Origins of Life and Evolution of Biospheres (2005) 35: 537–554

THE LAST COMMON ANCESTOR: WHAT'S IN A NAME?

LUIS DELAYE, ARTURO BECERRA and ANTONIO LAZCANO*
Facultad de Ciencias, UNAM Apdo. Postal 70-407 Cd. Universitaria, 04510 Mexico D.F. Mexico
*(*author for correspondence, e-mail: alar@correo.unam.mx)*

(Received 19 October 2004; accepted in revised form 9 April 2005)

Abstract. Twenty completely sequenced cellular genomes from the three major domains were analyzed using twice one-way BLAST searches in order to define the set of the most conserved protein-encoding sequences to characterize the gene complement of the last common ancestor of extant life. The resulting set is dominated by different putative ATPases, and by molecules involved in gene expression and RNA metabolism. DEAD-type RNA helicase and enolase genes, which are known to be part of the RNA degradosome, are as conserved as many transcription and translation genes. This suggests the early evolution of a control mechanism for gene expression at the RNA level, providing additional support to the hypothesis that during early cellular evolution RNA molecules played a more prominent role. Conserved sequences related to biosynthetic pathways include those encoding putative phosphoribosyl pyrophosphate synthase and thioredoxin, which participate in nucleotide metabolism. Although the information contained in the available databases corresponds only to a minor portion of biological diversity, the sequences reported here are likely to be part of an essential and highly conserved pool of proteins domains common to all organisms.

Keywords: last common ancestor, cenancestor, RNA/protein world, progenote

1. Introduction

One of the major achievements of molecular cladistics has been the evolutionary comparison of small subunit ribosomal RNA (rRNA) sequences, which has allowed the construction of an unrooted tree in which all known organisms can be grouped in one of three major (apparently) monophyletic cell lineages: the eubacteria, the archaebacteria, and the eukaryotic nucleocytoplasm, now referred to as new taxonomic categories, i.e., the domains *Bacteria*, *Archaea*, and *Eucarya*, respectively (Woese *et al.*, 1990). The variations of traits common to these major groups can be easily explained as the outcome of divergent processes from an ancestral lifeform that existed prior to the separation of the three major biological domains, i.e., the last common ancestor (LCA) or cenancestor (Fitch and Upper, 1987). No paleontological remains will bear testimony of its existence, as the search for a fossil of the cenancestor is bound to prove fruitless. However, insights on its nature can in principle be deduced from the molecular record.

From a cladistic viewpoint, the LCA is merely an inferred inventory of features shared among extant organisms, all of which are located at the tip of the branches of molecular trees. From an evolutionary point of view, however, it is reasonable

to assume that at some point in time the ancestors of all forms of life must have been less complex than even the simpler extant cells. However, the conclusion that the LCA was a progenote, i.e., a hypothetical biological entity in which phenotype and genotype still had an imprecise, rudimentary linkage relationship (Woese and Fox, 1977), was disputed some time ago when the analysis of homologous traits found among some of its descendants suggested that it was not a direct, immediate descendant of the RNA world, a protocell or any other pre-life progenitor system. Under the implicit assumption that lateral gene transfer (LGT) had not been a major driving force in the distribution of homologous traits in the three domains, it was concluded that the LCA was a complex organism, much like extant bacteria (Lazcano *et al.*, 1992; Lazcano, 1995).

A decade ago many were convinced that the LCA was very much like extant prokaryotes, but the inventory of shared traits based on sequence comparisons was small. It was surmised that the sketchy picture developed with the limited data bases would be confirmed by completely sequenced cell genomes from the three primary domains. This has not been the case: the availability of an increasingly large number of completely sequenced cellular genomes has sparked new debates, rekindling the discussion on the nature of the ancestral entity (Doolittle, 2000). This is shown, for instance, in the diversity of names that have been coined to describe it: progenote (Woese and Fox, 1977), cenancestor (Fitch and Upper, 1987), LUCA, a term first coined to describe the last universal common ancestor (Philippe and Forterre, 1999), and then as an acronym for the last universal cellular ancestor (Forterre, 2002), and LCC, last common community (Line, 2002), among others. These terms are not truly synonymous, and they reflect the current controversies on the nature of the universal ancestor and the evolutionary processes that shaped it.

1.1. Lateral gene transfer and the reconstruction of early cell evolution

In the past few years the analysis of an increasingly large number of completely sequenced cellular genomes has revealed major discrepancies with the topology of rRNA trees. Very often these differences have been interpreted as evidence of lateral gene transfer (LGT) events between widely separated species, questioning the feasibility of the reconstruction and proper understanding of early biological history (Doolittle, 1999, 2000). There is clear evidence that genomes have a mosaic-like nature whose components may come from many different phylogenetically separated donor species (Ochman *et al.*, 2000; Zhaxybayeva and Gogarten, 2004; Zhaxybayeva *et al.*, 2004). Depending on their different advocates, a wide spectrum of mix-and-match recombination processes have been described, ranging from the lateral transfer of few genes via conjugation, transduction or transformation, to cell fusion events involving organisms from the same or even different domains (Rivera and Lake, 2004).

Defining the nature of LCA is one of the central goals of the study of the early evolution of life on Earth, and several attempts have been made in this direction. Proper description of the LCA is still an unfinished task, mainly because of the complexity of the evolutionary process that connect extant organisms with it, the lack of a full understanding of the major evolutionary events that have taken place along the history of life on Earth, and the limitations inherent of the methodological attempts to reconstruct its nature. In this paper, we survey some of the difficulties encountered in the characterization of the last common ancestor, and summarize ongoing discussions on its nature. We also attempt a reconstruction of the gene complement of the LCA based on the conservation of proteins in a database of twenty completely sequenced cellular genomes from the three major domains, from which endosymbionts and obligate parasites were excluded.

2. Material and Methods

To avoid biases in the backtrack characterization of the LCA due to secondary gene losses, a sample of complete proteomes from the three domains of life from non-endosymbiotic or non parasitic species was downloaded from the Kyoto Encyclopedia of Genes and Genomes KEGG data base (ftp://ftp.genome.ad.jp/pub/kegg/) (Kanehisa *et al.*, 2000). The following species were analized: Bacteria, *Bacillus subtilis, Streptococcus pneumoniae, Thermoanaerobacter tengcongensis* (Firmicutes), *Escherichia coli* K-12 (Proteobacteria), *Fusobacterium nucleatum* (Fusobacteria), *Synechocystis* sp. (Cyanobacteria), *Aquifex aeolicus* (Aquificae), *Deinococcus radiodurans* (Deinococcus-Thermus); Archaea, *Archaeoglobus fulgidus, Methanococcus jannaschii, Halobacterium sp., Thermoplasma acidophilum, Pyrococcus horikoshii* (Euryarchaeota), *Aeropyrum pernix, Sulfolobus solfataricus, Pyrobaculum aerophilum* (Crenarchaeota); and Eucarya, *Caenorhabditis elegans* (Animalia), *Saccharomyces cerevisiae, Schizosaccharomyces pombe* (Fungi), *Arabidopsis thaliana* (Plantae).

