[Frontiers in Bioscience 4657-4666, May 1, 2008]

Gene content of LUCA, the last universal common ancestor

Arcady Mushegian

Stowers Institute for Medical Research, 1000 E 50th St., Kansas City, Missouri 64110, and Department of Microbiology, Molecular Genetics and Immunology, Kansas University Medical Center, Kansas City, Kansas, 66160, USA

TABLE OF CONTENTS

1. ABSTRACT

Comparative genomics and modern phylogenetic approaches allow us to infer the gene content of LUCA, the Last Universal Common Ancestor of all known currently living cellular organisms. Most of the estimates produce a putative LUCA with 500-1000 protein-coding genes and biochemically coherent metabolism, if the average rates of gene gains (gene emergence plus horizontal gene transfer) and gene losses per family are allowed to be close to each other. This estimate is not strongly sensitive to the topology of the Tree of Life, but the identity of the genes that are placed in LUCA may depend on the position of the deep branches and the root of the tree.

2. INTRODUCTION

The inference of the Last Universal Common Ancestor (LUCA) of all cellular species that inhabit Earth is a multifaceted problem, which can be approached in two diametrically opposite ways. First, one may wish to proceed "forward in time" and examine what is known about physical and chemical conditions of the prebiotic Earth, asking questions about the genetic systems that could emerge under these conditions, and about the progression from those primitive genomes to LUCA. In a complementary, "backward in time" approach, the information about currently living organisms is used to reconstruct the traits of LUCA. In this essay, I examine the latter class of approaches, which take us, as required, directly to the last ancestor, as opposed to perhaps some ancestor of such an ancestor.

The research program outlined by Pauling and Zuckerkandl (1) is to compare genes or gene products ("sense-carrying units") of the existing species and to infer, on the basis of this information, the set of characters that the ancestor had, as well as the details of the evolutionary process that transformed this ancestor into the currently living species. In this essay, I focus on the inference of the list of protein-coding genes that LUCA might have included. Several related themes, such as the physical layout of genes and of their control elements, and evolution of the primary structure of these "sense-carrying units", though of considerable interest, will not be covered here.

3. THE DATA: GENOME TREE, GENE TREES, LISTS OF ORTHOLOGS, AND PHYLETIC VECTORS

A preliminary statement of the problem is the following: for each protein-coding gene in every sequenced genome, we would like to know whether its ancestral gene was present in LUCA. Thus, we want each known gene to be labeled as either ancestral or non-ancestral. All current methods of backward-in-time LUCA reconstruction rely for this purpose on two types of data. First, we need to know the phylogeny of the species (i.e., genomes) included in the

analysis; the root of this phylogeny is assumed to be LUCA. Second, we need to know, for each gene, the list of its homologs, and, more specifically, orthologs (2), in all genomes that have been sequenced thus far.

Each set of aligned orthologous genes can be used for phylogenetic inference, to produce a gene tree. A set of orthologous genes can also be characterized by its phyletic vector, i.e., a string of ones and zeroes that encodes the presence and absence of these orthologs in each sequenced genome (Figure 1). Since most genes have orthologs in only a subset of completely sequenced genomes, most phyletic vectors contain many zeroes. A phyletic vector is a result of labeling of the species' tree tips with the species present in the gene tree, but it lacks the topology information that a gene tree contains.

With these data on hand, we can modify the problem as follows: for each set of orthologous genes, we want to know whether their common ancestor was present in LUCA. In the context of phylogenetic reconstruction, this is equivalent to inferring the character state of the ancestral gene in LUCA, where the possible states are "present" and "absent". Other modifications of this question are also possible, for example, we may wish to know, for each set of orthologous genes, what is the deepest node in the species' tree at which its ancestor can be inferred to have had the state "present", or we may wish to infer the state of all orthologs at all internal nodes (the latter would be equivalent to knowing the set of genes in every species from LUCA to the extant genomes).

A proper list of orthologous genes in the extant species is required for any inference of ancestral gene sets. There are two types of approaches to constructing such lists. One way is to collect all homologs (orthologs and paralogs) using appropriate similarity search programs; to delineate all homologous families; to build a gene tree for each family; to infer all duplication and speciation events in each gene tree based on the algorithmically defined comparison between this gene tree and the species tree; and to partition homologous families into orthologs and paralogs. Though this approach appears to rely on the same resources as the LUCA reconstruction itself, i.e., comparison of gene tree and species tree, the logic is not circular (see references 3 and 4 for thoughts on the theory and for a practical algorithm, respectively). The other type of approach uses the notion of symmetric best matches, sometimes also called bidirectional best hits, which are pairs of genes in two genomes, one gene in each, that are one another's top-ranked matches in a database search, such as BLAST or FASTA. These pairs can be algorithmically processed to form clusters, representing the sets of most similar genes across genomes – a heuristic approach that is not fully consistent in dealing with paralogs, but, when implemented using algorithms such as COGNITOR (5), OrthoMCL (6), or INPARANOID (7) appears to give a good approximation of the proper set of orthologs, especially for gene families with moderate level of paralogy, which represent the majority of all genes.

4. ANCESTRAL STATE INFERENCE: GENE GAIN TO GENE LOSS RATIO IS A CRUCIAL PARAMETER

Mirkin and co-authors (8), in a seminal reconstruction of the set of genes in LUCA, have collected phyletic vectors corresponding to each orthologous set of genes, represented in their case by the NCBI Clusters of Orthologous Groups (COGs; reference 5) found in 26 complete genomes of bacteria and archaea (eukaryotes can be omitted because they are considered to be derived life forms in most of the seriously discussed evolutionary scenarios). In extremely general terms, the pattern of presences and absences of each gene in existing species and in their ancestors is the sum of two processes: gene gains and gene losses. Each species may inherit a gene/COG from its immediate ancestor, and then either retain or lose this gene. Or, if a gene was not found in the ancestral species, then the descendant species may either continue without this gene, or gain this gene from some source. There are three sources of gene gains: duplication of an existing gene followed by divergence; *de novo* emergence of a new ORF from a non-coding sequence or by recoding; and gain of a gene from another organism by horizontal gene transfer.

