Published online 27 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1079–D1082***
doi:10.1093/nar/gks1217

LUCApedia: a database for the study of ancient life

Aaron David Goldman[1,*], Tess M. Bernhard[1], Egor Dolzhenko[2] and Laura F. Landweber[1,*]

[1]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, 08542, USA and [2]Department of Mathematics, University of South Florida, Tampa, FL, 33620, USA

Received August 15, 2012; Revised October 18, 2012; Accepted October 31, 2012

ABSTRACT

Organisms represented by the root of the universal evolutionary tree were most likely complex cells with a sophisticated protein translation system and a DNA genome encoding hundreds of genes. The growth of bioinformatics data from taxonomically diverse organisms has made it possible to infer the likely properties of early life in greater detail. Here we present LUCApedia, (http://eeb.princeton.edu/lucapedia), a unified framework for simultaneously evaluating multiple data sets related to the Last Universal Common Ancestor (LUCA) and its predecessors. This unification is achieved by mapping eleven such data sets onto UniProt, KEGG and BioCyc IDs. LUCApedia may be used to rapidly acquire evidence that a certain gene or set of genes is ancient, to examine the early evolution of metabolic pathways, or to test specific hypotheses related to ancient life by corroborating them against the rest of the database.

INTRODUCTION

All known life shares an ancestor at the root of the universal phylogenetic tree (1,2). This Last Universal Common Ancestor (LUCA) is a construct that may represent a single organism (1) or may represent populations of organisms capable of sharing large amounts of genetic information through horizontal gene transfer (3,4). Either way, organisms at the time of LUCA possessed many of the fundamental features present in modern organisms and likely exhibited a level of sophistication comparable with modern Bacteria or Archaea (5). Over the last decade, a number of studies have used bioinformatics databases to characterize the minimal set of features present in LUCA. The subjects of these surveys include gene families, protein architectures, protein domains and motifs and enzymatic functions.

Most of these studies identify a minimal set of hundreds of traits present in LUCA, which also most likely had a DNA genome, a cell membrane and a complete translation system. This complexity implies that a significant amount of evolutionary change must have taken place between the first life forms and LUCA. Multiple lines of evidence suggest that the earliest genetically encoded metabolism was produced by an RNA-only system in which RNA genes encoded ribozyme catalysts (6). Still more evidence suggests that protein translation arose from this RNA-only system (7,8) and that the DNA genome subsequently arose from the RNA-protein system, possibly just prior to the divergence of LUCA into the three domains of life (9,10). The capacity of an RNA-only system to support life has been studied by surveying the roles of naturally occurring ribozymes and synthesizing new ribozymes *in vitro* that have functions relevant to early life (11).

Non-genetically encoded catalysts, such as metal ions (12) or mineral surfaces (13,14) may also have played an important role in the production of large biomolecules both before and during the RNA-only era. Modern enzymes often use both organic and inorganic cofactors to impart catalysis. Some inorganic cofactors might reflect a pre-protein stage in which the reactions were catalysed by analogous ions and minerals (15). Similarly, nucleotide-derived cofactors may reflect a ribozyme precursor to modern protein enzymes that catalysed an analogous reaction.

Here we describe LUCApedia, which integrates these three lines of research into a unified framework provided by several well-established repositories of protein data. Users may query the database web server for a single protein in order to collect evidence of its antiquity from a broad range of studies. Downloadable database files may be used to evaluate the earliest components of modern pathways and to compare the antiquity of similar pathways to one another in an automated fashion. Users may also test the accuracy of previous studies and hypotheses implemented in the database by corroborating one of its data sets against the rest.

*To whom correspondence should be addressed. Tel: +1 609 258 6724; Fax: +1 609 258 1712; Email: adg@princeton.edu
Correspondence may also be addressed to Laura F. Landweber. Tel: +1 609 258 1947; Fax: +1 609 258 1682; Email: lfl@princeton.edu

OVERVIEW OF THE DATABASE

The central objective of the LUCApedia database is to integrate disparate data sets related to ancient life. Entry IDs from UniProt (16), BioCyc (17) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (18) supply the underlying framework. Uniprot was chosen because it captures a broad array of proteins and their relevant annotations. The Uniprot-based version of the database is appropriate for users wishing to examine the antiquity of a single protein, protein family or functional category. KEGG and BioCyc were used as alternative frameworks in order to facilitate studies of metabolic pathways. LUCApedia only stores Uniprot, KEGG and BioCyc identifiers, which are subsequently linked to early life data sets described below. Methods for mapping these data sets onto the three underlying databases are illustrated in Figure 1 and described in detail in the documentation available for download on the web server.