In order to construct the set of the most conserved set of proteins common to these proteomes, a one-way BLAST search strategy was performed using BLAST algorithm (Altschul, *et al.*, 1997) as summarized in Figure 1. The efficacy of this analysis depends on the ability of each BLAST search to find all homologs of the query sequence. Although this methodology may miss homologs common to all three lineages, it has the advantage of constructing a census of only the most conserved sequences, or of the most conserved domains in proteins. The order in which these genomes were first analyzed by one way BLAST search was as follows: *A. aeolicus, A. fulgidus, A. pernix, A. thaliana, B. subtilis, C. elegans, D. radiodurans, E. coli, F. nucleatum, Halobacterium sp., M. jannaschii, P. aerophilum, P. horikoshii, S. cerevisiae, S. pneumoniae, S. pombe, S. solfataricus, Synechocystis sp., T. acidophilum, T. Tengcongensis*. A second one way BLAST search was then performed in opposite direction. Only one false negative sequence was identified, b1740, that

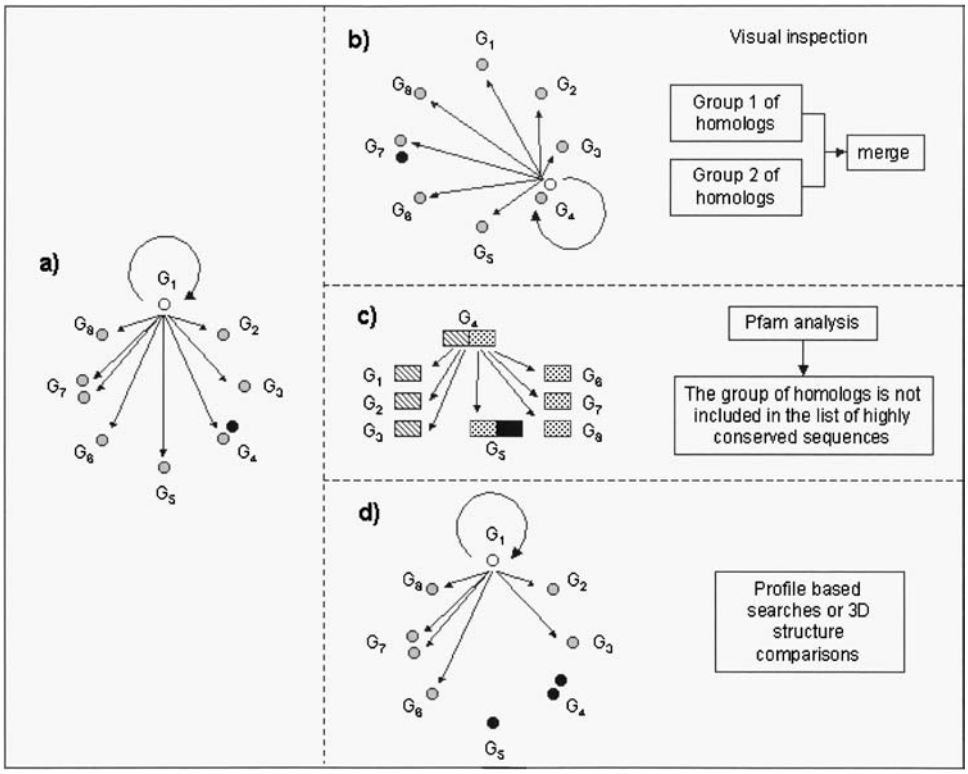

Figure 1. The one way BLAST search, (a) a given sequence (white circle) is considered as highly conserved if it has at least one homolog (gray circles) in all genomes (G_n) in our dataset. Divergent homologous sequences (black circles) may be mismatched or remain unidentified. Sequences detected as highly conserved (HC) by the BLAST search (e-value <0.0001), are represented by gray circles, are not used as query sequences in additional searches once they have been identified as HC; (b) in some cases, a previously undetected homolog will be matched again with the same family of homologous sequences. To avoid this kind of redundancy, visual inspection was performed to merge all the corresponding sequences in one single family; (c) due to protein domain fusions, in some cases false HC protein families will be constructed. A Pfam protein domain analysis was performed in *E. coli, M. jannaschii* and *S. cerevisiae* to identify them and eliminate the false HC family from the final dataset; and (d) simple BLAST searches are likely to miss some homologous sequences. In order to make a more comprehensive search, it would be necessary to use more sensitive methods like profile-based algorithms or three-dimensional structure comparisons, which are not feasible for the time being. As described in the text, to construct the database reported here, steps (a), (b) and (c) were applied.

corresponds to a NAD synthetase (glutamine-hydrolysing). Since domains are the structural and evolutionary units of proteins, we have extracted the highly conserved sequences from *E. coli, M. jannaschii* and *S. cerevisiae*, which are among the most well-studied organisms, and identified the protein domains conserved between each group of homologous sequences using the Pfam database (Bateman *et al.*, 2004) (http://www.sanger.ac.uk/Software/Pfam/). This allowed the identification

of multidomain proteins while avoiding the problem of false positives. To avoid the difficulties with sequences that can be classified under multiple categories, such as enolase (i.e., sugar metabolism or as a component of the degradosome), Table I follows, only in part, the cellular functional classes described by KEGG, with emphasis on a broad-scale description of sequences that interact with RNA.

3. Results

Because our interest is centered on the construction of a census of the most conserved proteins, only the genomes of *E. coli, S. cerevisiae* and *M. jannaschii* were analyzed in detail. The sequences in the resulting set have been classified according to functional categories which we have modified from those used in KEGG to avoid diluting sequences involved in RNA metabolism among other categories (Table I). There are 283 highly conserved proteins from *S. cerevisiae*, 245 from *E. coli*, and 145 from *M. jannaschii*. This set represents the most conserved sequences common to all genomes studied here. As shown in Table I, the list of highly conserved molecular traits includes sequences related to informational process like transcription and translation and several kinds of metabolic enzymes.

As expected from previous studies, different groups of genes involved in different RNAs (Klenk *et al.*, 1993) and translation (Olsen and Woese, 1997; Koonin, 2003; Harris *et al.*, 2003) exhibit a high level of conservation, while the only replication-related conserved ORFs are the bacterial clamp-loading protein complex (Edgell and Doolittle, 1997) (*dnaX*, b0470), and its archaeal/eukaryotic homologs (replication factor-C, MJ1422, MJ0884, YJR068W, YNL290W, YOL094C), which belongs to the AAA (ATPases Associated with diverse cellular Activities) family (Table I).

As shown in Table I, the set of sequences compiled using the methodology outlined here is overwhelmingly dominated by (a) molecules related to translation, RNA synthesis (i.e, transcription), translation and degradation; and (b) proteins with ATP-binding and hydrolyzing activities which can be grouped in relatively few discrete sets (ABC transporter subunits, RNA DEAD helicases, AAA-type ATPases, and HIT superfamily of nucleotide-binding proteins). Given the ubiquity and diversity of these different proteins with ATPase activity, it is perhaps not surprising that the sequences encoding the hydrophilic subunits of F-type ATP synthases and their homologs in the three domains are also highly conserved.

As reported by others (Mushegian and Koonin, 1996; Koonin, 2003; Harris *et al.*, 2003) the resulting repertoire includes few isolated sequences from incompletely represented basic biological processes, such as energy metabolism, nucleotide and amino acid biosynthetic pathways, transcription, translation, and folding of proteins, as well as some sequences related to replication, repair, and cellular transport. As discussed below, this pattern of primary sequence conservation can be explained by manifold processes that include polyphyletic losses, the metabolic idiosyncrasies of diverse species, and different rates of molecular evolution.

