose use and to get reducing power, essential for biosynthetic processes (Sillero et al. 2006).

Acetyl-CoA seems to be the center of a metabolic crossroads involved in the synthesis of cholesterol and/or steroids (3, 12, 15) and synthesis and degradation of fatty acids (10, 21, 132). The Krebs cycle is only represented by domains corresponding to α-ketoglutarate dehydrogenate (17) and fumarase (1).

Concerning nucleotide metabolism, it is worth noting the occurrence of a nucleoside phosphorylase domain (34). This is a key enzyme in the purine and pyrimidine salvage pathway. It catalyzes the phosphorolysis of a nucleoside, according to the reaction

$$\text{B-ribose} + \text{Pi} \rightarrow \text{B} + \text{ribose 1-P} \tag{4}$$

where B is a nitrogenous purine or pyrimidine base.

Two more enzymes are needed to complete the salvage nucleotide cycles: nucleotidases and B-phosphorybosyl transferases (35). The latter family of enzymes catalyzes the reactions

$$\begin{aligned}&\text{B} + \text{Phosphoryhosyl Pyrophosphate} \rightarrow\\&\quad\text{B-ribose-p} + \text{ppi}\end{aligned} \tag{5}$$

On the other hand, the cleavage of a mononucleotide to a nucleoside could be carried out by a specific nucleotidase or by nonspecific phosphatases. The latter topic brings us to the point of whether LUCA benefited from a de novo purine synthesis pathway, a complex metabolic route composed of 11 enzymes and yielding IMP as the final product. It seems as if only a dihydrofolate reductase activity (38) could be indirectly related to this pathway, through its participation in one-carbon transfer reactions. Here we are faced with two alternatives: either the LUCA synthesized nucleotides by the de novo pathways or it took them from the surrounding soup in the form of bases or, most probably, nucleosides, as some of the bases are rather insoluble in aqueous media.

Enzymes catalyzing interconversion of nucleoside monophosphates are also present as exemplified by adenylosuccinate lyase (1, 2, 5), an enzyme that catalyzes the formation of AMP from adenylosuccinate. If LUCA was able to capture the main bases or nucleosides from its surroundings, interconversion of the nucleotide pathway would not be strictly needed. The presence of a ribonucleotide reductase (46) points to the possibility of the LUCA transforming ribonucleotides into the corresponding deoxyribonucleotides; the occurrence of both RNA and DNA could then be suggested from the sole existence of this enzyme.

Ancient CATH superfamilies involved in DNA synthesis, repair, ligation, and modification are represented in LUCA (47–58) (Leipe 1999). Also, domains related to the synthesis of RNA and DNA transcription are also well represented (60–64). Particularly abundant are domains related to the ribosomal particle, to protein synthesis, and with aminoacyl-tRNA synthetases (65–112). Also noteworthy is the occurrence of proteins with characteristic chaperones (114, 117, 130) or with peptidyl-propyl isomerase activity (120).

There are also domains involved in the transfer of methyl (23, 25, 38, 42) and NH_3 groups (16, 131) and in the interconversion of amino acids (40, 45, 116).

Finally, there are groups of activities related to membranes and cell wall biogenesis (6, 19, 22, 121, 122, 129, 133), transduction of protein-protein signals and gene regulation (37, 59, 118, 124, 134), protein signal recognition for protein transport (113, 115), cell division (127), electron transport (9, 17, 42, 43, 119, 126), and ATP synthase (4, 7).

From the metabolic and functional properties of the universal ancient families described above, it can be inferred that the hypothetical LUCA was endowed with important functional properties such as synthesis of DNA, RNA, or proteins probably similar to those actually operating in bacteria.

Also relevant is the important role of glucose metabolism represented by particular CATH domains in each one of the metabolic steps corresponding to sugar interchange, biosynthesis of polysaccharides, and degradation of glucose to pyruvate. Through the latter pathway, the LUCA could get ATP in an anaerobic environment. In contrast, the Krebs cycle is poorly represented in the CATH superfamilies selected as universal. In our view, the LUCA was faced with two important challenges associated with the source of amino acids and purine/pyrimidine bases or nucleosides. Most of these compounds need complex pathways to be synthesized and our analyses suggest that these are not present in the LUCA. Based on that, we are more in favor of amino acids and nitrogenous bases being present in a possible primitive soup rather than being synthesized by the LUCA. Another possibility is that different parts of these synthesis pathways were split among disparate organisms integrated in a hypothetical ancestral ecosystem. In this regard, the occurrence of two enzymes essential for the recovery of purine and pyrimidine bases and nucleosides—nucleoside phosphorylase and base-phosphorybosyl transferase—is highly relevant.