The main observation of any large set of genes/COGs is that only a small proportion of them, less than 60 genes by the most current account (9), are found in every sequenced genome without exception. A straightforward and usually correct explanation for such a vector with all coordinates set to one ("present") is that these genes were in LUCA and were inherited by all its ancestors, including the extant species with completely sequenced genomes. There are also COGs that are found in almost all species, but are missing in a few of them: for example, about 100 COGs are found in 95-99% of the completely sequenced genomes (9). An intuitive, and not always wrong, explanation in this case is that such COGs were also found in LUCA and have been lost in a few lineages (gene losses in parasites, many of which are sequenced, are self-evident and well-studied, and examples of gene losses in large genomes of free-living prokaryotes have also been documented – e.g., references 10 and 11).

It should be noted, however, that these types of phyletic vectors account only for a small fraction of all COGs. In contrast, the majority of COGs are distributed sparsely: about 90% of COGs are found in 20% of genomes or less (9).This brings into sharp focus the need to explicate an evolutionary model that can account for such observations. A model that invokes gene gains at the root, followed by an occasional gene loss on the way to some of the present-day species, may look like a straightforward scenario for the COGs that were found in the vast majority of species, but it faces a paradox when applied to these sparse phyletic vectors. If we were to explain all of them by multiple losses in a large number of lineages, this is equivalent to assuming that gene losses are more common in evolution than gene gains by almost two orders of magnitude: as 90% of all COGs are found in 20 or less species out of 110 in the latest available version of the