TABLE I

List of highly conserved genes in *E. coli, M. jannaschii* and *S. cerevisiae* genomes. Sequences involved in nucleotide-, sugar-, or nucleic acid metabolism are located in the shaded part of the table. Conserved Pfam domains in the three genomes are shown. Sequences with the same domain organization and function are underlined. Sequences are grouped according to single functional categories

Description	Pfam domain	*E. coli*	*M. jannaschii*	*S. cerevisiae*
Transcription				
RNA polymerase β	RNA_pol_Rpb2_1; RNA_pol_Rpb2_2; RNA_pol_Rpb2_3; RNA_pol_Rpb2_6; RNA_pol_Rpb2_7	b3987	MJ1040, MJ1041	YOR151C, YOR207C, YPR010C
RNA polymerase β'	RNA_pol_Rpb1_1; RNA_pol_Rpb1_2; RNA_pol_Rpb1_3; RNA_pol_Rpb1_4; RNA_pol_Rpb1_5	b3988	MJ1042, MJ1043	YDL140C, YOR116C, YOR341W
Translation				
aminoacyl-tRNA synthetases classI	tRNA-synt_1c; tRNA-synt_1c_C; Arg_tRNA_synt_N; tRNA-synt_1d; tRNA-synt_1d_C; tRNA-synt_1	b2400, b0144, b1876, b3384, b2114, b0642, b0026, b4258, b0526	MJ1377, MJ0237, MJ1415, MJ1263, MJ0947, MJ1007, MJ0633	YOL033W, YGL245W, YOR168W, YDR341C, YHR091C, YDR268W, YOL097C, YGR264C, YGR171C, YPL040C, YBL076C, YGR094W, YLR382C, YPL160W, YNL247W
aminoacyl-tRNA synthetases classII	tRNA-synt_2d; tRNA_anti; tRNA-synt_2; tRNA-synt_2b; HGTP_anticodon; tRNA-synt_2c; DHHA1	b1714, b1713, b0930, b1866, b4129, b0893, b0194, b2697, b2890, b4155, b1719, b2514	MJ0487, MJ1108, MJ1555, MJ1077, MJ0564, MJ1238, MJ0228, MJ1197, MJ1000	YFL022C, YPR047W, YLR060W, YHR019C, YLL018C, YCR024C, YPL104W, YDR023W, YHR011W, YKL194C, YOR335C, YNL040W, YDR037W, YNL073W, YER087W, YIL076W, YHR020W, YPR033C
tRNA pseudouridine 55 synthase	TruB_N	b3166	MJ0148	YNL292W, YLR175W
dimethyladenosine transferase	RrnaAD	b0051	MJ1029	YPL266W
elongation factors (EG-G, EF-Tu, and other GTP binding proteins)	GTP_EFTU; GTP_EFTU_D2;	b3340, b3339, b3980, b4375, b2569, b3871, b3168, b3590, b2751	MJ1048, MJ0324, MJ0495, MJ0262, MJ1261, MJ0325	YDR385W, YOR133W, YBR118W, YPR080W, YLR069C, YJL102W, YOR187W, YNL163C, YKL173W, YDR172W, YKR084C, YLR289W, YAL035W, YOL023W, YER025W,
methionine aminopeptidase	Peptidase_M24	b0168, b2385, b3847, b2908	MJ1329, MJ0806	YLR244C, YBL091C, YER078C, YFR006W
ribosomal-protein-alanine acetyltransferase	Acetyltransf_1	b4373, b2434**, b1448, b4012	MJ1530, MJ1207	YHR013C
Sua5, RNA-binding	Sua5_yciO_yrdC	b3282	MJ0062	YGL169W
several different RNA-binding proteins contanining the S1 domain	S1	b3164, b0911	MJ0117	YJR007W, YMR229C
ribosomal proteins (small subunit) *	Ribosomal_S5; Ribosomal_S5_C; Ribosomal_S2; KH_2; Ribosomal_S3_C; S4; Ribosomal_S7; Ribosomal_S8; Ribosomal_S9; Ribosomal_S10; Ribosomal_S11; Ribosomal_S12; Ribosomal_S13; Ribosomal_S19	b3303, b0169, b3314, b3296, b3341, b3306, b3230, b3321, b3297, b3342, b3298, b3316	MJ0475, MJ0982, MJ0461, MJ0190, MJ1047, MJ0470, MJ0195, MJ0322, MJ0191, MJ1046, MJ0189, MJ0180	YGL123W, YBR251W, YLR048W, YGR214W, YHL004W, YNL178W, YPL081W, YBR189W, YNL137C, YHR148W, YJR123W, YJL190C, YLR367W, YDL083C, YMR143W, YBR146W, YHL015W, YJL191W, YCR031C, YNR036C, YGR118W, YPR132W, YDR450W, YML026C, YNL081C, YOL040C, YNR037C
ribosomal proteins (large subunit) *	Ribosomal_L1; Ribosomal_L2; Ribosomal_L2_C; Ribosomal_L6; Ribosomal_L11_N; Ribosomal_L11; Ribosomal_L5; Ribosomal_L5_C; Ribosomal_L14	b3984, b3317, b3305, b3983, b3308, b3310	MJ0510, MJ0179, MJ0176, MJ0471, MJ0469, MJ0466	YGL135W, YPL220W, YEL050C, YFR031C-A, YIL018W, YGR220C, YNL067W, YGL147C, YDR237W, YGR085C, YPR102C, YKL170W, YBL087C, YER117W

(*Continued on next page*)

TABLE I

(*Continued*)