Our analysis also suggests that the LUCA may have functions related to membrane and wall structures biogenesis and, associated with them, machinery to carry out redox reactions coupled to electron transfer and synthesis of ATP. Also interesting is the presence of chaperones, domains involved in protein-protein recognition, signal transduction, gene regulation, cell division, protein signal recognition, and transport.

Ancestry and Evolutionary Temperature

The degree of universality observed for superfamilies in each functional group is shown in Fig. 2a. Superfamilies in the highest universality range (90%–100%) do not change significantly when eukaryotes are included in the analysis (data not shown), suggesting that ancestral and essential superfamilies in prokaryotes also tend to be ancestral and essential for all three kingdoms. Metabolism shows the highest representation in all the universal classes, emphasizing the important role of metabolism in bacterial adaptation as a means of generating new functional variants (Fig. 2a).

If we consider the degree of universality exhibited by different functional groups, it is clear that families involved in translation show a significant bias toward high universality compared to other functional groups (see Fig. 2b and Supplementary Table 2b and Fig. 2b), confirming the ancestral origin and essentiality of this cellular system. Metabolism shows an almost-homogeneous distribution throughout all universality ranges, while the poorly characterized group shows a significant bias toward lower universality ranges (Fig. 2b and Supplementary Table 2b and Fig. 2b).

Woese introduced the term *evolutionary temperature* to denote the number of changes accepted by a given system and the diversity of variants that result. Although it is possible to infer from Fig. 1b that translation is the most ancestral system, as proposed by Woese (1998, 2002), followed by the rest of the functional groups, differences in universality could also be explained by the differences in evolutionary temperatures or gene variation exhibited by these functional groups, as suggested by Doolittle (2000). Higher duplication rates and therefore genetic variation would make it harder to detect homologues across species and result in some families having artificially low universality values. This clearly makes the dating of cellular systems based on sequence conservation among distant species problematic.

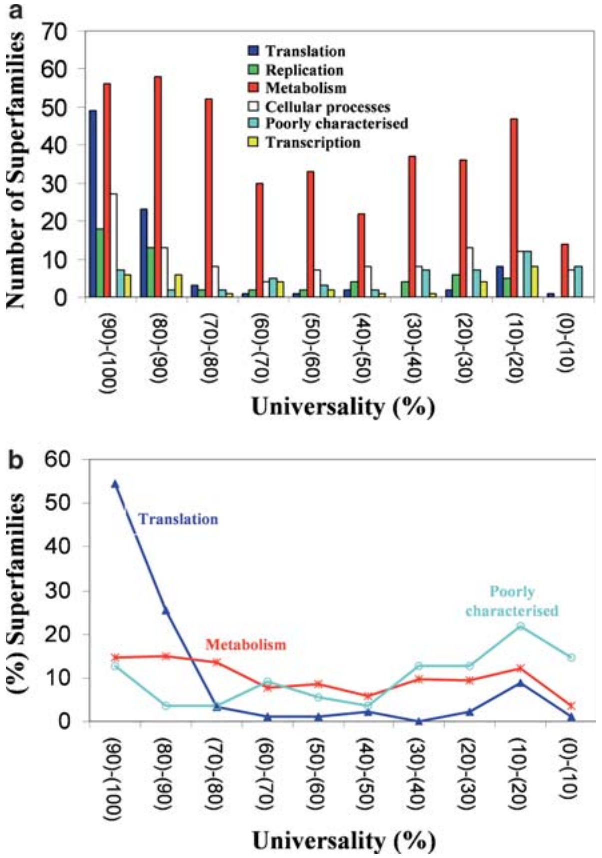

Fig. 2. Universality percentage distribution analysis of the six functional groups in bacteria, for the 726 functionally annotated superfamilies. **a** Number of superfamilies for each of the six functional groups (*y*-axis), distributed by their universal percentage values through 10 classes of universal value ranges of 10% size (*x*-axis). **b** Percentage of superfamilies (*y*-axis) distributed by universal distribution percentage classes (*x*-axis). Functional groups are also indicated.

To explore this further we devised some new approaches to measuring evolutionary temperature, which we have applied to the large set of completed prokaryote genomes (100 genomes). In contrast to eukaryotic genes, accurate identification of ORFs is possible in prokaryotes, because there is less noise from domain rearrangements, less complicated gene architectures, and an almost-complete absence of noncoding regions. In prokaryotes, there is also a strong correlation between the number of ORFs and the genetic and functional complexity, with an almost-complete absence of nonfunctional genes or pseudogenes (Mira et al. 2001; Moran 2002; Giovannoni et al. 2005). Additionally, there is a higher deletion pressure on functionally redundant genes, compared to eukaryotes, resulting in a higher correlation between the number of gene variants and functional diversification (Morett et al. 2003). All these features make prokaryotes an ideal sample for calculating the evolutionary temperature of superfamilies.