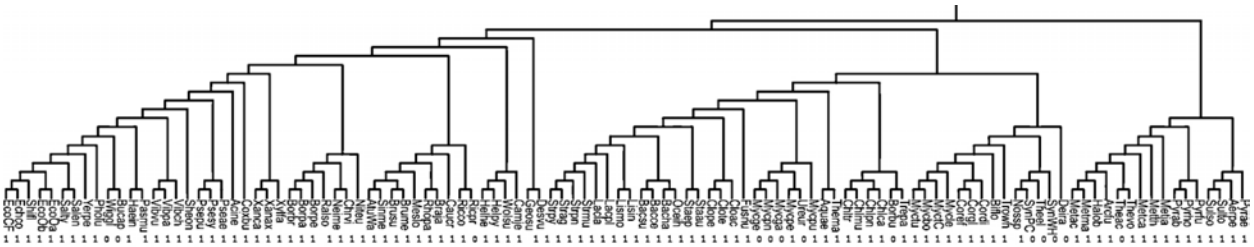

Figure 1. Species tree and phyletic vector of one gene. The maximum-likelihood tree of 110 species included in the recent release of the NCBI COG database (Yu. Wolf and E.Koonin, personal communication) was rooted between Bacteria and Archaea. The phyletic vector, shown under the tips of the tree, is for COG00350, Methylated DNA-protein cysteine methyltransferase. Species abbreviations are as follows: Acine, *Acinetobacter sp*. ADP1 (*Gammaproteobacteria*); Aerpe, *Aeropyrum pernix* (*Crenarchaeota*); AtuWa, *Agrobacterium tumefaciens* str. C58 (*Alphaproteobacteria*); Aquae, *Aquifex aeolicus* VF5 (*Aquificales-Thermotogales*); Arcfu, *Archaeoglobus fulgidus* DSM 4304 (*Euryarchaeota*); Bacce, *Bacillus cereus* ATCC 14579 (*Firmicutes*); Bacha, *Bacillus halodurans* (*Firmicutes*); Bacsu, *Bacillus subtilis* subsp. *subtilis* (*Firmicutes*); Biflo, *Bifidobacterium longum* NCC2705 (*Actinobacteria*); Borbr, *Bordetella bronchiseptica* RB50 (*Betaproteobacteria*); Borpa, *Bordetella parapertussis* (*Betaproteobacteria*); Borpe, *Bordetella pertussis* (*Betaproteobacteria*); Borbu, *Borrelia burgdorferi* (*Spirochaetes*); Braja, *Bradyrhizobium japonicum* USDA 110 (*Alphaproteobacteria*); Brume, *Brucella melitensis* 16M (*Alphaproteobacteria*); Brusu, *Brucella suis* 1330 (*Alphaproteobacteria*); Bucap, *Buchnera aphidicola* str. APS (*Gammaproteobacteria*); Camje, *Campylobacter jejuni* subsp. *jejuni* NCTC 11168 (*Epsilonproteobacteria*); Caucr, *Caulobacter crescentus* CB15 (*Alphaproteobacteria*); Chlmu, *Chlamydia muridarum* (*Chlamydiae*); Chltr, *Chlamydia trachomatis* (*Chlamydiae*); Chlca, *Chlamydophila caviae* GPIC (*Chlamydiae*); Chlpn, *Chlamydophila pneumoniae* AR39 (*Chlamydiae*); Chrvi, *Chromobacterium violaceum* ATCC 12472 (*Betaproteobacteria*); Cloac, *Clostridium acetobutylicum* (*Firmicutes*); Clope, *Clostridium perfringens* str. 13 (*Firmicutes*); Clote, *Clostridium tetani* E88 (*Firmicutes*); Cordi, *Corynebacterium diphtheriae* (*Actinobacteria*); Coref, *Corynebacterium efficiens* YS-314 (*Actinobacteria*); Corgl, *Corynebacterium glutamicum* (*Actinobacteria*); Coxbu, *Coxiella burnetii* RSA 493 (*Gammaproteobacteria*); Deira, *Deinococcus radiodurans* R1 (*Deinococcus-Thermus* group); Desvu, *Desulfovibrio vulgaris* subsp. *vulgaris* str. Hildenborough (*Deltaproteobacteria*); EcoCF, *Escherichia coli* CFT073 (*Gammaproteobacteria*); Echco, *Escherichia coli* K12 (*Gammaproteobacteria*); EcoOa, *Escherichia coli* O157:H7 (*Gammaproteobacteria*); EcoOb, *Escherichia coli* O157:H7 (*Gammaproteobacteria*); Fusnu, *Fusobacterium nucleatum* subsp. *nucleatum* ATCC 25586 (*Firmicutes*); Geosu, *Geobacter sulfurreducens* PCA (*Deltaproteobacteria*); Haein, *Haemophilus influenzae* Rd KW20 (*Gammaproteobacteria*); Halob, *Halobacterium sp*. NRC-1 (*Euryarchaeota*); Helhe, *Helicobacter hepaticus* ATCC 51449 (*Epsilonproteobacteria*); Helpy, *Helicobacter pylori* 26695 (*Epsilonproteobacteria*); Lacpl, *Lactobacillus plantarum* WCFS1 (*Firmicutes*); Lacla, *Lactococcus lactis* subsp. *lactis* (*Firmicutes*); Lisin, *Listeria innocua* (*Firmicutes*); Lismo, *Listeria monocytogenes* EGD-e (*Firmicutes*); Meslo, *Mesorhizobium loti* (*Alphaproteobacteria*); Metja, *Methanocaldococcus jannaschii* (*Euryarchaeota*); Metca, *Methanopyrus kandleri* AV19 (*Euryarchaeota*); Metac, *Methanosarcina acetivorans* C2A (*Euryarchaeota*); Metma, *Methanosarcina mazei* Goe1 (*Euryarchaeota*); Metth, *Methanothermobacter thermautotrophicus* str. Delta H (*Euryarchaeota*); Mycbo, *Mycobacterium bovis* subsp. *bovis* AF2122/97 (*Actinobacteria*); Mycle, *Mycobacterium leprae* (*Actinobacteria*); MyctC, *Mycobacterium tuberculosis* CDC1551 (*Actinobacteria*); Myctu, *Mycobacterium tuberculosis* H37Rv (*Actinobacteria*); Mycga, *Mycoplasma gallisepticum* R (*Mollicutes*); Mycge, *Mycoplasma genitalium* (*Mollicutes*); Mycpe, *Mycoplasma penetrans* (*Mollicutes*); Mycpn, *Mycoplasma pneumoniae* (*Mollicutes*); Mycpu, *Mycoplasma pulmonis* (*Mollicutes*); Neime, *Neisseria meningitides* (*Betaproteobacteria*); Niteu, *Nitrosomonas europaea* ATCC 19718 (*Betaproteobacteria*); Nossp, *Nostoc sp*. PCC 7120 (*Cyanobacteria*); Oceih, *Oceanobacillus iheyensis* HTE831 (*Firmicutes*); Pasmu, *Pasteurella multocida* (*Gammaproteobacteria*); Pholu, *Photorhabdus luminescens* subsp. *laumondii* TTO1 (*Gammaproteobacteria*); Pseae, *Pseudomonas aeruginosa* PAO1 (*Gammaproteobacteria*); Psepu, *Pseudomonas putida* KT2440 (*Gammaproteobacteria*); Psesy, *Pseudomonas syringae* pv. tomato str. DC3000 (*Gammaproteobacteria*); Pyrae, *Pyrobaculum aerophilum* str. IM2 (*Crenarchaeota*); Pyrab, *Pyrococcus abyssi* (*Euryarchaeota*); Pyrfu, *Pyrococcus furiosus* DSM 3638 (*Euryarchaeota*); Pyrho, *Pyrococcus horikoshii* (*Euryarchaeota*); Ralso, *Ralstonia solanacearum* (*Betaproteobacteria*); Rhopa, *Rhodopseudomonas palustris* CGA009 (*Alphaproteobacteria*); Ricco, *Rickettsia conorii* (*Alphaproteobacteria*); Ricpr, *Rickettsia prowazekii* (*Alphaproteobacteria*); Salen, *Salmonella enterica* subsp. *enterica* serovar *typhi* (*Gammaproteobacteria*); Salty, *Salmonella typhimurium* LT2 (*Gammaproteobacteria*); Sheon, *Shewanella oneidensis* MR-1 (*Gammaproteobacteria*); Shifl, *Shigella flexneri* 2a str. 301 (*Gammaproteobacteria*); Sinme, *Sinorhizobium meliloti* (*Alphaproteobacteria*); Staau, *Staphylococcus aureus* subsp. *aureus* Mu50 (*Firmicutes*); Staep, *Staphylococcus epidermidis* ATCC 12228 (*Firmicutes*); Strag, *Streptococcus agalactiae* 2603V/R (*Firmicutes*); Strmu, *Streptococcus mutans* UA159 (*Firmicutes*); Strpn, *Streptococcus pneumoniae* TIGR4 (*Firmicutes*); Strpy, *Streptococcus pyogenes* (*Firmicutes*); Sulso, *Sulfolobus solfataricus* (*Crenarchaeota*); Sulto, *Sulfolobus tokodaii* (*Crenarchaeota*); SynWH, *Synechococcus sp*. WH 8102 (*Cyanobacteria*); SynPC, *Synechocystis sp*. PCC 6803 (*Cyanobacteria*); Theac, *Thermoplasma acidophilum* (*Euryarchaeota*); Thevo, *Thermoplasma volcanium* (*Euryarchaeota*); Theel, *Thermosynechococcus elongatus* BP-1 (*Cyanobacteria*); Thema, *Thermotoga maritima* (*Thermotogales*); Trepa, *Treponema pallidum* (*Spirochaetes*); Trowh, *Tropheryma whipplei* TW08/27 (*Actinobacteria*); Ureur, *Ureaplasma urealyticum* (*Mollicutes*); Vibch, *Vibrio cholerae* (*Gammaproteobacteria*); Vibpa, *Vibrio parahaemolyticus* RIMD 2210633 (*Gammaproteobacteria*); Vibvu, *Vibrio vulnificus* CMCP6 (*Gammaproteobacteria*); Wiggl, *Wigglesworthia glossinidia* (*Gammaproteobacteria*); Wolsu, *Wolinella succinogenes* (*Epsilonproteobacteria*); Xanax, *Xanthomonas axonopodis* pv. *citri* str. 306 (*Gammaproteobacteria*); Xanca, *Xanthomonas campestris* pv. *campestris* str. ATCC 33913 (*Gammaproteobacteria*); Xylfa, *Xylella fastidiosa* 9a5c (*Gammaproteobacteria*); Yerpe, *Yersinia pestis* CO92 (*Gammaproteobacteria*)

COG resource, each of these COGs would have experienced more than 90 gene losses per one gene gain in this scenario. This estimate of loss-to-gain ratio in evolution is not supported by any evidence. Moreover, this scenario also means that no new genes emerged after LUCA, and that the LUCA genome contained the ancestor of every COG, resulting in a large genome size of 14000 genes, hardly imaginable for a prokaryote.