Description	Pfam domain	*E. coli*	*M. jannaschii*	*S. cerevisiae*
Metabolism				
thymidylate kinase [EC:2.7.4.9]	Thymidylate_kin	b1098	MJ0293	YJR057W
dihydroorotate oxidase [EC:1.3.3.1]	DHO_dh	b0945, b2147**	MJ0654	YKL216W
orotate phosphoribosyltransferase [EC:2.4.2.10]	Pribosyltran	b3642	MJ1109, MJ1655	YML106W, YMR271C
aspartate [EC:2.1.3.2] and ornithine [EC:2.1.3.3] carbamoyl- transferase catalytic chain	OTCace_N; OTCace	b4245, b4254, b0273, b2870**	MJ1581, MJ0881	YJL130C, YJL088W
carbamoyl-phosphate synthase small chain [EC:6.3.5.5]	CPSase_sm_chain; GATase	b0032, b3360, b2507, b1263	MJ1019, MJ0238, MJ1575, MJ1131	YJL130C, YOR303W, YKL211C, YMR217W
ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	Pribosyltran	b1207	MJ1366	YER099C, YBL068W, YHL011C, YKL181W, YOL061W
IMP dehydrogenase [EC:1.1.1.205] and hypothetical proteins	IMPDH (1rst. half); CBS, CBS; IMPDH (2nd. half)	b2508	MJ1616, MJ0188, MJ1232, MJ0653, MJ0100, MJ1225, MJ0922, MJ0392, MJ0868, MJ1404, MJ0556	YAR073W, YHR216W, YML056C
adenylo succinate lyase [EC:4.3.2.2]	Lyase_1	b1131, b3960, b1611, b4139	MJ0929, MJ0791	YLR359W, YPL262W
amidophosphoribosyltransferase [EC:2.4.2.14]	GATase_2; Pribosyltran	b2312	MJ0204	YMR300C
pyruvate kinase [EC:2.7.1.40]	PK; PK_C	b1676, b1854	MJ0108	YAL038W, YOR347C
thioredoxin reductase and other reductases [EC:1.8.1.9]	Pyr_redox	b0888, b0606, b0116, b3500, b0304, b3962, b3365, b2711, b2763, b2542	MJ1536, MJ0649, MJ0551	YHR106W, YDR353W, YFL018C, YPL091W, YPL017C, YJR137C
glucosamine--fructose-6-phosphate aminotransferase (isomerizing) [EC:2.6.1.16]	GATase_2; SIS; SIS	b3729, b3371**	MJ1420, MJ1116	YKL104C, YMR085W, YMR084W
metal dependent hydrolase superfamily [EC:3.5.-.-]	Amidohydro_1	b2873**	MJ1490	YIR027C
UDP-galactose 4-epimerase [EC:5.1.3.2] and others	Epimerase	b0759, b3619, b2041, b3788	MJ0211, MJ1055	YBR019C
nucleotidyltransferase activity [EC: 2.7.7.-]	NTP_transferase; Hexapep	b2039, b3789, b1236, b2042, b3730, b3430	MJ1101, MJ1334	YDL055C, YDR211W
hypothetical nucleoside-triphosphatase [EC:3.6.1.15]	Ham1p_like	b2954**	MJ0226	YJR069C**
enolase [EC:4.2.1.11]	Enolase_N; Enolase_C	b2779	MJ0232, MJ0198**	YGR254W, YHR174W, YMR323W, YOR393W, YPL281C
phosphoglycerate kinase [EC:2.7.2.3]	PGK	b2926	MJ0641	YCR012W
phosphomannomutase [EC:5.4.2.8]	PGM_PMM_I; PGM_PMM_II; PGM_PMM_III; PGM_PMM_IV	b2048, b0688, b3176	MJ1100, MJ0399	YMR278W, YMR105C, YKL127W
sugar transferases	Glycos_transf_1	b2044, b3631	MJ1607, MJ1178, MJ1059, MJ1069	YPL175W
sugar transferases	Glycos_transf_2	b2254, b2351, b0363, b1022, b3615	MJ1222, MJ0544	YPL227C, YPR183W
nucleotide-binding proteins	HIT	b1103**	MJ0866**	YDL125C, YDR305C
phosphoglycerate dehydrogenase [EC:1.1.1.95]	2-Hacid_dh; 2-Hacid_dh_C; ACT	b2913, b1380, b3553, b2320, b1033	MJ1018	YER081W, YIL074C, YOR388C, YNL274C, YGL185C, YPL113C
NAD synthetase [EC:6.3.1.5, 6.3.5.1]	NAD_synthase	b1740	MJ1352	YHR074W
flavoprotein enzymes	Flavoprotein	b3639	MJ0913	YKL088W, YKR072C, YOR054C
UPP synthetase [EC:2.5.1.-]	Prenyltransf	b0174	MJ1372	YMR101C, YBR002C
tryptophan synthase (β-chain) [EC:4.2.1.20]	PALP	b1261, b3117, b2421, b3772, b2871, b2414	MJ1037, MJ1465	YGL026C, YCL064C, YKL218C, YGR155W, YER086W, YGR012W
tryptophan synthase (α-chain) [EC:4.2.1.20]	Trp_syntA	b1260	MJ1038	YGL026C
histidinol-phosphate aminotransferase [EC:2.6.1.9]	Aminotran_1_2	b2021, b2379, b0600, b2290, b1439, b1622, b4340	MJ0955, MJ0001, MJ1391, MJ0684, MJ1479	YIL116W, YJL060W, YDR111C, YLR089C
Polyprenyl synthetase [EC: 2.5.1-]	polyprenyl_synt	b0421, b3187	MJ0860	YJL167W, YPL069C, YBR003W
probable glyoxylase II [EC:3.1.2.6]	Lactamase_B	b0927**, b0212	MJ0888**	YDR272W
probable peroxiredoxin [EC:1.6.4.-]	AhpC-TSA	b0605, b2480	MJ0736	YIL010W, YBL064C, YML028W, YDR453C

(*Continued on next page*)

TABLE I
(*Continued*)

Description	Pfam domain	*E. coli*	*M. jannaschii*	*S. cerevisiae*
DEAD helicases	DEAD; Helicase_C	b0797, b3162, b1343, b3780, b2576, b3822, b1653	MJ0669, MJ1401, MJ1574, MJ0294, MJ0383, MJ1124	YJL138C, YKR059W, YDR021W, YPL119C, YOR204W, YGL078C, YNL112W, YDL160C, YLL008W, YHR065C, YDL084W, YHR169W, YJL033W, YDR243C, YOR046C, YBR237W, YMR290C, YGL171W, YFL002C, YDL031W, YDR194C, YLR276C, YBR142W, YKR024C, YGL064C, YNR038W, YDR291W, YGL251C, YER172C, YMR190C, YGR271W
ATP synthesis (*atpA*, *atpB*)	ATP-synt_ab_N; ATP-synt_ab; ATP-synt_ab_C	b3734, b3732, b1941	MJ0217, MJ0216	YBL099W, YJR121W, YDL185W, YBR127C
Replication, recombination and repair factors				
ATPase family proteins (clamp-loading, γ τ subunits)	AAA	b0470, b3178, b0892,	MJ1422, MJ0884, MJ1156, MJ1176, MJ1494	YJR068W, YNL290W, YOL094C, YMR089C, YER017C, YDL126C, YBR080C, YPR024W, YGR270W, YPR173C, YKL145W, YGL048C, YOR259C, YDL007W, YOR117W, YDR394W, YLR397C, YNL329C, YLL034C, YKL197C, YGR028W, YPL074W, YER047C, YDR375C, YBR186W
ribonuclease HII	RNase_HII	b0183	MJ0135	YNL072W
endonuclease III	HhH-GPD (1rst. half); HHH; HhH-GPD (2nd. half);	b1633, b2961	MJ1434, MJ0613	YAL015C, YOL043C
DNA topoisomerase I and III	Toprim; Topoisom_bac	b1274, b1763	MJ1652, MJ1512	YLR234W
ABC transporters	ABC_tran	b0448, b1290, b1291, b1496, b1682, b1709, b1756, b2201, b3479, b4058, b4096, b0066, b0127, b0151, b0199, b0262, b0366, b0449, b0490, b0495, b0588, b0652, b0760, b0794, b0809, b0820, b0829, b0855, b0864, b0879, b0886, b0887, b0914, b0933, b0949, b1117, b1126, b1246, b1247, b1318, b1441, b1483, b1484, b1513, b1858, b1900, b1917, b2129, b2149, b2180, b2306, b2422, b2547, b2677, b3201, b3271, b3352, b3450, b3454, b3455, b3463, b3480, b3486, b3540, b3541, b3567, b3725, b3749, b4035, b4087, b4097, b4106, b4228, b4287, b4391	MJ1023, MJ1088, MJ1242, MJ1267, MJ1367, MJ1508, MJ1572, MJ1662	YKR104W, YLL015W, YNR070W, YOR011W, YOR328W, YPL058C, YPL147W, YCR011C, YDR091C, YFR009W, YGR281W, YHL035C, YKL209C, YLL048C, YLR188W, YLR249W, YMR301C, YNL014W, YOL075C, YOR153W, YPL226W, YPL270W
Protein management				
signal recognition particle protein	SRP54_N; SRP54; SRP_SPB	b2610, b3464	MJ0101, MJ0291	YPR088C, YDR292C
chaperonin Cpn60	Cpn60_TCP1	b4143	MJ0999	YLR259C, YJR064W, YJL111W, YDL143W, YIL142W, YDR188W, YJL014W, YDR212W, YJL008C

*Different groups of homologous are joined in a single functional category.
**Hypothetical ORF.

4. Discussion

The methodological approach developed here is straightforward: we searched for sequences present in all genomes analyzed that have changed slowly enough to be still recognizable using the one-way BLAST strategy described above. Because a simple BLAST search may not find all possible homologs of a given sequence, the results shown here represent a first approximation of the gene complement of the LCA as described from the standpoint of *S. cerevisiae, E. coli* and *M. jannaschii*, three species that have been selected because of the considerable information that exists on their biology. A more complete census of highly conserved traits would require a detailed analysis of each set of homologous proteins using more sensitive approaches like profile-based methods and, eventually, information derived from tertiary structure databases.