Traditionally, variation within a superfamily is measured as the number of point changes detected in protein or DNA multiple sequence alignments (MSAs). Analysis of sequence variation gives an estimate of the mutation acceptance in terms of entropy or percentage conservation in the alignments (Valdar 2002). By contrast, little has been done to estimate the tolerance to genetic versatility of families

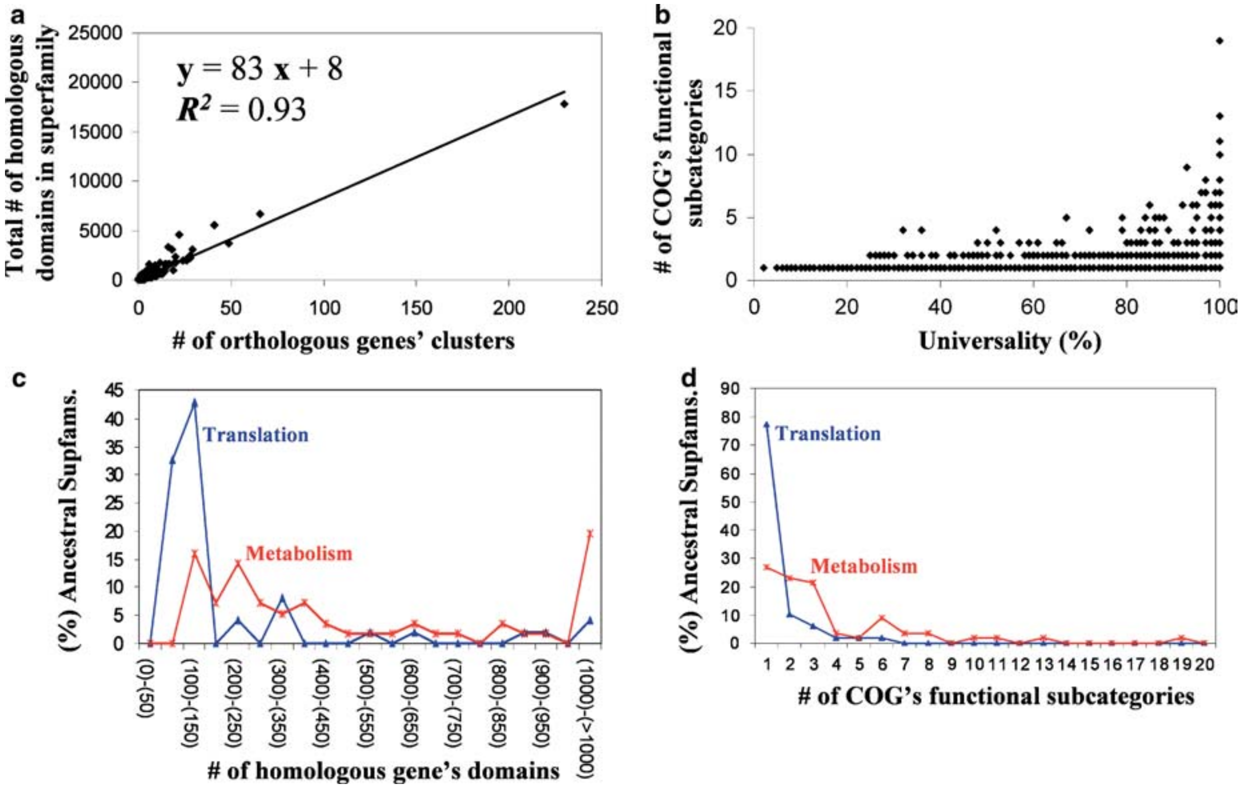

Fig. 3. Duplication rates, functional diversification, and universal percentage distribution for the 726 functionally annotated superfamilies. **a** Number of homologous domains (y-axis) versus number of different COG's orthologous gene clusters (x-axis) in each superfamily. **b** Number of COG functional subcategories (y-axis) versus universality percentage (x-axis) for each superfamily. For the ancestral set of the 46 superfamilies in metabolism (red line) and the 48 superfamilies in translation (blue line). **c** Percentage of superfamilies (y-axis) distributed by homologous domain number classes (x-axis). **d** Percentage of ancestral superfamilies (y-axis) distributed by number of COG functional subcategories (x-axis).

or functional systems, and so we have sought to take this into account in devising new measures of evolutionary temperature. Here we assess four parameters for measuring the evolutionary temperature of superfamilies and functional groups. The relationship between the evolutionary temperature and the ancestry of the different groups is analyzed below.

Superfamily duplication rates and functional diversification. The number of homologues within a superfamily can be used as an estimate of the superfamily's duplication rate during evolution. When the numbers of homologous domains are plotted against the number of different COG orthologue clusters (Fig. 3a), for the 726 prokaryotic superfamilies annotated in the COG database, we see a high correlation between superfamily duplication rates and their functional diversification, suggesting that gene duplication is highly correlated with the provision of new functional variants in prokaryotic cells and, therefore, a superfamily's evolutionary temperature.