The problem is remedied, if only partially, by taking into account the fact that many of the phyletic vectors with many zeroes display "clustered", rather than "patchy", phyletic patterns, whereby a COG is found in a clade, not in an evolutionarily disperse group of species. It should be noted that by the very procedure of COG construction, only sets of orthologous genes from distinct lineages are counted as COGs: three orthologs, if they are found in three distinct lineages of *Proteobacteria*, qualify as a COG, but three orthologs of the same gene from three strains or closely related species of *Enterobacteriaceae* do not, so each COG consists of genes separated by millions of years of evolution. Even so, many of the COGs are distributed only within a subtree of the genome tree. Such COGs may be explained by gene gain (either by duplication or *de novo* formation) in the last common ancestor of the species that currently have this COG, perhaps followed by some gene losses. This reduces the excess of gene losses over gene gains, paints a more realistic picture of continuous emergence of new genes, and reduces the number of genes in LUCA several-fold, but still give estimates close to the upper bound of gene numbers in the known prokaryotic genomes. Thus, the single-gain scenarios described above give implausible estimates of gene loss-to-gain ratios, inflate the number of genes in LUCA, and do not explain "patchy" phyletic vectors. A more realistic scenario has to incorporate multiple gene gains along different branches of the species tree in order to overcome these difficulties.

All this brings the question of horizontal gene transfer (HGT) to the forefront of any effort of inferring ancestral gene sets. HGT appears to be the only plausible mechanism for repeated emergence of orthologs in several different parts of the species' tree. Indeed, parallel or convergent emergence of orthologs can be ruled out by everything that is known about evolution of biological sequences (12), with the only possible exception in the form of convergent origin of simple or repeated sequences, which, however, are not prominent in COGs and have modest influence on the reconstruction. Another objection to HGT scenarios alleges the errors of ortholog definition: a patchy phyletic vector is either said to include paralogous or even unrelated genes, and therefore the COG has to be split, or there is unrecognized orthology of two COGs, in which case two or more COGs have to be merged. In the minds of some critics, this would reduce the problem either precisely or nearly to the single-gain, many-losses scenario. In my opinion, the burden of proof of any statistical bias in any database of orthologous genes has not been met by these critics (see reference 9 for further discussion). Thus, a considerable fraction of phyletic vectors will have to be explained by some combination of three factors: the first emergence of this gene at a particular node of the tree, transfer of this gene between branches of the tree, and losses of this gene in some lineages.

One principled way to know the relative contribution of all these factors to the composition and evolution of the ancestral genome would be to estimate all the relevant parameters from the data, perhaps in the maximum-likelihood framework. This has not been done yet, but Mirkin and co-authors (8) have provided algorithms that reconstruct the ancestral state of each COG on the basis of a modified minimum-evolution principle. The results, as expected, turned out to be most sensitive to the value of a single parameter, g, the "gain penalty", which assigns relative weights to the gene gains and gene losses in the equation that determines the "amount of evolution" needed to explain each phyletic vector given the species tree (the importance of this factor has been emphasized in an earlier work by Snel and co-authors (10), whose main attention, however, was at the dynamics of gains and losses along various branches of the species tree, rather than at the reconstruction of the ancestral state of genes). When $g >> 1$, many gene losses count the same as one gene gain, and when $g << 1$, many gene gains count the same as one gene loss. When $g >> 1$, all phyletic vectors will tend to be explained mostly by gene losses, because those are relatively cheap, and this will also have the effect of pushing the first emergence of the gene back in time, closer to LUCA. When $g << 1$, all vectors will tend to be explained by gene gains (i.e., extensive HGT), with the effect of pushing the first gene gain away from LUCA. Interestingly, at $g \approx 1$, the amount of evolution needed to account for all phyletic vectors in the dataset is at a pronounced minimum, suggesting that the most concise explanation of the observed gene content in the existing genomes is achieved when the rate of gene gain and gene loss is about the same. (This is slightly different from saying that rate of gene loss and HGT is the same in evolution, because gene gain includes both HGT and gene appearance *de novo*). The distribution of COGs by the number of events required to explain their evolution peaks around 3, suggesting that most genes may have experienced at most two events after their birth, and that a substantial fraction of all genes may therefore have been horizontally transferred either once or never. On the other hand, a considerable fraction of all genes must have been transferred more than once in their lifetime. This HGT distribution turned out to be in agreement with the later observations, at vastly different evolutionary scales, of HGT rates in proteobacteria (11) and in double-stranded DNA bacteriophages (13).

The arguments for a special significance of $g \approx 1$ are based on parsimony, but, as memorably said by T.Cavalier-Smith, life is not parsimonious with respect to losses of genes (14). Even though Mirkin *et al.* was minimized the function representing the number of all evolutionary events, and the amount of losses was not minimized separately, one wonders whether the estimate makes any biological sense. But it turns of that the model with $g \approx 1$ has two more special properties, which fit biological sensibilities. Namely, as g is grows and becomes

closer to 1, the number of genes in LUCA increases as expected (reduced number of independent gains means that more of them need to have occurred earlier), and the sharpest increase is observed when *g* is changed from 0.9 to 1. Moreover, if we examine the list of biological functions represented by the COGs that are placed into LUCA under these growing values of *g*, it is the LUCA at *g*=1 (called LUCA1.0) that for the first time becomes metabolically coherent (see next section).

It should be noted that the reconstruction proposed by Mirkin *et al.* was made on the basis of ~3000 COGs in 26 genomes. More recently, Ouzounis *et al.* inferred the gene set of LUCA using a similar in spirit, but different in technical details, algorithmic approach (15). That latter effort was based on the OFAM database that contained more than 37,000 orthologous families from 168 genomes. Despite this significantly larger dataset, the estimate for the minimally plausible LUCA genome was not dramatically different from the smallest plausible LUCA of Mirkin *et al.* (669 families in the former vs. 572 COGs in the latter, compatible with perhaps overly conservative delineation of orthologs in OFAM compared to COGs), suggesting that the majority of ancestral genes that are surviving to this day have been already sampled by the genome projects.

