Orig Life Evol Biosph (2007) 37:415–418
DOI 10.1007/s11084-007-9092-3

Question 7: Comparative Genomics and Early Cell Evolution: A Cautionary Methodological Note

Sara Islas · Ricardo Hernández-Morales · Antonio Lazcano

Received: 4 March 2007 / Accepted: 11 April 2007 /
Published online: 28 June 2007

Abstract Inventories of the gene content of the last common ancestor (LCA), i.e., the cenancestor, include sequences that may have undergone horizontal transfer events, as well as sequences that have originated in different pre-cenancestral epochs. However, the universal distribution of highly conserved genes involved in RNA metabolism provide insights into early stages of cell evolution during which RNA played a much more conspicuous biological role, and is consistent with the hypothesis that extant living systems were preceded by an RNA/protein world. Insights into the traits of primitive entities from which the LCA evolved may be derived from the analysis of paralogous gene families, including those formed by sequences that resulted from internal elongation events. Three major types of paralogous gene families can be recognized. The importance of this grouping for understanding the traits of early cells is discussed.

Keywords Cenancestor · Paralogous duplications · Gene elongation events · RNA/protein world

The universal distribution of the genetic code, the same essential features of genome replication and gene expression, basic anabolic reactions, and membrane-associated ATPase mediated energy production suggests that they were already present in the last common ancestor (LCA), i.e., the cenancestor, of all living beings. It is of course unlikely that such traits were already present in the first forms of life, whose actual nature can only be surmised. Comparative genomics can provide important insights on intermediate stages analysis provides important insights on evolutionary stages that may have existed prior to the LCA and the separation of the three major cell lineages. However, such information cannot be extrapolated into older evolutionary stages, including the events that may have taken place on the prebiotic soup, nor on the RNA world itself. At the time being, the applicability of molecular cladistics and

Presented at: International School of Complexity – 4th Course: Basic Questions on the Origins of Life; "Ettore Majorana" Foundation and Centre for Scientific Culture, Erice, Italy, 1–6 October 2006.

S. Islas · R. Hernández-Morales · A. Lazcano (✉)
Facultad de Ciencias, UNAM,
Apdo. Postal 70-407, Cd. Universitaria, 04510 Mexico, D.F., Mexico
e-mail: alar@correo.unam.mx

comparative genomics cannot be extended beyond a threshold that corresponds to a period of cellular evolution in which protein biosynthesis was already in operation.

Studies of deep phylogenies provide important insights into the nature of the LCA itself. It could be argued that the most parsimonious characterization of the cenancestor could be achieved by summarizing the features of the oldest recognizable nodes of universal cladograms. The rooting of universal cladistic trees determines the directionality of evolutionary change and allows the recognition of ancestral from derived characters, i.e., primitive characters should appear in older, basal branches than do their derived counterparts. Determination of the rooting point of a tree normally imparts polarity to most or all characters. However, large-scale studies based on the availability of genomic data have revealed major discrepancies with the rRNA tree topology. Very often these differences have been interpreted as evidence of horizontal gene transfer events between different species and even domains, questioning the feasibility of the reconstruction and proper understanding of early biological history. Moreover, it is important to distinguish between ancient and primitive organisms. Species located near the root of universal rRNA-based trees are cladistically ancient, but they are not endowed with a primitive molecular genetic apparatus, nor seem to be more primitive in their metabolic abilities than their aerobic counterparts.

Reconstructions of gene complements of distant ancestors are mere statistical approximations of biological past, since their accuracy depends on manifold factors, including horizontal gene transfer, polyphyletic gene losses, the significant variations in substitution rates of different proteins, as well as methodological caveats, including the possible biases in the construction of genome databases (Becerra et al. 1997). Medical and veterinarian interests have shaped the nature of extant genome databases from which many species are absent, and that exclude, for the time being, representatives of all major biological groups. However, there is a significant overlap in the inventories of highly conserved sequences reported by different authors. Sequences involved in RNA metabolism, i.e., ORFs whose products synthesize, degrade, or interact with RNA, are among the most highly conserved sequences common to all known genomes, and provide insights into an early stage in cell evolution during which RNA played a much more conspicuous biological role. The conservation of this set of sequences is consistent with the proposal that the extant DNA/RNA/protein world was preceded by an RNA/protein world, an evolutionary stage in which ribonucleotide reduction and DNA genomes had not yet evolved (Becerra et al. 2007).