The number of different COG subcategory functions exhibited by domains within a superfamily, as an estimation of the functional diversification of the superfamily, is plotted against the number of homologous domains for each of the 726 superfamilies functionally annotated in prokaryotes (Fig. 3b). Superfamilies with higher universal distribution percentages exhibit a wide range of functional diversification values, while the less universal or more specific superfamilies show lower levels of functional diversification (see the x-axis in Fig. 3b). Figure 3b shows that some ancestral superfamilies (with high universality) have remained in the same functional niche, throughout evolution, while others have remained highly creative, providing new functional variants.

To understand these different evolutionary behaviors in the most universal protein families, we compared the ancestral superfamilies involved in translation and metabolism, as these two groups represent almost 70% of all ancestral superfamilies, and the number of superfamily domains (48 metabolic and 46 translation; see Table 1) are quite similar and large enough to allow significant statistical comparison.

When the number of superfamily homologues (equivalent to duplication rates) is compared between the ancestral set of metabolic and translational superfamilies, there is a clear difference between these two functional groups (Fig. 3c). Ancestral metabolic superfamilies show a significant bias (with a confident $p < 0.001$, Kolmogorov-Smirnov test; see Materials and Methods) toward duplication rates higher than those associated with translation (Fig. 3c). The frequency distributions in different COG classes, for these two ancestral groups, also show significantly different shapes (with a confident $p < 0.001$; Fig 3d). Ancestral superfamilies in translation show low functional diversification, with about 80% of the superfamilies having only one COG functional subcategory assigned to them (Tatusov 1997; Koonin 2003; Ranea 2004). This contrasts with metabolism, where 73% of the ancestral superfamilies have functionally expanded into two or more COG functional subcategories. These results reveal disparate rates of genetic duplication and functional diversification of these two ancestral groups, throughout evolution.

Superfamily occurrence profiles and genome size correlation. Although increasing prokaryotic genome size implies increasing genetic and functional complexities, genome size is not related to species lineage or ancestry in prokaryote (Ochman et al. 2000). That is, prokaryotes with the most complex genomes are not necessarily the most evolved. For example, *Buchnera* is a species closely related to *E. coli*, but its genome is four times smaller. *Buchnera* has recently adapted its genome to efficiently exploit a symbiotic relationship with aphids (Shigenobu et al. 2000), by a process described as evolution by reduction (Dobrindt and Hacker 2001; Moran 2002). Although *Buchnera* is four times less complex than *E. coli*, this bacterium is not four time more ancient. Rather, bacteria change their genome size in response to different environments or reproduction strategies (Dufresne et al. 2005; Giovannoni et al. 2005; Ranea 2005).

However, genome size variation is the final consequence of gene gains and losses at the superfamily and functional group levels (Ranea et al. 2005). Therefore analyzing the correlation of genetic variation with genome size can provide another estimation of superfamily evolutionary temperature, since it measures the tolerance to genetic variation of the superfamilies and functional groups as a function of the genome size adaptation in each organism.

To explore this, the correlation between superfamily occurrence profiles and bacterial genome sizes was calculated using Pearson's correlation coefficient. The largest superfamilies, with high numbers of homologues, tend to have expansions highly correlated with genome size (Fig. 4a), while superfamilies varying independently of genome size (low size correlation coefficients, e.g., < 0.2) show low duplication rates (Nimwegen 2003; Ranea 2004).

Except for translation, the superfamily frequency distribution in genome size correlation bins shows a clear bias away from the random distribution expected for all the functional groups (Fig. 4b and Supplementary Table 2c and Fig. 2c). Replication shows a distribution intermediate between random and remaining distributions, while metabolic distribution is the most distant from random and most closely correlated with genome size variation.

The separation between translation and metabolism is even more pronounced when we consider only the ancestral superfamilies (Fig. 4c), indicating that the dynamics of superfamily expansion in translation has been practically independent of genome size variation since before the separation of the three kingdoms, while in the ancestral metabolic set, genetic variation has generally contributed to prokaryotic genome size adaptation (Ranea et al. 2004).

Superfamily coefficient of interspecies gene variation. Although two superfamilies can show similar duplication rates, and therefore similar average numbers of homologues, they can differ significantly in the distributions of homologous genes throughout different species (see examples of two real superfamilies in Figs. 5a and b). The deviation from average in the number of paralogues in each organism can be estimated by dividing the standard deviation by the mean. This statistical ratio is called the coefficient of variation (CV) (Wayne 1995), and in the context of this work it is referred to as the coefficient of interspecies gene variation (CIGV) (see Materials and Methods).