Reconstructions of gene complements of distant ancestors are mere statistical approximations of biological past, since their accuracy depends on manifold factors including the possible biases in the construction of genome databases, the levels of horizontal gene transfer, the significant variations in substitution rates of different proteins, and the degree of secondary looses, as well as methodological caveats. As argued here, in spite of these limitations the available data provides significant insights into (a) the existence of an ancient RNA/protein world; (b) the biological complexity of the LCA; and (c) and evidence pertaining to the chemical nature of the cenancestral genome (i.e, RNA or DNA).

4.1. The high proportion of cenancestral RNA-related ORFs suggest the prior-existence of an RNA/protein world

In order to avoid the bias introduced by secondary gene losses (Becerra *et al.*, 1997), we have not included in our analysis genomes from obligate parasites or endosymbiotic organisms, which would lead to an underestimation of the number of genes inherited from the LCA. As demonstrated by other analyses, proteins that interact with RNA in one way or other are among the most highly conserved sequences (Delaye and Lazcano, 2000; Anantharaman *et al.*, 2002). This is shown in Figure 2, where more than 80% of the conserved domains correspond to proteins that interact directly with RNA (such as ribosomal proteins, DEAD-type helicases, aminoacyl tRNA synthetases, and elongation factors, among others), or take part in RNA and nucleotide biosyntheses, including the DNA-dependent RNA polymerase β and β' subunits, dimethyladenosine transferase, adenyl-succinate lyases, dihydroorotate oxidase, and ribose-phosphate pyrophosphokinase, among many others (Table I). This percentage includes sugar metabolism-related sequences (see below).

Nonetheless, few metabolic genes are part of the conserved ORF product set. These include many sugar metabolism-related sequences, such as the enolase-encoding genes noted above, as well as homologs of thioredoxin (*trxB*, mj1536), phosphoribosyl-pyrophosphate synthase (*prs*, b1207), and UDP-galactose

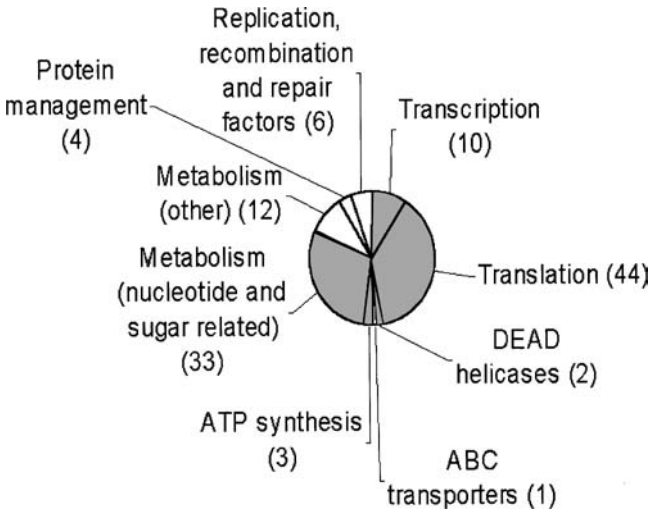

Figure 2. Prevalence of highly conserved sequences and protein domains related to RNA metabolism in the dataset estimated here.

4-epimerase (*galE*, b0759) genes. Very likely, the evolutionary conservation of the *trxB* and *prsA* genes is best understood in terms of the key roles they play in nucleotide biosynthesis. The role of UDP-galactose 4-epimerase in complex carbohydrate synthesis via the interconversion of the galactosyl and glucosyl groups is well-known. Although the uniqueness of the enzyme mechanism has been acknowledged, it is possible that the conservation of UDP-galactose 4-epimerase is due to an undescribed participation in other basic processes, as in the case of enolase.

ATP-dependent RNA helicases are universally-distributed, highly conserved proteins which participate in a variety of cellular functions involving the unwinding and rearrangement of RNA molecules, including translation initiation, RNA splicing, ribosome assembly, mRNA nucleocytoplasmic transport, and degradosome-mediated mRNA decay (Schmid and Linder, 1992). The degradosome is a multienzymatic complex involved in mRNA processing and breakdown, that includes polynucleotide phosphorylase (which shares an RNA-binding domain with RNAse E), polyphosphate kinase (PPK), ATP-dependent DEAD/H-type RNA helicase (RhlB), and enolase, a glycolytic enzyme that catalyzes the conversion of 2-phosphoglycerate to phosphoenolpyruvate and water (Blum *et al.*, 1997). Reports showing that PPK is not essential for *E. coli* survival and may be a later evolutionary addition involved in degradosome regulation (Blum *et al.*, 1997) are consistent with the absence of the corresponding gene from the set of highly similar ORFs.

Although RNA hydrolysis is an exergonic process, degradosome-mediated mRNA turnover plays a key role as a regulatory mechanism for gene expression in both prokaryotes and eukaryotes (Blum *et al.*, 1997). A possible explanation for the conservation of DEAD-type RNA helicases may lie in their role in protein

biosynthesis and in mRNA degradation. This possibility is supported by the phylogenetic relatedness of the RhlB and DeaD sequences (Schmid and Linder, 1992) and by the surprising conservation of the *eno*-like sequences. If this interpretation is correct, then it could be argued that degradosome-mediated mRNA turnover is an ancient control mechanism at RNA level that was established prior to the divergence of the three primary kingdoms. Together with other lines of evidence, including the observation that the most highly conserved gene clusters in several (eu)bacterial genomes are regulated at RNA-level (Siefert *et al.*, 1997), the results reported here are fully consistent with the hypothesis that during early stages of cell evolution RNA molecules played a more conspicuous role in cellular processes.

4.2. THE LCA, A PROGENOTE OR A BACTERIAL-LIKE CENANCESTOR?

Our ability to reconstruct the gene complement of the LCA depends in part on the levels of lateral gene transfer during early cell evolution, as well as on the degree of differential gene losses across cellular lineages (Becerra *et al.*, 1997). If lateral gene transfer was rampant during early evolution, then there is a risk of overestimating the number of genes of the LCA. The opposite outcome can be expected if, on the other hand, differential losses have been much more common in evolution.

Analysis of an increasingly large number of genes and genomes has revealed major discrepancies with the topology of rRNA trees. As summarized by Brown (2003), very often these differences have been interpreted as evidence of LGT events between different species, questioning the feasibility of the reconstruction and proper understanding of early biological history (Doolittle, 1999). There is evidence that genomes have a mosaic-like nature whose components come from a wide variety of sources (Ochman *et al.*, 2000). However, not all sequences are equally prone to such phenomenon, nor has the evolution of all prokaryotic species been equally affected by LGT (Zhaxybayeva *et al.*, 2004). Most transfers take place between closely related species, and interdomain LGT events are quite rare. Accordingly, if the species that carries the ancestral sequence was not the organismal ancestor itself, then very likely was its close relative (Zhaxybayeva and Gogarten, 2004).

Driven in part by the impact of lateral gene acquisition as revealed by the discrepancies of different gene phylogenies with the rRNA tree, Woese (1998) proposed that the LCA was not a single organism, but rather a highly diverse population of metabolically complementary, cellular progenotes endowed with multiple, small linear chromosome-like genomes that benefited from massive multidirectional horizontal transfer events. According to this idea, which is reminiscent of a similar hypothesis proposed independently by Kandler (1994), the essential features of translation and the development of metabolic pathways took place before the earliest branching event, but what led to the three domains was not a single ancestral lineage, but a rapidly differentiating community of genetic entities. This communal ancestor occupied as a whole the node located at the bottom of the universal

tree, in which the decrease of sequence exchange and increasing genetic isolation eventually lead to the observed tripartite division of the biosphere.