The available information suggests that the cenancestor was not an immediate descendant of the RNA world, a protocell, nor any other pre-life progenitor system, but that it was a complex organism, much alike extant prokaryotes. However, it should be kept in mind that inventories of LCA gene content include sequences that may have undergone horizontal transfer events, as well genes (or domains) that have originated in different pre-cenancestral epochs. For instance, invariant motifs that exhibit a surprising degree of conservation, such as GHVDHGKT, DTPGHVDF, and GAGKSTL (Goto et al. 2002), and the RNA-binding domains found in highly conserved genes (Delaye and Lazcano 2000), which may well be among the oldest recognizable polypeptides found in databases, and are very likely much more ancient that some of the proteins in which they are present.

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 phylogenies. 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 surprisingly the resulting repertoire is formed by relatively few features and by incompletely represented biochemical processes (Becerra et al. 2007). Analysis of some of the most likely a priori

candidates for strict universality, such as the molecular machinery involved in DNA replication, have turned out to be of polyphyletic origin (Edgell and Doolittle 1997). It has been argued that polymerases and topoisomerases may have an ultimate viral origin (Forterre 2006). However, not all the components of multidomain enzymes are equally ancient. This appears to the case of the catalytic palm subdomain of the Klenow fragment of DNA polymerase I, which appears to be a vestige of the RNA/protein world replicase (Becerra et al. 2007).

Understanding of the evolution of central metabolic pathways during pre-LCA epochs is hampered by the absence of one or more biosynthetic genes in the genomes of manifold free-living prokaryotes that have been sequenced. It is not easy to explain the troublesome absence of a number of biosynthetic genes in the genomes of manifold free-living prokaryotes that have been sequenced. Addressing this issue will require not only the identification and proper annotation of highly conserved open reading frames found in all cell genomes, but also more complete tertiary structure databases. The possibility that some of the enzymes of archaic pathways may have survived in unusual organisms suggest that considerable prudence should also be exerted when attempting to describe the physiology of ancestral organisms.

It is also possible that extant enzymes participated in metabolic pathways which no longer exist or remain to be discovered (Zubay 1993; Becerra and Lazcano 1998), a possibility that has begun to be explored by computer searches for alternative reaction pathways (Goto et al. 1996). The discovery that carbamate kinase, which participates in fermentative ATP production, catalyzes the formation of carbamoyl phosphate in the archaea *Pyrococcus furiosus* and *P. abyssi* (Alcántara et al. 2000) shows that considerable attention should be given to the possibility that significant variations of the basic biosynthetic pathways may have existed in the past.

Clues to the genetic organization and biochemical complexity of primitive entities from which the LCA evolved may also be derived from the analysis of paralogous gene families. The number of sequences that have undergone such duplications prior to the divergence of the three lineages includes genes encoding for a variety of enzymes that participate in widely different processes such as translation, DNA replication, biosynthetic pathways, and energy-producing processes. As noted elsewhere (Becerra et al. 2007), a survey of the available information shows that sequences that have resulted from early pre-ancestral paralogous expansion may be classified in three major groups:

(a) sequences formed by two tandemly arranged homologous modules which underwent fusion events, such as the (1) protein disulfide oxidoreductase (Ren et al. 1998), (2) large subunit of carbomoyl phosphate synthetase (Alcántara et al. 2000), and (3) HisA, an isomerase that forms part of the histidine biosynthetic pathway (Alifano et al. 1996);
(b) gene families which have undergone a major expansion of sequences, such as ABC transporters and other enzymes involved in membrane transport phenomena (Clayton et al. 1997); and
(c) families formed by a relatively small number of paralogous sequences. These includes, among others, the pair of homologous genes encoding the EF-Tu and EF-G elongation factors, (Iwabe et al. 1989) as well as the duplicated sequences encoding the F-type ATPase hydrophilic alpha and beta subunits (Gogarten et al. 1989).