Six of the early life data sets are derived from studies in which features of LUCA were inferred by surveying a taxonomically broad range of organisms for universal traits:

'Harris *et al.* (19)'. The COG database (20) was used to identify 80 gene families that are universally distributed and have a similar phylogenetic pattern to ribosomal RNAs.

'Mirkin *et al.* (21)'. The COG database was used to generate models of gene gain and loss as well as horizontal gene transfer. The model corresponding to a gain penalty of one was used to reconstruct a LUCA gene set of 571 families.

'Delaye *et al.* (22)'. All-against-all BLAST (23) searches were performed in both directions between the genomes of 20 taxonomically diverse organisms. The resulting highly conserved genes were surveyed, and 115 common Pfam (24) domains were identified.

'Yang *et al.* (25)'. A phylogeny of 174 taxonomically diverse organisms was produced using a quantitative classification system based on protein domain content. The method identified 66 universal protein superfamilies (defined by SCOP) (26).

'Wang *et al.* (27)'. A phylogeny of protein folds (defined by SCOP) was generated using a quantitative classification system based on genomic surveys, and a branch on the resulting phylogeny was identified that represents the divergence of LUCA into the three taxonomic domains. The 165 deeper branching protein folds were predicted to have been present in LUCA.

'Srinivasan and Morowitz (28)'. Metabolic pathways (defined by KEGG) for three chemoautotrophic bacteria and one chemoautotrophic archaean were compared, and 286 common reactions were identified.

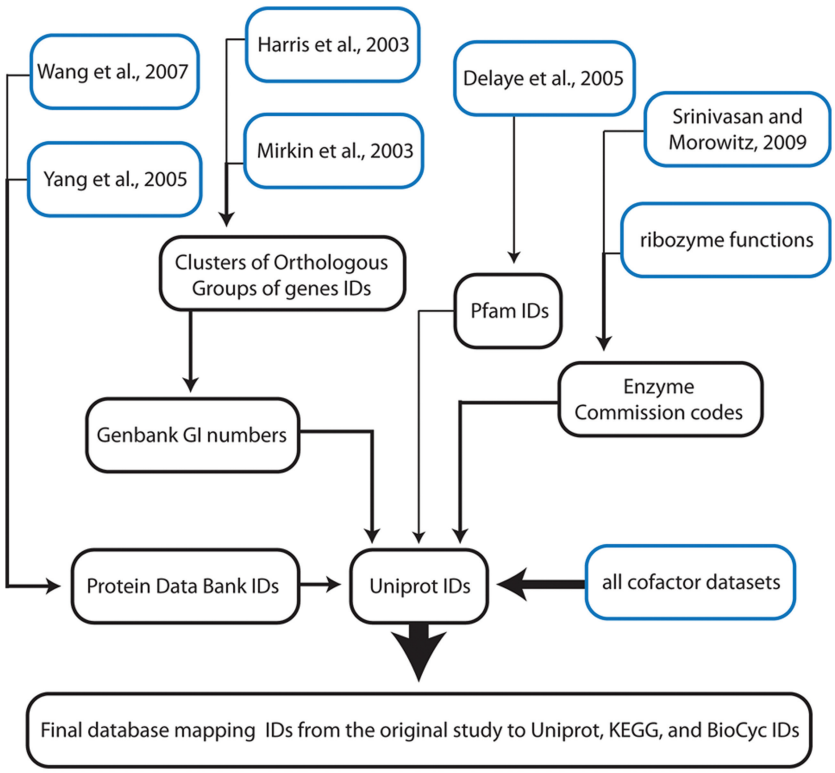


Figure 1. An overview of methods used to map the eleven early life data sets onto the underlying database framework of Uniprot, KEGG and Biocyc IDs. Initial data sets extracted either from previous studies or generated by the authors are highlighted in blue. Line thickness corresponds directly to the number of files at each methodological step. A complete description of ID mapping for each data set can be found in the LUCApedia documentation available for download from the web server.

Five other early life data sets are original to this database and pertain to published hypotheses regarding the origin and early evolution of life:

'Ribozyme functional analogs'. The RNA world hypothesis (6) predicts that most essential enzyme functions were performed by ribozymes prior to the establishment of protein translation. Thirty-three *in vivo* and *in vitro* ribozyme functions were collated through literature review.