Superfamilies with high CIGV values seem to be more adaptable (hotter evolutionary temperature), changing the number of gene variants with the specific functional requirement of each species, expanding the number of homologous gene domains in some bacteria and reducing their representation in others (Fig. 5a). On the contrary, superfamilies with an almost-constant number of homologues per species (low CIGV values) exhibit low flexibility to accept or reject genetic variants (colder superfamilies; Fig. 5b). The different degrees of genetic versatility and, therefore, adaptability, measured by the CIGV, also therefore reflect evolutionary temperature.

The superfamily distribution of the functional groups in seven different CIGV classes shows similar shapes except for translation (Supplementary Table 2d and Fig. 2d) (Ranea 2004). In the translation group the distribution is significantly biased toward the lowest CIGV values. These disparate distribution shapes are particularly pronounced when the ancestral sets of superfamilies in the translation and metabolic groups are compared (Fig. 5c).

The ancestral metabolic superfamily distribution shows a significant bias toward higher CIGV values compared to the ancestral translation set. Since both ancestral subsets have apparently evolved throughout a similar evolutionary time, these differences seem to

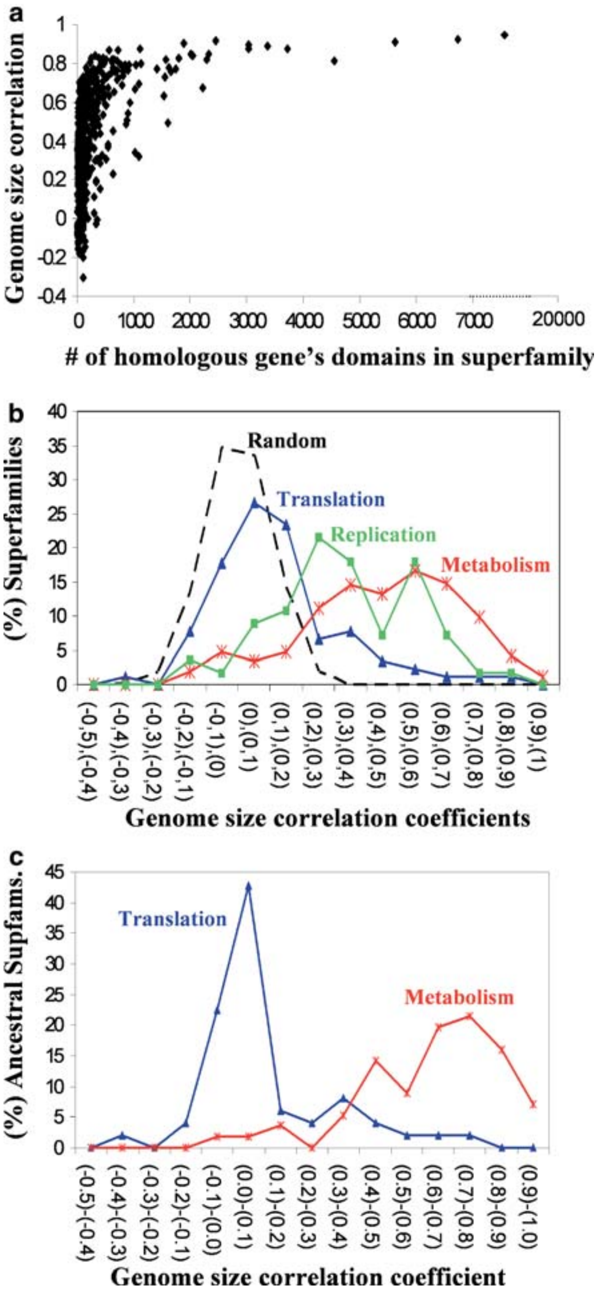

Fig. 4. Size correlation analysis. **a** For the 726 functionally annotated superfamilies: genome size correlation coefficient values (*y*-axis) plotted against the number of domains (*x*-axis) for each superfamily. **b** For different functional groups: superfamily percentages (*y*-axis) distributed by genome size correlation coefficient classes (*x*-axis) for the random distribution of all superfamilies (black dashed line; see Materials and Methods), translation (blue line), replication (green line), and metabolism (red line). **c** For the ancestral set of the 46 superfamilies in metabolism (red line) and the 48 superfamilies in translation (blue line): percentage of ancestral superfamilies (*y*-axis) distributed by genome size correlation class (*x*-axis).

be mainly due to the higher genetic adaptability of the ancestral metabolic superfamilies to the specific functional requirements of different organisms. On the contrary, translation shows a frozen profile with an extremely low deviation in the number of homologous genes per species (Ranea 2004).

Rates of superfamily innovation in the functional groups. Although the universal presence of a given superfamily is proof of its ancestry, low representation is not proof to the contrary, as a superfamily could be present in the ancestor but maintained in only a few cellular lineages and lost by deletion in the rest (Doolittle 2000). However, although percentage universality is not a good measure of how old a superfamily is, it could be a good indicator of the degree of species specificity of a superfamily.