The identification of sequences formed by tandemly fused homologous modules provides direct evidence of the existence during early Precambrian times of smaller, functional genes. Moreover, the families of paralogous duplicates also imply that the LCA was preceded by simpler cells with a smaller genome in which only one copy of each of

these genes existed, i.e., by cells in which, for instance, protein synthesis involved only one elongation factor, and with ATPases with limited regulatory abilities. Paralogous families of metabolic genes also support the proposal that anabolic pathways were assembled by the recruitment of primitive enzymes that could react with a wide range of chemically related substrates, i.e., the so-called patchwork assembly of biosynthetic routes (Jensen 1976). Such relatively slow, unspecific enzymes may have represented a mechanism by which primitive cells with small genomes could have overcome their limited coding abilities. How early cells overcame the bottlenecks imposed by such limitations is still an open problem than can be addressed experimentally.

Acknowledgments Support from CONACYT-México (Project 50520-Q) to A.L. is gratefully acknowledged.

References

Alcántara C, Cervera J, Rubio V (2000) Carbamate kinase can replace in vivo carbamoyl phosphate synthetase. Implications for the evolution of carbamoyl phosphate biosynthesis. FEBS Lett 484:261–264

Alifano P, Fani R, Liò P, Lazcano A, Bazzicalupo M, Carlomagno MS, Bruni CB (1996) Histidine biosynthetic pathway and genes: structure, regulation, and evolution. Microbiol Rev 60:44–69

Becerra A, Lazcano A (1998) The role of gene duplication in the evolution of purine nucleotide salvage pathways. Origins Life Evol Biosph 28:539–543

Becerra A, Islas S, Leguina JI, Silva E, Lazcano A (1997) Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. J Mol Evol 45:115–118

Becerra A, Delaye L, Islas S, Lazcano A (2007) Very early stages of biological evolution related to the nature of the last common ancestor of the three major cell domains. Ann Rev Ecol Evol Syst 38 (in press)

Clayton RA, White O, Ketchum KA, Venter CJ (1997) The genome from the third domain of life. Nature 387:459–462

Delaye L, Lazcano A (2000) RNA-binding peptides as molecular fossils. In: Chela-Flores J, Lemerchand G, Oró J (eds) Origins from the big-bang to biology: Proceedings of the First Ibero-American School of Astrobiology. Klüwer, Dordrecht, pp 285–288

Edgell RD, Doolittle WF (1997) Archaea and the origins of DNA replication proteins. Cell 89:995–998

Forterre P (2006) 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 U S A 103:3669–3674

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson ML, Poole RJ, Date T, Oshima T, Konishi L, Denda K, Yoshida M (1989) Evolution of the vacuolar H^+-ATPase, implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86:6661–6665

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, Kanehisa M (1996) Organizing and computing metabolic pathway data in terms of binary relations. In: Altman RB, Dunker K, Hunter L, Klein TE (eds) Pacific Symposium on Biocomputing '97 (World Scientific, Singapore), pp 175–186

Goto N, Kurokawa K, Yasunaga T (2002) Finding conserved amino acid sequences among prokaryotic proteomes. Genome Inform 13:443–444

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Jensen RA (1976) Enzyme recruitment in the evolution of new function. Annu Rev Microbiol 30:409–425

Ren B, Tibbelin G, de Pascale D, Rossi M, Bartolucci S, Ladenstein R (1998) A protein disulfide oxidoreductase from the archaeon *Pyrococcus furiosus* contains two thioredoxin fold units. Nat Struct Biol 7:602–611

Zubay G (1993) To what extent do biochemical pathways mimic prebiotic pathways? Chemtracts Biochem Mol Biol 4:317–323