5. WHICH GENES ARE IN LUCA AND WHICH ARE NOT

LUCA1.0, with 572 genes, has the complete translation apparatus, except for glycyl-tRNA synthetase. This is in full agreement with the preponderance of the essential genes and generally vertical pattern of inheritance among the proteins involved in translation. The most likely cause of failure to recover glycyl-tRNA synthetase is non-orthologous gene displacement (see below), rather than a peculiar version of genetic code in LUCA. LUCA1.0 also had transcription machinery consisting of the basal RNA polymerases subunits, transcription termination factors and several helix-turn-helix regulators. Energy supply needs in LUCA1.0 were met by almost complete glycolysis (only phosphoglyceromutase was missing, in another well-known case of gene displacement), complete TCA cycle, and the complete set of the H+-ATPase subunits. Intermediary metabolism was represented by nucleotide salvage (interestingly, this biosynthetic module is the first to emerge upon the increase of *g* from 0.1 to higher numbers), nucleobase biosynthesis, and substantially full pathways for biosynthesis of amino acids. On the other hand, only salvage of complex coenzymes is represented in LUCA1.0, without any complete *de novo* coenzyme biosynthesis pathway, and the repertoire of membrane transporters in LUCA1.0 was narrow.

Of great interest are not only those genes present in LUCA, but those that are absent as well. Two essential biopolymers are physically contiguous between the generations of prokaryotic cells: DNA and membrane lipids. The enzymes that have to do with the maintenance of these two classes of molecules are missing in most LUCA reconstructions. In particular, there are no replicative DNA polymerase, helicase and replication initiation ATPase in LUCA1.0, and the enzymes for lipid side chain biosynthesis are also missing. This raises the question about the status of the DNA genome and cytoplasmic membrane in LUCA.

The question of the chemical composition of the LUCA genome is not settled. It has been hypothesized that LUCA might have had an RNA genome and that DNA replication could have been invented (and processive DNA polymerase with accompanying ATPases recruited) twice independently, once in a lineage leading to Bacteria, and again in the stem of Archaea/Eukarya (16). This, however, does not explain the presence in LUCA of several enzymes that are involved in biosynthesis of deoxyribonucleotides, such as flavin-dependent thymidylate synthase and two subunits of ribonucleotide reductase. To reconcile these facts, it has been proposed that DNA genomes were invented by virus-like parasites and then appropriated by cellular genomes two or three times (17). Another theory is that LUCA had an RNA genome that replicated via a DNA intermediate, similarly to present-day retroid viruses (16, 18), and this strategy was superseded by modern-type DNA replication on two occasions.

As for lipids, the main conundrum is the following. In bacteria and eukaryotes, the side chains of membrane lipids consist of fatty acids that are synthesized using essentially one and the same pathway. In archaea, membrane lipids lack fatty acids, which are replaced by isoprenoids. Eukaryotes and most bacteria also produce isoprenoids, but in these organisms, isoprenoids play no direct role as lipid side chains. Isoprenoids can be synthesized either by the mevalonate pathway that is currently seen in eukaryotes, in some bacteria, and in modified form in archaea, or by the methylerithrytol phosphate pathway that is restricted to bacteria (and plant chloroplasts). The evolutionary history of these pathways remains to be fully understood. As far as LUCA is concerned, a recent hypothesis states that ancient cells may have not needed lipid membranes, as they lived in inorganic microcompartments, where the role of the cell membrane was essentially performed by minerals (18, 19). The “escape into cellular world” may have occurred twice: once by an organism with bacteria-like DNA replication, enabled by invention of fatty acids biosynthesis, and another time by the archaeal lineage, perhaps following the invention of the mevalonate pathway of isoprenoid biosynthesis (20, 21).

Thus, there may be good reasons to infer the LUCA without lipid side chains and with an unconventional mode of genome replication. But there is also a more mundane possibility that some genes found in the existing organisms cannot be reliably placed in LUCA because of insufficient information in their phyletic vectors. In particular, if a gene is retained only in a group of the closely related species, we may be unable to infer it beyond the common ancestor of this set of species. This problem is closely linked to the phenomenon of displacement of orthologous genes (9, 22), i.e., the fact that isofunctional proteins are not always orthologous, and sometimes are not

even homologous. For example, if the same function is performed by three non-orthologous proteins, one in archaea, another in gammaproteobacteria, and yet another in actinomycetes and spirochetes, none of these COGs stands a particularly good chance to be placed into LUCA, even if one of them is in fact ancestral. This is the main reason for underestimation of the number of genes and proteins in LUCA.

6. THE EFFECTS OF TREE TOPOLOGY AND OF HORIZONTAL GENE TRANSFER

The inference of LUCA gene set requires a rooted species tree. The topology of the Tree of Life, however, is still a subject of debate, and the position of the root in that tree is sometimes also contested (see reference 23 for a summary and further references). Mirkin *et al.* (8) noted that the gene set of LUCA was sensitive to the changes to the tree topology that occurred when different datasets were used for inference of the species tree (e.g., ribosomal RNA vs. universally conserved proteins) and when different principles of tree building were employed (e.g., maximum parsimony vs. maximum likelihood). Generally speaking, if one tree has several deep branches, but in another tree these branches are clustered, then genes that are present only in these branches will tend to be removed from LUCA in the latter tree, because their origin can be placed no further back than the root of the cluster. This emphasizes the importance of high resolution of the species tree for LUCA reconstruction and suggests that all inferences of LUCA ideally have to be explicitly probabilistic, taking into account all uncertainties in the data as well as in the evolutionary model.

Recently, Dagan and Martin (24) studied the effect of varying root positions on the gene sets in LUCA. Their main conclusion was that moving the root of the Tree of Life from its canonical position (between Bacteria and Archaea) into new locations within proteobacteria, actinomycetes, or mollicutes had only a trivial effect on the number of genes in LUCA, provided that the procedure of ortholog definition and the gain penalty (in their approach, HGT penalty) stayed the same. In contrast, what would change dramatically with the changed position of the root is the identity of genes that make the list. For example, if the root of the Tree of Life is within proteobacteria (as almost no one except T.Cavalier-Smith would like it - see reference 14), then the problems of DNA and lipid biosynthesis discussed in the previous section would go away – LUCA would have a bacterial-type replisome with class C DNA polymerase and fatty acids for the lipid side chains. If the root of the Tree is within archaea, we would likewise see a DNA replication apparatus, albeit in this case of archaeo/eukaryal type, and lipid side chains, albeit made of isoprenoid and synthesized by modified mevalonate pathway (a phosphomevalonate decarboxylase activity appears to be needed for completion of this pathway in archaea but has not been identified thus far). These alternative positions of the root either within Bacteria or within Archaea also cause major changes in the repertoire of ancestral metabolic enzymes. For example, a proteobacterial placement of the root leads to the ancestor that was capable of synthesizing the peptidoglycan cell wall, whereas archaeal root results in chemoautotrophy (25).