Figure 6 shows the cumulative proportion of superfamilies within each functional group, in different percentage universality bins. The distributions (referred to here as ‘‘innovation lines’) observed for each functional set reflect the versatility of the functional groups in adapting to speciation by creating new, more species-specific protein families. Diagonals would be expected if the superfamily innovation has been directly proportional, during evolution, to the different speciation processes, while large deviations above this central diagonal suggest lower rates of innovation than proportional (colder functional groups), and deviations below the diagonal mean higher rates of innovation than proportional (hotter functional groups).

As reported by Woese, the majority of superfamilies involved in translation (triangles in Fig. 6) seem to have an ancestral origin and few new species-specific superfamilies seem to have appeared for a long time. The innovation line for translation shows a sharp fall, from 80% to 20% universality, reflecting the bias between the number of eubacterial and the number of archaeal species in the sample used, and mirrors the differences in translation between these two phyla. After translation, replication shows slightly more flexibility in accepting superfamily innovation. Metabolism, cellular processes, and transcription show innovation rates closer to the proportional-trend central diagonal line (Fig. 6). However, the transcription set must be carefully considered due to its low statistical weight (few superfamily domains associated with this set; Table 1) and its particular superfamily composition in this work. The transcriptional group as defined here, in addition to the RNA-polymerase enzymes and related domains considered in former analyses (Woese 1998; Koonin 2003), also includes superfamilies involved in gene regulation such as different transcription factor domains (e.g., λ-repressor, winged helix repressor, and the homeodomain-like DNA binding domain).

From their innovation rates, metabolism and cellular processes, in particular, seem to have had an important and constant role in speciation at all levels

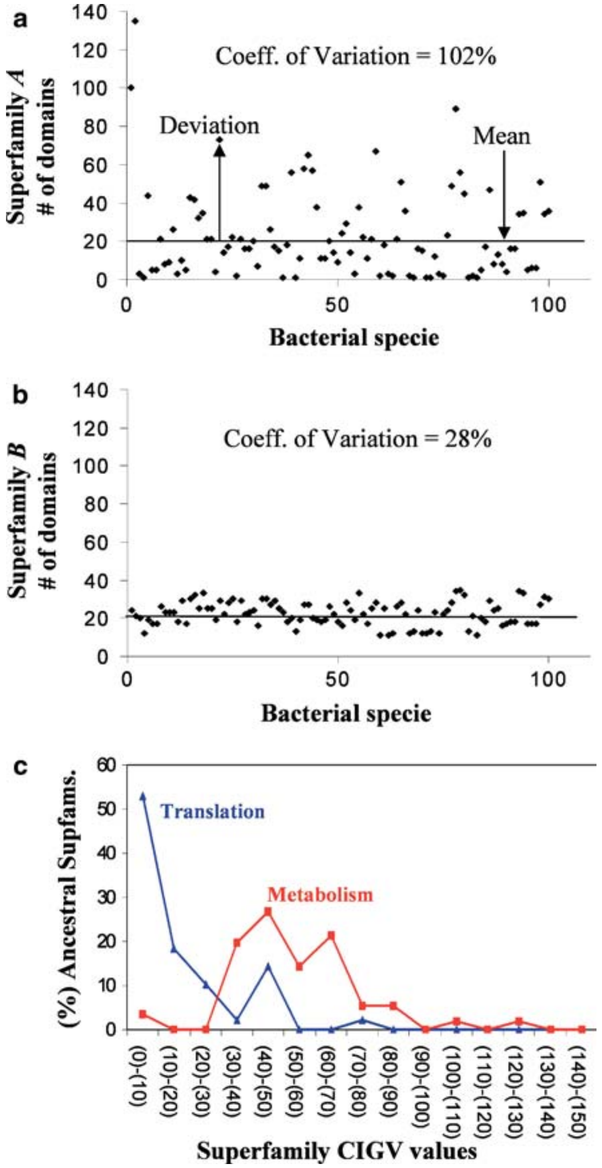

Fig. 5. Superfamily coefficient of interspecies gene variation (CIGV) analysis. **a, b** Examples of two real specific superfamilies, *A* and *B*, with similar numbers of domains (similar means) and different standard deviation values. (a) Superfamily *A*: chaperone (heat shock protein 90-like) superfamily domain (CATH code, 3.30.565.10; CATH-PDB code, 1ah600). (b) Superfamily *B*: nucleic acid-binding protein superfamily domain (CATH code, 2.40.50.140; CATH-PDB code, 1ckmA2). (a, b) Plots: number of domains (*y*-axis) in each of the different 100 prokaryote species (x-axis). **c** For the ancestral set of the 46 superfamilies in metabolism (red line) and the 48 superfamilies in translation (blue line): percentage of ancestral superfamilies (*y*-axis) distributed by superfamilies' CIGV percentage classes (*x*-axis).

of the species tree in bacteria. Poorly characterized superfamilies show an innovation line slope below the constant rate, pointing to higher rates for superfamily specificity than any other group. Again, ancestry is related to colder evolutionary temperatures reflected by low superfamily innovation rates, identified here, particularly, in translation.