'Nucleotide cofactor usage'. Enzyme functions requiring nucleotide-derived cofactors may reflect a transition from an RNA-only system to a system of RNA and protein enzymes (29). Cofactors derived from nucleotides were identified through literature review from the complete pool of cofactors used in Uniprot annotations.

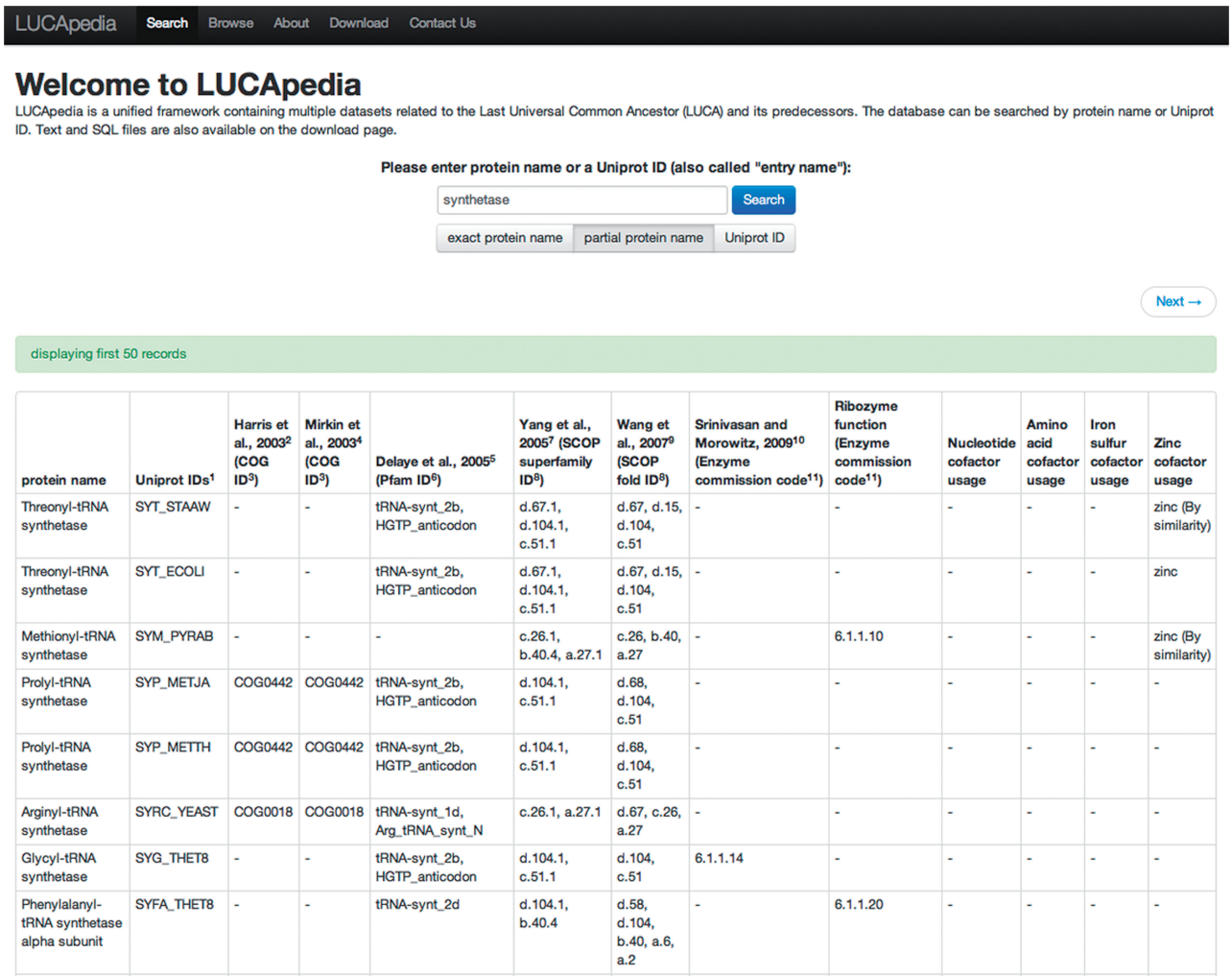
'Amino acid cofactor usage'. Amino acid cofactors may have played an early role in the transition out of an RNA-only system, initiating the transition to protein enzymes (30). Cofactors derived from amino acids were identified as above.