Though the overall dynamics of gene gain, loss, and horizontal transfer has become better understood in recent years, the effect of HGT on the placement of individual genes into LUCA remains to be studied in more detail. Algorithmic methods of HGT inference are being actively developed. Many, if not all of them, rely on detecting differences between different gene trees, or between a gene tree and the species tree (26, 27). For example, if a gene is shared by proteobacteria and archaea, then in the absence of HGT it could be placed into LUCA, given the canonical position of the root and methods discussed in Section 4. But if analysis of the gene family tree indicates that this gene has been gained by the proteobacterial clade by horizontal transfer from archaea, this in effect changes the gene from widely to narrowly distributed (i.e., the patchy phyletic vector turns into a clustered one), and as a result, the origin of this gain is inferred at the root of archaea, not in LUCA. Methods that determine the acts of HGT transfer should be applied to find the genes that have been disseminated by HGT between distant branches of the Tree of Life, and to reinvestigate the ancestral status of these genes.

7. ANCESTRAL GENES AND "LUCA-LIKENESS"OF THE PRESENT-DAY GENOMES

With the tentative list of ancestral COGs in hand, we can study their distribution in the extant genomes. I used the list of COGs assigned to LUCA1.0 in reference 8 and examined them in the context of the latest available COG resource, with 110 species and $1.4 \cdot 10^3$ COGs. As expected, the distribution of these genes by the number of genomes is very different from the COG database as a whole, with 90% of LUCA1.0 COGs found in more than half of all genomes.

Several other trends also agree with common sense. Larger genomes with more COGs in them tend to have more ancestral COGs, as well as more COGs supported by a larger number of genomes (Figure 2 and data not shown). Interestingly, however, not a single genome contains every ancestral COG (the largest number of ancestral COGs in any genome is 537, in *Salmonella enterica*), indicating that gene loss can be visible even against the countervailing evolutionary drive towards the growth of genome size. On the other hand, each species, without exception, contains ancestral COGs that are found in less than half of all species. For example, the smallest Gram-positive-like genome in the dataset, a mollicute *Ureaplasma urealyticum*, and the smallest Gram-negative genome, a gammaproteobacterium *Buchnera aphidicola*, each contain a distinct set of four such rare ancestral COGs.

"LUCA-likeness", i.e., the proportion of the ancestral COGs in the present-day genomes, ranges from just under 18% for the largest genomes in the dataset to more than 50% for the smallest genomes (Figure 2). Though the number of conserved genes (COGs) in the

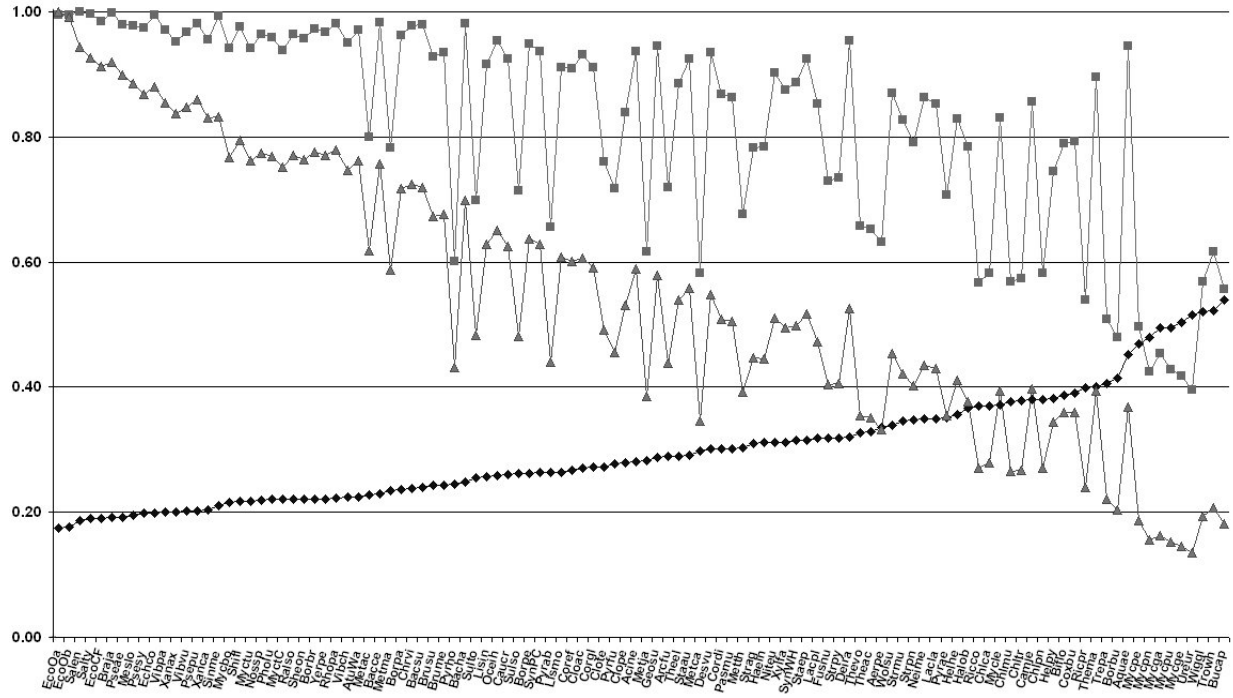

Figure 2. LUCA-likeness of genomes**.** Squares indicate the number of ancestral COGs in each genome, triangles indicate the total number of COGs in each genome, and black diamonds indicate the ratio of ancestral COGs to total COGs in each genome (LUCA-likeness). The top two curves are scaled so that the maximal number of COGs equals 100%. Species' abbreviations are as in Figure 1.