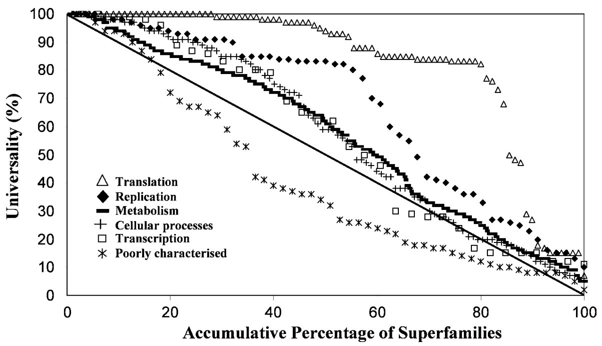

Fig. 6. Analysis of the innovation rates of superfamilies for the six functional groups. Considering all superfamilies for each of the six functional groups, the plot displays the universal distribution percentage of the superfamilies (*y*-axis) against the accumulated percentage of superfamilies (*x*-axis). Different functional groups, codes, and the proportional trend line (–45% slope diagonal) are also indicated in the plot (black line).

Conclusions

The assignment of completed genome sequences to CATH structural domain superfamilies has provided a sensitive method to derive a more realistic distribution of superfamilies in distant species. From this annotation we know that the LUCA, or the primitive community that constituted this entity, was functionally and genetically complex (Table 1, Fig. 1, Supplementary Table 3), supporting the theory that life achieved its modern cellular status long before the separation of the three kingdoms (Doolittle 2000). Contrary to analyses based purely on sequence conservation and universal ubiquity throughout all species, which suggested a simple LUCA with translation and few other genes (Koonin 2003), with the application of a more sensitive method to detect remote homology, we can affirm that the LUCA held representatives in practically all the essential functional niches currently present in extant organisms, with a metabolic complexity similar to translation in terms of domain variety. The criteria applied to select ancestral superfamilies are stringent in order to ensure a confident sample of ancestral representatives. The selected 140 ancestral domains are analogous to spots in a "connect-the-dots" picture, suggesting the presence of other hidden partners in the LUCA's functional composition. Likewise, the true genetic and functional content of the LUCA has, with all probability, been underestimated. Even if the ancestral domain set in the LUCA was much larger than the set considered here, the functional analysis of this selected sample reveals that the LUCA comprised functions for (i) replication, transcription, and translation; (ii) the use of glucose and other sugars; (iii) the assimilation of amino acids and nucleosides/bases; (iv) the synthesis of ATP both by substrate-level phosphorylation and through redox reactions coupled to membranes; (v) signal transduction and gene regulation; (vi) protein modification; (vii) protein signal recognition, transport, and secretion; (viii) protein folding assistance; and (ix) cell division.

The comparison of evolutionary temperature among the ancestral functional groups in the LUCA, particularly between translation and metabolism, indicates that different functional systems in LUCA have undergone disparate rates of functional diversification and genetic expansion through the same evolutionary period (Figs. 3c and d, 4c, and 5c). As a result of these differences, some ancestral superfamilies have stayed in the same functional niche, while others have remained highly creative, providing many new functional variants throughout similar evolutionary timescales.

As already discussed, these data call into question the methodology applied for dating the emergence of cellular systems based on Woese's (1998, 2002) annealing hypothesis, since these differences seem to correspond more to the specific features and adaptabilities of the different functional systems than to the time during which the systems have been evolving since they emerged in evolution. For example, the use of a complex interlinked machine for protein biosynthesis—the ribosome—compared with the more modular metabolic pathways, is likely to have been a restraint on protein variation in the translational system. However, metabolism shows the highest representation in all the universal classes, emphasizing the important role of metabolism in prokaryotic adaptation as a means of generating new functional variants (Figs. 2a and 6). In other words, prokaryotic metabolism will probably never evolve toward any highly integrated and complex single system universally distributed as has occurred in translation, but it will always be at the core of prokaryote competition and diversification. Furthermore, whatever was the real order of appearance of the different cellular systems in evolution, the progressive evolution and subsequent annealing of the different functional systems may have occurred prior to a time point we can reliably trace back to, at least with current genome comparison analyses.