We suggest a different scenario. The LCA was of course not alone: company must have been kept by its siblings, a population of entities similar to it that existed throughout the same period. They may have not survived, but some of their genes did if they became integrated via lateral transfer into the LCA genome. The cenancestor is one of the last evolutionary outcomes of a tree trunk of unknown length, during which the history of a long but not necessarily slow (Lazcano and Miller, 1994) series of ancestral events including lateral gene transfer, gene losses, and duplications probably played a significant role in the accretion of complex genomes (Lazcano *et al.*, 1992; Castresana, 2001; Snel *et al.*, 2002).

The gene complement of the LCA can also be considered a mosaic in that not all universally distributed genes are of equal antiquity. For instance, the evidence suggesting that DNA evolved after RNA and proteins (Lazcano *et al.*, 1988; Freeland *et al.*, 1999) implies that the translational machinery is older than ribonucleotide reductases or DNA polymerases. However, the extraordinary similarity of the basic traits shared by all extant cells suggest that they must have been integrated by the time of the LCA.

The genetic entities that formed the communal ancestor proposed by Woese (1998) may have been extremely diverse, but an indication of their ultimate monophyletic origin from a sole progenitor is provided by universally distributed features such as the genetic code and the basic features of the gene expression machinery. Did this hypothetical communal progenote ancestor diverged sharply into the three domains soon after the appearance of the code and the establishment of translation? Not necessarily. The origin of the mutant sequences ancestral to those found in all extant species, and the divergence of the Bacteria, Archaea, and Eucarya were not synchronous events, i.e., the separation of the primary domains took place later, perhaps even much later, than the appearance of the genetic components of their last common ancestor. Moreover, by definition, the node located at the bottom of the cladogram is the root of a phylogenetic tree, and corresponds to the common ancestor of the group under study. But names may be misleading. What we have been calling the root of the universal tree is in fact the tip of its trunk: inventories of LCA genes include sequences that originated in different pre-cenancestral epochs (Delaye and Lazcano, 2000; Becerra-Bracho *et al.*, 2000; Anantharaman *et al.*, 2002). As noted by Fox et al. (1982), a major phylogenetic issue is the relationship between the timing of the transition from the age of progenotes and the branching order between the three cell domains. The gene complement of the LCA described here corresponds to a cellular entity that evolved after the age of progenote had come to an end, but prior to the divergence of the Archaea, Bacteria and Eucarya.

It is currently difficult to propose a unifying hypothesis. However, the scheme outlined here is supported by gene content trees, which exhibit a broad-level agreement with rRNA-based phylogenies (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999; Tekaia, *et al.*, 1999). Such trees are not cladograms but phenograms, i.e.,

they are merely hierarchical representations of similarities and differences in gene content, where the presence or absence of a sequence is counted as a character. Since different lineages evolve at different rates, such overall similarity may be an equivocal indicator of genealogical relationships. Nevertheless, these trees are consistent with rRNA phylogenies, and do not support the hypothesis of massive LGT between distant species. The robustness exhibited by these different methodologies indicates that although LGT has played an important role in cellular evolution, it has not obliterated the early history of life (Glansdorff, 2000), and that the role of reticulate evolution in defining the LCA as a progenote swarm may have been overstated.

4.3. THE LCA, A DNA OR A RNA GENOME?

Since all extant cells are endowed with DNA genomes, the most parsimonious conclusion is that this genetic polymer was already present in the cenancestral population. Woese (1983, 1987) has suggested otherwise, arguing for a progenote-like universal ancestor endowed with a rapidly evolving genome formed by disaggregated, small-sized RNA molecules. This possibility appeared to be supported by the findings of Mushegian and Koonin (1996), who argued that the absence of eucaryal or archaeal homologs of key components of DNA replication and nucleotide biosynthesis in the minimal gene set which resulted from the comparison of the *Haemophilus influenzae* and *Mycoplasma genitalium* genomes suggested that the cenancestor had used RNA as genetic polymer. Such conclusion is weakened by the limited data set analyzed, which consisted of only two parasitic bacterial genomes that have undergone extensive polyphyletic gene losses (Becerra *et al.*, 1997). In a subsequent publication, however, Koonin and his collaborators analyzed a large set of primases, replicative polymerases, and other proteins involved in DNA replication, and suggested an alternative scheme with a hybrid RNA/DNA cenancestral genetic system whose complex replication cycle involved reverse transcription (Leipe *et al.*, 1999).

The idea that RNA preceded DNA as cellular genetic material has been proposed independently by many authors (see, for instance, Oparin, 1961; Rich, 1962; Haldane, 1965; Reanney, 1979). However, it is likely that double-stranded DNA genomes had become firmly established prior to the divergence of the three primary domains. The major arguments supporting this possibility are:

(a) in sharp contrast with other energetically favorable biochemical reactions (such as phosphodiester backbone hydrolysis or the transfer of amino groups), the direct removal of the oxygen from the 2′-C ribonucleotide pentose ring to form the corresponding deoxy-equivalents is a thermodynamically much less-favored reaction, considerably reducing the likelihood of multiple, independent origins of biological ribonucleotide reduction;
(b) demonstration of the monophyletic origin of ribonucleotide reductases (RNR) is greatly complicated by their highly divergent primary sequences and the

different mechanisms by which they generate the substrate 3'-radical species required for the removal of the 2'-OH group. However, sequence analysis and biochemical characterization of archaebacterial RNRs have shown their similarities with their eubacterial and eukaryotic counterparts, suggesting that the most distributed enzymes are of monophyletic origin (Tauer *et al.*, 1996; Riera *et al.*, 1997; Freeland *et al.*, 1999); and

(c) sequence similarities shared by many ancient, large proteins found in all three domains suggest that considerable fidelity existed in the operative genetic system of their common ancestor, but such fidelity is unlikely to be found in RNA-based genetic systems (Lazcano *et al.*, 1992).

While accepting a DNA component in the LCA genome, Leipe *et al.*, (1999) have underlined the highly divergent character of the main components of the (eu)bacterial replication machinery when compared with their archaeal/eukaryotic counterpart. Although it is possible to recognize the evolutionary relatedness of various orthologous DNA informational proteins (i.e., ATP-dependent clamp loader proteins, topoisomerases, gyrases, and 5'–3' exonucleases) across the entire phylogenetic spectrum, comparative proteome analysis has shown that (eu)bacterial replicative polymerases and primases lack homologues in the two other primary kingdoms. As argued by Leipe *et al.* (1999) these observations can be explained by assuming a dual, independent origin of the DNA replication machineries of the Bacteria, on the one hand, and of the Archaea/Eucaryal on the other.

We think this is unlikely. Nucleic acid replication enzymatic machinery requires, at the very least, a replicase, a primase, and a helicase (Forterre, 1999), which are currently described as non-orthologues between the bacterial and the archaea/eukaryotic branches. Given the central role that is assigned to nucleic acid replication in mainstream definitions of life (Koshland, 2002), the lack of conservation and polyphyly of several of its key enzymatic components is somewhat surprising. However, we believe that there may be an explanation for the evolution of the DNA replication machinery simpler that the one advocated by Leipe *et al.*, (1999). Our hypothesis implies that this progenitor DNA polymerase was originally involved in the replication of the LCA genome, until its (eu)bacterial descendants underwent a non-orthologous displacement by the ancestor of the *Escherichia coli* replicative DNA pol III (DNA pol C) and its homologs. Based on the conservation and versatily of functions of the palm domain of DNA polymerase II and its homologs, we suggest that the Archaeal-Eucaryal replication machinery is in fact older than the current Bacterial one.