'Iron–sulfur cofactor usage'. Iron–sulfur cofactors have been proposed to reflect an important early role of iron–sulfur mineral surfaces in producing small molecules and facilitating polymerization (14). Cofactors containing iron–sulfur clusters were identified as above.

'Zinc cofactor usage'. Zinc cofactors have been proposed to reflect an important early role of zinc ions in nucleic acid chemistry and energy production (31). Zinc cofactors were identified as above.

THE LUCAPEDIA WEB SERVER

The web server is designed for users interested in quickly collecting evidence of deep ancestry for a small number of protein families. Proteins can be searched by name or Uniprot ID (Figure 2). A single name may correspond to multiple Uniprot IDs representing orthologs of the same protein in different species. A single Uniprot ID may also correspond to multiple names that are directly synonymous with one another. Proteins can also be found



protein name	Uniprot IDs[1]	Harris et al., 2003[2] (COG ID[3])	Mirkin et al., 2003[4] (COG ID[3])	Delaye et al., 2005[5] (Pfam ID[6])	Yang et al., 2005[7] (SCOP superfamily ID[8])	Wang et al., 2007[9] (SCOP fold ID[8])	Srinivasan and Morowitz, 2009[10] (Enzyme commission code[11])	Ribozyme function (Enzyme commission code[11])	Nucleotide cofactor usage	Amino acid cofactor usage	Iron sulfur cofactor usage	Zinc cofactor usage
Threonyl-tRNA synthetase	SYT_STAAW	-	-	tRNA-synt_2b, HGTP_anticodon	d.67.1, d.104.1, c.51.1	d.67, d.15, d.104, c.51	-	-	-	-	-	zinc (By similarity)
Threonyl-tRNA synthetase	SYT_ECOLI	-	-	tRNA-synt_2b, HGTP_anticodon	d.67.1, d.104.1, c.51.1	d.67, d.15, d.104, c.51	-	-	-	-	-	zinc
Methionyl-tRNA synthetase	SYM_PYRAB	-	-	-	c.26.1, b.40.4, a.27.1	c.26, b.40, a.27	-	6.1.1.10	-	-	-	zinc (By similarity)
Prolyl-tRNA synthetase	SYP_METJA	COG0442	COG0442	tRNA-synt_2b, HGTP_anticodon	d.104.1, c.51.1	d.68, d.104, c.51	-	-	-	-	-	-
Prolyl-tRNA synthetase	SYP_METTH	COG0442	COG0442	tRNA-synt_2b, HGTP_anticodon	d.104.1, c.51.1	d.68, d.104, c.51	-	-	-	-	-	-
Arginyl-tRNA synthetase	SYRC_YEAST	COG0018	COG0018	tRNA-synt_1d, Arg_tRNA_synt_N	c.26.1, a.27.1	d.67, c.26, a.27	-	-	-	-	-	-
Glycyl-tRNA synthetase	SYG_THET8	-	-	tRNA-synt_2b, HGTP_anticodon	d.104.1, c.51.1	d.104, c.51	6.1.1.14	-	-	-	-	-
Phenylalanyl-tRNA synthetase alpha subunit	SYFA_THET8	-	-	tRNA-synt_2d	d.104.1, b.40.4	d.58, d.104, b.40, a.6, a.2	-	6.1.1.20	-	-	-	-

Figure 2. A screen shot of the LUCApedia web server search function. Protein names can be entered into the search field and the search will return all corresponding Uniprot IDs along with evidence of their relevance to ancient life. Searches may be conducted for either exact or partial protein names. If a name search does not return any results, Uniprot IDs may also be directly queried.

by way of the 'Browse' page, where they are listed by name in alphabetical order. The 'About' page features an abridged documentation explaining the core organization of the database and a description of each data set. More advanced users interested in performing detailed analysis may use the 'Download' page to acquire flat text files of each data set mapped to Uniprot, KEGG and Biocyc IDs, as well as the complete database documentation and MySQL dumps of the tables used to serve the database and to implement the web server's search function.

CONCLUSION

Inferring likely characteristics of early life forms with any statistical confidence prior to and during the stage of the last universal common ancestor has only recently become a tractable problem. Even so, it is usually not possible to prove the conclusions of studies in this discipline. In lieu of definitive proof, an understanding of ancient life can be built from the consensus of diverse and independent methods. LUCApedia creates an unprecedented ability to corroborate the results from independent studies, to evaluate early life hypotheses, and to direct future experiments toward understudied areas, all in an objective, quantitative and systematic manner.