In summary, by combining analyses of multiple features, including correlation of duplication rates with functional diversity, flexibility of a superfamily's gains and losses in contributing to genome size and species adaptation, and, finally, the ability of a functional group to be innovative and contribute new species-specific functional variants, we have revealed a spectrum of superfamilies and functional groups from the evolutionarily cooler ones involved in translation through to the hotter metabolic and poorly characterized ones which have enriched the functional repertoire of bacteria in response to environmental stimuli. However, this property, by giving rise to greater genetic and functional diversity, is associated with greater sequence divergence, making it harder to trace evolutionary relationships in hotter groups (e.g., metabolism) and suggesting that the predicted absence of some of these superfamilies in the LUCA may simply reflect their hotter evolutionary temperature.

The metaphor arising from this analysis is of a LUCA comprised of "soft" and "hard" parts just as human bodies are made of bones and tissues. When anthropologists discovered ancient bones they did not surmise our ancestors' appearance as skeletons. They imagined bones surrounded by muscles and other tissues more susceptible to decomposition during time. Even using very conservative selection criteria to identify ancestral superfamilies, the LUCA shows a similar number of metabolic and translational protein domains. Since metabolic superfamilies generally have hotter evolutionary temperatures, it is quite probable that an underestimation of true ancestral superfamilies affects this group more significantly. Conversely, it is highly probable that the metabolic domains were thus more represented than the translational ones in the genetic-based organisms of the most primitive times (Castresana 2001).

Acknowledgments. We would like to thank Beatriz Simas Magalhaes for her useful advice and comments, Stathis Sideris for help with the figures, and Corin Yeats for text review. This work was supported by grants from the MRC (Christine A. Orengo) and European Union (Juan A. G. Ranea). A.S. was a visiting professor at UCL (from UAM) aided by the Spanish Ministry of Education and Science and supported by grants from Direccion General de Investigacion Cientifica y Tecnica (08/0021.1/2001) and Instituto de Salud Carlos III, RMN (C03/08) Madrid, Spain.

References

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. Nucleic Acids Res 30:276–280

Buchan DW, Rison SC, Bray JE, Lee D, Pearl F, Thornton JM, Orengo CA (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res 31:469–473

Castresana J (2001) Comparative genomics and bioenergetics. Biochim Biophys Acta 1506:147–162

Dobrindt U, Hacker J (2001) Whole genome plasticity in pathogenic bacteria. Curr Opin Microbiol 4:550–557

Doolittle WF (2000) The nature of the universal ancestor and the evolution of the proteome. Curr Opin Struct Biol 10:355–358

Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6:R14

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245

Koonin EV (2003) Comparative genomics minimal gene-sets and the last universal commonancestor. Nat Rev Microbiol 1:127–136

Lee D, Grant A, Buchan D, Orengo CA (2003) Structural perspective on genome evolution. Curr Opin Struct Biol 13:359–369

Leipe DD, Aravind L, Koonin EV (1999) Did DNA replication evolve twice independently? Nucleic Acids Res 27: 3389–3401

McGuffin LJ, Street SA, Bryson K, Sorensen SA, Jones DT (2004) The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. Nucleic Acids Res 32:D196–D199

Metzler DE, ed (2002) Biochemistry. The chemical reactions of living cells, 2nd ed. Academic Press, New York

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583–586

Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P (2003) Systematic discovery of analogous enzymes in thiamine biosynthesis. Nat Biotechnol 21:790–795

Nelson DL, Cox MM, eds (2000) Lehninger principles of biochemistry, 3rd ed. Worth, New York

Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19:479–484

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Orengo CA (1999) CORA—topological fingerprints for protein structural families. Protein Sci 8:699–715

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108

Ranea JA, Buchan DW, Thornton JM, Orengo CA (2004) Evolution of protein superfamilies and bacterial genome size. J Mol Biol 336:871–887

Ranea JA, Grant A, Thornton JM, Orengo CA (2005) Microeconomic principles explain an optimal genome size in bacteria. Trends Genet 21:21–25

Ranea JA (2005) Micro(be)-economics. Heredity 96:337–338

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp APS. Nature 407:81–86

Siegel S, Castellan N (1988) Nonparametric statistics for the behavioural sciences, 2nd ed. Anker JD (ed). McGraw-Hill International Editions, Singapore

Sillero A, Selivanov VA, Cascante M (2006) Pentose phosphate and Calvin cycles: similarities and three-dimensional views. Biochem Mol Biol Edu 34:275–277

Sillitoe I, Dibley M, Bray J, Addou S, Orengo C (2005) Assessing strategies for improved superfamily recognition. Protein Sci 14:1800–1810

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28

Taylor WR, Orengo CA (1989) Protein structure alignment. J Mol Biol 208:1–22

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies from a structural perspective. J Mol Biol 307:1113–1143

Valdar WS (2002) Scoring residue conservation. Proteins 48:227–241

Voet D, Voet J, eds (2004) Biochemistry, 3rd ed. Wiley & Sons, New York

Wayne WD (1995) Biostatistics 6rd ed. Wiley, New York

Whitfield J (2004) Origins of life: born in a watery commune. Nature 427:674–676

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci USA 99:8742–8747