5. Conclusions and Outlook

The variations of traits common to extant species can be easily explained as the outcome of divergent processes from an ancestral life form that existed prior to the

separation of the three major biological domains, i.e., the last common ancestor (LCA) or cenancestor. However, if the term "universal distribution" is restricted to its most obvious sense, i.e., that of traits found in all completely sequenced genomes, then quite unexpectedly the resulting repertoire is formed by relatively few features and by incompletely represented biochemical processes (Tatusov *et al.*, 1997; Tekaia *et al.*, 1999; Brown *et al.*, 2001; Delaye *et al.*, 2002). Quite surprisingly, some of the most likely *a priori* candidates for strict universality, such as those sequences involved in DNA replication, have also turned out to be not only poorly preserved but also, in some cases, of polyphyletic origin (Edgell and Doolittle, 1997; Olsen and Woese, 1997; Böhlke *et al.*, 2000).

The traits described here as inherited from the LCA represent only those sequences that can be identified at the primary structure level. Because it is known that tertiary structure level is more conserved across evolutionary distances, attempts to reconstruct the LCA gene complement using a fold-recognition algorithm may enhance the census of such cenancestral molecular traits. For instance, the lack of detection of the complete set of sequences encoding F-type ATPases does not indicate these multimeric enzymes were absent in the LCA (Gogarten and Taiz, 1992; Castresana *et al.*, 1994), but should be interpreted instead as an indication of the different rates of evolution of the sequences and the limits of the methodologies described here. Nonetheless, the fact that are no major inconsistencies between the data presented here and those reported by other authors who have used different methodologies (Delaye and Lazcano, 2000; Anantharaman *et al.*, 2002; Harris *et al.*, 2003; Koonin, 2003; Delaye *et al.*, 2004) underlines the robustness of our own results.

The dataset reported in Table I includes (a) genes that have undergone lateral transfer and (b) sequences that although highly conserved, have originated in different evolutionary epochs. However, the over-representation of highly conserved sequences related to RNA metabolism, i.e., ORFs whose products synthesize, degrade, or interact with polyribonucleotides (Table I), is best understood in terms of an early evolutionary period during which RNA played a more prominent role in biological processes, i.e., an RNA/protein world. Degradasome components (DEAD helicase and enolase) are as highly conserved as molecules involved in RNA biosynthesis. It can also be argued that the conservation of enolase and DEAD-type RNA helicases discussed here constitutes additional evidence of the early development of gene expression control mechanisms at the RNA level. These conclusions, however, do not imply that the cenancestor was endowed with an RNA genome, nor support the possibility of an RNA-based origin of life. Indeed, the conservation of genes involved in ribonucleotide reduction such as *trxB*, combined with other independent lines of evidence, argues for the presence of a DNA genome in the LCA.

The analysis of the dataset reported here is consistent with a prokaryotic root of universal phylogenies, and indicates that the cenancestor was much more complex that expected for a progenote. There is no contradiction between this conclusion

and the relatively few metabolic genes that are conserved. In fact, most of them synthesize or interact with ribonucleotides or sugar compounds. Conserved sequences related to metabolic pathways include homologs of phosphoribosyl pyrophosphate synthase and thioredoxin, among others, which are involved in nucleotide biosynthesis. Although the information contained in the available databases corresponds only to a minor portion of biological diversity, the sequences reported here are likely to be part of an essential and highly conserved pool of proteins common to all organisms.

Acknowledgments

Work reported here was supported by project DGPA IN-111003 (UNAM, Mexico). We are deeply indebted to two anonymous reviewers for their many suggestions and help in improving the results presented here. This paper was completed during a leave of absence in which one of us (A.B.) enjoyed the hospitality of Dr. Janet Siefert (Rice University, Houston) thanks to the support of DGPA-UNAM. Due thanks are given to the Departamento de Supercómputo, DGSCA-UNAM, for its constant support and technical assistance.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, Z., Miller, W. and Lipman, D.J.: 1997, Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs, *Nucleic Acid Res.* **25**, 3389–3402.

Anantharaman, V., Koonin, E.V. and Aravind, L.: 2002, Comparative genomics and evolution of proteins involved in RNA metabolism, *Nucleic Acid Res.* **30**, 1427–1464.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R.: 2004, The Pfam Protein Families Database, *Nucleic Acids Res.* **32**, 138–141.

Becerra, A., Islas, S., Leguina, J.I., Silva, E. and Lazcano, A.: 1997, Polyphyletic Gene Losses Can Bias Backtrack Characterizations of the Cenancestor, *J. Mol. Evol.* **45**, 115–118.

Becerra-Bracho, A., Velasco, A.M., Islas, S., Silva, E., Lloret, S. and Lazcano, A.: 2000, Molecular Biology and the Reconstruction of Microbial Phylogenies: Des Laisions Dangereuses? In J. Chela-Flores, G. Lemerchand, and J. Oró (eds), *Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology*, Klüwer Academic Publishers, Dordrecht, pp. 135–150.

Blum, E., Py, B., Carpousis, A.J. and Higgins, C.F.: 1997, Polyphosphate Kinase is a Component of the *Escherichia coli* RNA degradosome, *Mol. Microbiol.* **26**, 387–398.

Böhlke, K., Pisani, F.M., Vorgias, C.E., Frey, B., Sobek, H., Rossi, M. and Antranikian, G.: 2000, PCR Performance of the B-type DNA Polymerase from the Thermophilic Euryarchaeon *Thermococcus aggregans* Improved by Mutations in the Y-GG/A Motif, *Nucleic Acid Res.* 28: 3910–3917.

Brown, J.R.: 2003, Ancient Horizontal Gene Transfer, *Nature Rev. Genet.* **4**, 121–132.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J.: 2001, Universal Trees Based on Large Combined Protein Sequence Data Sets, *Nature Genet* **28**, 281–285.

Castresana, J.: 2001, Comparative Genomics and Bioenergetics, *Biochem. Biophys. Acta* **1506**, 147–162.

Castresana, J., Lubben, M., Saraste, M. and Higgins, D.G.: 1994, Evolution of Cytochrome Oxidase, and Enzyme Older than Atmospheric Oxygen, *EMBO J.* **13**, 2516–2525.

Delaye, L. and Lazcano, A.: 2000, RNA-Binding Peptides as Molecular Fossils in J. Chela-Flores, G. Lemerchand, and J. Oró (eds), *Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology*, Kluwer Academic Publishers, Dordrecht, pp. 285–288.

Delaye, L., Becerra, A. and Lazcano, A.: 2002, The Nature of the Last Common Ancestor in Lluis Ribas de Pouplana (ed), *The Genetic Code and the Origin of Life*, Landes Bioscience, Georgetown, in press.

Doolittle, W.F.: 1999, Phylogenetic Classification and the Universal Tree, *Science* 284: 2124–2128.

Doolittle, W.F.: 2000, The Nature of the Universal Ancestor and the Evolution of the Proteome, *Curr. Opinion Struct. Biol.* **10**, 355–358.

Edgell, D.R. and Doolittle, W.F.: 1997, Archaea and the Origins, of DNA Replication Proteins, *Cell* **89**, 995–998.

Fitch, W.M. and Upper, K.: 1987, The Phylogeny of tRNA Sequences Provides Evidence of Ambiguity Reduction in the Origin of the Genetic Code, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 759–767.

Fitz-Gibbon, S.T. and House, C.H.: 1999, Whole Genome-Based Phylogenetic Analysis of Free-Living Organisms, *Nucleic Acids Res.* **27**, 4218–4222.