genome is inversely proportional to LUCA-likeness and tends to be directly proportional to the number of ancestral COGs, these trends are relatively weak among genomes with low LUCA-likeness. These large genomes may differ by as many as ~800 COGs (33% of total COG number), yet the difference in their ancestral COGs does not exceed 10%. Thus, in these complex genomes (which are not a monophyletic group, but rather a mix of actinomycetes and several divisions of proteobacteria), most of the difference between gene repertoires is due to the evolutionarily novel genes: there is just not enough ancestral genes to provide for such large changes. At the other extreme, the genomes of parasitic microorganisms with reduced biosynthetic capacity show a "jump" in the LUCA-likeness value. This is hardly the sign of their ancient origin – indeed, most of them are not deep clades in the tree, but either derived proteobacteria or derived Gram-positive bacteria – but rather is the reflection of the fact that small genomes tend to be strongly enriched in essential genes and in omnipresent genes, the categories that have a large overlap with ancestral genes.

The highest LUCA-likeness among free-living species is observed in two thermophilic bacteria, *Aquifex aeolicus* and *Thermotoga maritima.* However tempting this might be to use this observation as the evidence of the thermophilic lifestyle of LUCA, I do not think this is a correct explanation. Indeed, these two genomes appear to have acquired an unusually high number of genes by horizontal transfer from several thermophilic archaea (28, 29), undoubtedly facilitated by their co-habitation in thermal vents. This results in placing into LUCA many genes that are shared by archaea and one or both of these species, even though many such genes should properly be placed no further than the common ancestor of all Archaea. This is where the re-evaluation of HGT discussed in the previous section needs to play a role. The identity of the free-living species that has vertically inherited the largest proportion of ancestral COGs remains to be established.

8. LUCAS INSTEAD OF LUCA?

The "backward-in-time" reconstructions of gene content in LUCA that I discussed in the previous sections all share one general assumption, namely that LUCA was a cellular organism with spatially self-contained genome and metabolism (even if separation from the environment was provided by a porous wall made of mineral, and not by a lipid membrane). A different view (30, 31) states that there may have been no such a cell, and the inhabitants of the "compartments" are best viewed as the communal assembly of genetic and metabolic molecules – the Last Universal Common Ancestral State (LUCAS; ref. 32) – which teems with genes or gene fragments that are free to recombine and form loose associations. Postulating this stage in the evolution of Life is attractive because it helps to explain in the same framework several major, and apparently extremely rapid, evolutionary transitions, such as the origin of genetic code and, later, explosive generation of several major phylae of both Bacteria and

Archaea (31, 33). Some of the implications of such a hypothesis for the reconstruction of the set of ancestral genes are relatively trivial, while others are more profound.

An ancestral gene either has survived in some of the existing genomes, or it is missing from all of them. As far as any individual gene is concerned, the issue is exactly the same regardless of whether the ancestral life form was a cell-like LUCA or a community-like LUCAS: if the gene is missing from all lineages of Life (or, more to the point, from all the sampled lineages, i.e., from the sequence databases), it is unavailable for analysis, and its ancestral status will never be inferred. As we do not know whether LUCAS could have had a vastly different number of genes than a LUCA, it is hard to know what we are missing in either case.

A deeper problem is that in the case of LUCAS, the notions of gene gain and gene loss close to the root of the Tree (and indeed, the notion of such a root itself) become poorly defined, so that the models and algorithms discussed in this paper need a major revision.

9. CONCLUSIONS

In a short span of a few years in this century, the inference of ancestral gene content on the basis of comparative analysis of existing genomes has become a well-established area of research. In the same spirit as the inferences of LUCA discussed in this essay, reconstructions of more recent ancestors of various groups of unicellular organisms, such as lactobacteria (34), archaea (25), and fungi (35) have been described. In addition to the lists of genes, these studies produce technical advances, such as improved algorithms and more accurate estimations of important parameters (primarily the rate of gene loss and various types of gene gain). In principle, similar approaches are applicable to inference of the ancestral status of various non-protein-coding genes and domains, such as various functional regions in rRNA (36).

Recently, grave doubts have been expressed in our ability to obtain a highly resolved, accurately rooted Tree of Life (23, 33). The main issue here is not that frequent horizontal gene transfer in the past may have turned the hierarchical tree into a reticulated network: other than the technical issues of algorithmic complexity, there are no reasons why a cyclic graph should be less worthy of reconstruction, or less amenable to it, than an acyclic one. A more difficult problem is that the data may not contain sufficient signal to support any topology of the earliest branches of prokaryotic life, and that "Big Bang-like" origin of major evolutionary clades may preclude such resolution in principle (33). This would be too bad, because the list of the ancestral genes is dependent on the position of the tree root after all (see section 6).

If the root of the Tree of Life is placed conventionally, i.e., between Bacteria and Archaea, the ancestral gene set appears to be coherent but does not contain any smoking guns, which would give away the phenotype of LUCA and the mode of its interactions with the environment. We see a most likely heterotroph with the capacity of de novo biosynthesis of amino acids, nucleotides, and many sugars, but with reduced ability of coenzyme biosynthesis and with a limited repertoire of transporters, sensors and signal transduction systems. Displacement of orthologous genes or whole biological modules may have erased parts of the evolutionary record for these systems, but the hope springs eternal that continued deep sampling of the diversity of life on Earth will fill in the blanks and help the evolutionary signal to regain its strength. Even in the absence of the ultimately resolved tree of life, the ancestral gene list may be refined by simulation, resampling, and other statistical approaches. Moreover, densely covered space of orthologous sequences may afford us with the possibility to reconstruct the primary structures of the ancestors of at least some genes that were present in LUCA, and to study their properties by direct biochemical experimentation.

10. ACKNOWLEDGMENTS

The author is grateful to Céline Brochier, Patrick Forterre and Simonetta Gribaldo for organizing the meeting "LUCA: Ten years after" (Fondation des Treilles, 4-9 September 2006), to all the participants of that meeting for inspiring discussions, and to David Kristensen for help with the manuscript.