ACKNOWLEDGEMENTS

The authors thank the members of the Landweber lab for useful comments and for help testing the web server. The authors also thank John Baross and Ram Samudrala for early discussions of the LUCApedia concept.

FUNDING

NASA postdoctoral fellowship (to A.D.G.); National Science Foundation (NSF) [0900544 to L.F.L]. Funding for open access charge: NSF [0900544].

Conflict of interest statement. None declared.

REFERENCES

1. Woese,C. (1998) The universal ancestor. *Proc. Natl Acad. Sci. USA*, **95**, 6854–6859.
2. Theobald,D. (2010) A formal test of the theory of universal common ancestry. *Nature*, **465**, 219–222.
3. Doolittle,W.F. (2000) The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.*, **10**, 355–358.
4. Zhaxybayeva,O. and Gogarten,J.P. (2004) Cladogenesis, coalescence and the evolution of the three domains of life. *Trends Genet.*, **20**, 182–187.
5. Becerra,A., Delaye,L., Islas,S. and Lazcano,A. (2007) The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annu. Rev. Ecol. Evol. Syst.*, **38**, 361–379.
6. Gilbert,W. (1986) The RNA world. *Nature*, **319**, 618.
7. Freeland,S.J., Knight,R.D. and Landweber,L.F. (1999) Do proteins predate DNA? *Science*, **286**, 690–692.
8. Goldman,A.D., Samudrala,R. and Baross,J.A. (2010) The evolution and functional repertoire of translation proteins following the origin of life. *Biol. Direct.*, **5**, 15.
9. Forterre,P. (2002) The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.*, **5**, 525–532.
10. Goldman,A.D. and Landweber,L.F. (2012) Oxytricha as a modern analog of ancient genome evolution. *Trends Genet.*, **28**, 382–388.
11. Landweber,L.F., Simon,P.J. and Wagner,T.A. (1998) Ribozyme engineering and early evolution. *BioScience*, **48**, 94–103.
12. Mulkidjanian,A.Y. (2009) On the origin of life in the Zinc world: 1. Photosynthesizing, porous edifices built of hydrothermally precipitated zinc sulfide as cradles of life on Earth. *Biol. Direct.*, **4**, 26.
13. Ferris,J.P., Hill,A.R. Jr, Liu,R. and Orgel,L.E. (1996) Synthesis of long prebiotic oligomers on mineral surfaces. *Nature*, **381**, 59–61.
14. Wächtershäuser,G. (1990) Evolution of the first metabolic cycles. *Proc. Natl Acad. Sci. USA*, **87**, 200–204.
15. White,H.B. (1976) Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.*, **7**, 101–104.
16. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
17. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al.* (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **40**, D742–D753.
18. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res.*, **40**, D109–D114.
19. 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.
20. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
21. Mirkin,B.G., Fenner,T.I., Galperin,M.Y. and Koonin,E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, **3**, 2.
22. Delaye,L., Becerra,A. and Lazcano,A. (2005) The last common ancestor: what's in a name? *Orig. Life Evol. Biosph.*, **35**, 537–554.
23. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
24. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
25. Yang,S., Doolittle,R.F. and Bourne,P.E. (2005) Phylogeny determined by protein domain content. *Proc. Natl Acad. Sci. USA*, **102**, 373–378.
26. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2007) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
27. Wang,M., Yafremava,L.S., Caetano-Anollés,D., Mittenthal,J.E. and Caetano-Anollés,G. (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res.*, **17**, 1572–1585.
28. Srinvasan,V. and Morowitz,H.J. (2009) The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. *Biol. Bull.*, **216**, 126–130.
29. Kyrpides,N.C. and Ouzounis,C.A. (1995) Nucleic acid-binding metabolic enzymes: living fossils of stereochemical interactions? *J. Mol. Evol.*, **40**, 564–569.
30. Szathmáry,E. and Smith,J.M. (1995) The major evolutionary transitions. *Nature*, **374**, 227–232.
31. Mulkidjanian,A.Y. and Galperin,M.Y. (2009) On the origin of life in the Zinc world. 2. Validation of the hypothesis on the photosynthesizing zinc sulfide edifices as cradles of life on Earth. *Biol. Direct.*, **4**, 27.