Forterre, P.: 1999, Displacement of Cellular Proteins by Functional Analogues from Plasmids or Viruses Could Explain Puzzling Phylogenies of Many DNA Informational Proteins, *Mol. Microbiol.* **33**, 457–465.

Fox, G.E., Luehrsen, K.R. and Woese, C.R.: 1982, Archaebacterial 5S Ribosomal RNA, *Zbl. Bakt. Hyg. I Abt. Orig.* **C3**, 330–345.

Forterre, P.: 2002, The Origin of DNA Genomes and DNA Replication Proteins, *Curr. Opin. Microbiol.* **5**, 525–532.

Freeland, S.J., Knight, R.D. and Landweber, L.F.: 1999, Do Proteins Predate DNA? *Science* **286**, 690–692.

Glansdorff, N.: 2000, About the Last Common Ancestor, the Universal Life-Tree and Lateral Gene Transfer: A Reappraisal, *Mol. Microbiol.* **38**, 177–185.

Gogarten, J.P. and Taiz, L.: 1992, Evolution of Proton-Pumping ATPase: Rooting the Tree of Life, *Photosyn. Res.* 33: 137–146.

Haldane, J.B.S.: 1965, Data Needed for the Blueprint of the First Organism. *In Fox*, S. W. (ed), *The Origin of Prebiological Systems and their Molecular Matrices*, Academic Press, New York, pp. 11–15.

Harris, J.K., Kelley, S.T., Spiegelman, G.B. and Pace, N.R.: 2003, The Genetic Core of the Universal Ancestor, *Genome Res.* 13: 407–412.

Kandler, O.: 1994, The early diversification of life. In Stefan Bengtson: (ed), *Early Life on Earth: Nobel Symposium No. 84*, Columbia University Press/Nobel Foundation, New York, pp. 152–160.

Kanehisa, M. and Goto, S.: 2000, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res* 28: 27–30.

Klenk, H.-P., Palm, P., Zillig, W.: 1993, DNA-Dependent RNA Polymerases as Phylogenetic Markers Molecules, *Syst. Appl. Microbiol.* **16**, 138–147.

Koonin, E.V.: 2003, Comparative Genomics, Minimal Gene-Sets and the Last Universal Common Ancestor, *Nature Reviews* **1**, 127–136.

Koshland, D.E.: 2002, The Seven Pillars of Life, *Science* **295**, 2215–2216.

Lazcano, A., Guerrero, R., Margulis, L. and Oró, J.: 1988, The Evolutionary Transition from RNA to DNA in Early Cells, *J. Mol. Evol.* **27**, 283–290.

Lazcano, A.: 1995, Cellular Evolution During the Early Archean: What Happened Between the Progenote and the Cenancestor? *Microbiologia SEM* **11**, 185–198.

Lazcano, A., Fox, G.E. and Oró, J.: 1992, Life before DNA: The Origin and Early Evolution of Early Archean Cells. In R.P. Mortlock: (ed), *The Evolution of Metabolic Function*, CRC Press, Boca Raton, FL, pp. 237–295.

Lazcano, A. and Miller, S.L.: 1994, How Long did it Take for Life to Begin and Evolve to Cyanobacteria? *J. Mol. Evol.* **39**, 546–554.

Leipe, D.D., Aravind, L. and Koonin, E.V.: 1999, Did DNA Replication Evolve Twice Independently? *Nucleic Acid Res.* **27**, 3389–3401.

Line, M.A.: 2002, The Enigma of the Origin of Life and its Timing, *Microbiology* **148**, 21–27.

Mushegian, A.R. and Koonin, E.V.: 1996, A Minimal Gene Set for Cellular Life Derived by Comparison of Complete Bacterial Genomes, *Proc. Natl. Acad. Sci. USA* **93**, 10268–10273.

Ochman, H., Lawrence, J.G. and Groisman, E.A.: 2000, Lateral Gene Transfer and the Nature of Bacterial Innovation, *Nature* **405**, 299–304.

Olsen, G.J. and Woese, C.R.: 1997, Archael Genomics: An Overview, *Cell* **89**, 991–994.

Oparin, A.I.: 1961, *Life: Its nature, origin and development:* Oliver and Boyd, Edinburgh,

Philippe, H. and Forterre, P.: 1999, The Rooting of the Universal Tree of Life is not Reliable, *J. Mol. Evol.* **49**, 509–523.

Reanney, D.C.: 1979, RNA Splicing and Polynucleotide Evolution, *Nature* **227**, 597–600.

Rich, A.: 1962, On the Problems of Evolution and Biochemical Information Transfer. In Kasha, M. and Pullman, B.: (eds), *Horizons in Biochemistry*, Academic Press, New York, pp. 103–126.

Riera, J., Robb, F.T., Weiss, R. and Fontecave, M.: 1997, Ribonucleotide Reductase in the Archaeon Pyrococcus Furiosus: A Critical Enzyme in the Evolution of DNA Genomes, *Proc. Natl. Acad. Sci. USA* **94**, 475–478.

Rivera, M.C. and Lake, J.A.: 2004, The Ring of Life Provides Evidence for a Genome Fusion Origin of Eukaryotes, *Nature* **431**: 152–155.

Schmid, S.R. and Linder, P.: 1992, D-E-A-D Protein Family of Putative RNA Helicases, *Mol. Microbiol.* **6**, 283–292.

Seifert, J.L., Martin, K.A., Abdi, F., Wagner, W.R. and Fox, G.E.: 1997, Conserved Gene Clusters in Bacterial Genomes Provide Further Support for the Primacy of RNA. *J. Mol. Evol.* **45**, 467–472.

Snel, B., Bork, P. and Huynen, M.A.: 1999, Genome Phylogeny Based on Gene Content, *Nature Genet.* **21**: 108–110.

Snel, B., Bork, P. and Huynen, M.A.: 2002, Genomes in Flux: The Evolution of Archaeal and Proteobacterial Gene Content, *Genome Res.* **12**: 17–25.

Tauer, A. and Benner, S.A.: 1996, The B_{12}-Dependent Ribonucleotide Reductase from the Archaebacterium *Thermoplasma acidophila:* An Evolutionary Solution to the Ribonucleotide Reductase Conundrum, *Proc. Natl. Acad. Sci. USA* **94**: 53–58.

Tatusov, R.L., Koonin, E.V. and Lipman, D.J.: 1997, A Genomic Perspective on Protein Families, *Science* **278**: 631–637.

Tekaia, F., Lazcano, A. and Dujon, B.: 1999, The Genomic Tree as Revealed from Whole Proteome Comparisons, *Genome Res.* **9**, 550–557.

Woese, C.R. and Fox, G.E.: 1977, The Concept of Cellular Evolution, *J. Mol. Evol.* **10**, 1–6.

Woese, C.R.: 1983, The Primary Lines of Descent and the Universal Ancestor. In D. S. Bendall (ed), *Evolution from Molecules to Men*, Cambridge University Press, Cambridge, pp. 209–233.

Woese, C.R.: 1987, Bacterial Evolution. Microbiol, *Reviews* **51**, 221–271.

Woese, C.R.: 1998, The Universal Ancestor, *Proc. Natl. Acad. Sci. USA* **95**, 6854–6859.

Woese, C.R., Kandler, O. and Wheelis M.L.: 1990, Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya, *Proc. Natl. Acad. Sci. U S A.* **87**, 4576–4579.

Zhaxybayeva O. and Gogarten, J.P.: 2004, Cladogenesis, Coalescence and the Evolution of the Three Domains of Life, *Trends Genet.* **20**, 182–187.

Zhaxybayeva O., Lapierre, P. and Gogarten, J. P.: 2004, Genome Mosaicism and Organismical Lineages, *Trends Genet.* **20**, 254–260.