11. REFERENCES

1. E. Zuckerkandl and L. Pauling: Molecules as documents of evolutionary history. *J Theoret Biol* 8, 357-366 (1965)

2. W.M. Fitch: Homology: a personal view on some of the problems. *Trends Genet* 16, 227-231 (2000)

3. W.M. Fitch: Uses for evolutionary trees. *Philos Trans R Soc Lond B Biol Sci* 349, 93-102 (1995)

4. C.M. Zmasek and S.R.Eddy: A simple algorithm to infer gene duplication and speciation events on a phylogenetic tree. *Bioinformatics* 17, 821-828 (2001)

5. R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova and E.V. Koonin: The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29, 22-28 (2001)

6. A.J. Enright, S. Van Dongen and C.A. Ouzounis: An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30, 1575-1584 (2002)

7. K. O'Brien, M. Remm and E. Sonnhammer E: Inparanoid: The eukaryotic ortholog database. *Nucleic Acids Res* 33, D476-D480 (2005)

8. B.G. Mirkin, T.I. Fenner, M.Y. Galperin and E.V. Koonin: 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 (2003)

9. Arcady Mushegian. Foundations of Comparative Genomics. Amsterdam, Boston : Academic Press (2007)

10. B. Snel, P. Bork and M. A. Huynen: Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res* 12, 17-25 (2002)

11. E .Lerat, V. Daubin, H. Ochman and N.A. Moran: Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* 3, e130 (2005)

12. R.F. Doolittle: Convergent evolution: the need to be explicit. *Trends Biochem Sci* 19, 15-18 (1994)

13. G. Glazko, V. Makarenkov, J. Liu and A. Mushegian: Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol Direct* 2, 36 (2007)

14. T. Cavalier-Smith: The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52, 7-76 (2002)

15. C.A. Ouzounis, V. Kunin, N. Darzentas and L. Goldovsky: A minimal estimate for the gene content of the last universal common ancestor – exobiology from a terrestrial perspective. *Res Microbiol* 157, 57-68 (2006)

16. D.D. Leipe, L. Aravind and Koonin EV: Did DNA replication evolve twice independently? *Nucleic Acids Res* 27, 3389-3401 (1999)

17. P. Forterre: Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain. *Proc Natl Acad Sci USA* 103, 3669-3674 (2006)

18. E.V. Koonin, T.G. Senkevichand V.V Dolja: The ancient Virus World and evolution of cells. *Biol Direct* 1, 29 (2006)

19. E.V. Koonin and W. Martin: On the origin of genomes and cells within inorganic compartments. *Trends Genet* 21, 647-654 (2005)

20. A. Smit and A. Mushegian: Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. *Genome Res* 10, 1468-1484 (2000)

21. J. Peretó, P. López-García and D. Moreira: Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem Sci* 29, 469-477 (2004)

22. E.V. Koonin, A.R. Mushegian and P. Bork: Non-orthologous gene displacement. *Trends in Genetics* 12, 334-336 (1996)

23. T. Dagan and W. Martin: The tree of one percent. *Genome Biol* 7, 118 (2006)

24. T. Dagan and W. Martin: Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104, 870-875 (2007)

25. K.S. Makarova, A.V. Sorokin, P.S. Novichkov, Y.I. Wolf and E.V. Koonin: Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct* 2, 33 (2007)

26. V. Makarenkov, P. Legendre: From a phylogenetic tree to a reticulated network. *J Comput Biol* 11, 195-212 (2004)

27. C. Than, D. Ruths, H. Innan and L. Nakhleh: Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J Comput* Biol. 14, 517-535 (2007)

28. L. Aravind, R.L. Tatusov, Y.I. Wolf, D.R. Walker and E.V. Koonin: Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles.*Trends Genet* 14, 442-444 (1998)

29. K.E. Nelson, R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen, O. White, S.L. Salzberg, H.O. Smith, J.C. Venter and C.M. Fraser: Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature*. 399, 323-329 (1999)

30. C.R. Woese: On the evolution of cells. *Proc Natl Acad Sci U S A*. 99, 8742-8747 (2002)

31. K. Vetsigian, C. Woese, and N. Goldenfeld: Collective evolution and the genetic code. *Proc Natl Acad Sci U S A*. 103, 10696-10701 (2006)

32. P. O'Donoghue, A. Sethi, C.R. Woese, and Z.A. Luthey-Schulten: The evolutionary history of Cys-tRNACys formation. *Proc Natl Acad Sci U S A*. 102, 19003-19008 (2005)

33. E.V. Koonin: The Biological Big Bang model for the major transitions in evolution. *Biol Direct* 2, 21 (2007)

34. K. Makarova, A. Slesarev, Y. Wolf, A. Sorokin, B. Mirkin, E. Koonin, A. Pavlov, N. Pavlova, V. Karamychev, N. Polouchine, V. Shakhova, I. Grigoriev, Y. Lou, D. Rohksar, S. Lucas, K. Huang, D.M. Goodstein, T. Hawkins, V. Plengvidhya, D. Welker, J. Hughes, Y. Goh, A. Benson, K. Baldwin, J.H. Lee, I. Díaz-Muñiz, B. Dost, V. Smeianov, W. Wechter, R. Barabote, G. Lorca, E. Altermann, R. Barrangou, B. Ganesan, Y. Xie, H. Rawsthorne, D. Tamir, C. Parker, F. Breidt, J. Broadbent, R. Hutkins, D. O'Sullivan, J. Steele, G. Unlu, M. Saier, T. Klaenhammer, P. Richardson , S. Kozyavkin, B. Weimer, and D. Mills: Comparative genomics of the lactic acid

bacteria. *Proc Natl Acad Sci USA.* 103, 15611-15616 (2006)

35. I. Wapinski, A. Pfeffer, N. Friedman, A. Regev: Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449, 54-61 (2007)

36. A. Mushegian: Protein content of minimal and ancestral ribosome. *RNA* 11, 1400-1406 (2005)

Abbreviations: LUCA, last universal common ancestor; COG, cluster of orthologous groups; HGT, horizontal gene transfer

Key Words: LUCA, Last Universal Common Ancestor, Horizontal Gene Transfer, Phylogeny, Review

Send correspondence to: Arcady Mushegian, Stowers Institute for Medical Research, 1000 E 50th St., Kansas City MO 64110, USA, Tel: 816-926-4021, Fax: 816-926-2041, E-mail: arm@stowers-institute.org

http://www.bioscience.org/current/vol13.htm