A Guide to LUCApedia

Why use LUCApedia

Many topics within the study of the origin and early evolution of life are amenable to computational research strategies. LUCApedia was established to aggregate data from studies and hypotheses about the Last Universal Common Ancestor (LUCA) and its predecessors. These data are unified onto a common framework so that they can be compared and corroborated. The most recent update of the LUCApedia database (July 2025) is available via the web server https://lucapedia.org/, where it can be searched, browsed, and downloaded.

What is in LUCApedia

Underlying database framework

The LUCApedia database consists of seventeen datasets that can be categorized as representing four distinct stages of early evolution: Prebiotic Chemistry, the RNA World, LUCA Protein Domains, and LUCA Protein Families. Each dataset within the four categories represents either the results of a single empirical study or a previously published hypothesis about the origin or early evolution of life. Each of these studies or hypotheses is mapped onto UniProt database accessions[1] (representing single proteins). A separate version of the LUCApedia database is created by aggregating these individual proteins into protein families represented by the EggNOG database[2]. Either of these versions of the database can be searched, browsed, or downloaded from the LUCApedia web server.

Descriptions of individual datasets

Prebiotic Chemistry Datasets

Iron-Sulfur Cluster (7296 Uniprot accessions, 351 EggNOG accessions): Iron-sulfur cofactors have been previously proposed to reflect the potential role of iron-sulfur mineral surfaces as a possible setting for the origin of life and an important catalyst for very early life forms[3,4]. This dataset is composed of all proteins that use an iron-sulfur cofactor.

Zinc Cofactor (13948 Uniprot accessions, 586 EggNOG accessions): Previous publications have proposed that zinc and zinc-sulfide catalysts may have played an important role in early nucleic acid chemistry and energy metabolism[5] and that the use of zinc cofactors in extant proteins reflects this early role in prebiotic chemistry[6]. This dataset is composed of all proteins that use an iron-sulfur cofactor.

Goldford (5249 Uniprot accessions, 280 EggNOG accessions): A previous study by Goldford and colleagues[7] generated a phosphate-free metabolic network by starting with simple prebiotically available compounds and using a network expansion

algorithm that assumed the evolution of catalysts or enzymes that could convert available metabolites into new metabolites. The dataset is composed of all proteins with enzymatic functions matching those present in this hypothetical metabolic network.

Goldford-thio (15723 Uniprot accessions, 699 EggNOG accessions): The minimal phosphate-free metabolic network developed by Goldford and colleagues assumed that only energetically favorable reactions could take place without the presence of ATP as an energy currency. The Goldford-thio dataset is composed of a larger network in which prebiotically-plausible thioesters are used as an energy currency that predated ATP. The dataset is composed of all proteins with enzymatic functions matching those present in this expanded hypothetical metabolic network.

RNA World Datasets

Ribozyme function analogs (3730 Uniprot accessions, 100 EggNOG accessions): The RNA World hypothesis[8] has motivated synthetic biochemists to develop an increasingly broad array of ribozymes (i.e., catalytic RNAs) that can perform enzymatic functions now performed by proteins. This dataset is composed of proteins that perform a function that can also be performed by a natural or artificial ribozyme. The dataset was created by reviewing ribozyme literature, assigning Enzyme Commission (EC) numbers[9] to the ribozyme function, and identifying protein enzymes associated with the same EC number.

Nucleotide Cofactor (5059 Uniprot accessions, 294 EggNOG accessions): Many of the most important coenzymes in metabolism, e.g., ATP, NADH, Coenzyme A, and S-Adenosyl Methionine, are composed of or derived from nucleotides[10]. This observation led to the hypothesis that these coenzymes are remnants of earlier RNA World catalysts. This dataset is composed of all proteins that use these coenzymes.

Amino Acid Cofactors (894 Uniprot accessions, 109 EggNOG accessions): The use of amino acids as ribozyme cofactors was previously proposed as a preadaptation that facilitated the origin of the genetic code. Specifically, sequences of nucleotides that bound amino acid cofactors could have evolved into the codons and anticodons that now comprise the genetic code[11]. This dataset is composed of all proteins that use cofactors derived from amino acids.

RNA Aptamer (10 Uniprot accessions, 10 EggNOG accessions): Blanco and colleagues[12] identified proteins with known structures that bound artificial RNA aptamers and used these data to better characterize RNA-protein binding and the types of amino acids that are typically found in such interactions. This dataset is composed of the proteins used in this study.

LUCA protein domains

Yang (1936 Uniprot accessions, 1141 EggNOG accessions): Yang and colleagues[13] developed a species phylogeny based on patterns of the presence or absence of protein

folds as catalogued in the SCOP database. They identified 66 SCOP superfamilies that were present in all proteomes, which they consider to be the set of protein folds inherited from the proteome of the LUCA.

Wang (6318 Uniprot accessions, 3304 EggNOG accessions): Wang and colleagues[14] created a phylogeny of protein folds based on accessions in the SCOP database[15]. Phylogenetic relationships were determined based on both the taxonomic breadth of each fold and its frequency within individual proteomes. This phylogeny was used to identify 165 SCOP folds that emerged prior to the LUCA.

Delaye (2375 Uniprot accessions, 719 EggNOG accessions): Delaye and colleagues[16] identified protein domains and motifs, as defined by the Pfam database[17], that were found to be universal across bacterial and archaeal proteomes. The study identified 115 such universal Pfam domains and motifs.

Ranea (10343 Uniprot accessions, 3861 EggNOG accessions): Ranea and colleagues[18] used structural comparisons of proteins across bacteria and archaea to identify universal ancestral domain architectures as defined by the CATH database[19]. The study identified 114 CATH domain superfamilies to have been present in the LUCA proteome.

LUCA protein families

Harris (29085 Uniprot accessions, 104 EggNOG accessions): Harris and colleagues[20] surveyed clusters of proteins defined by the now defunct COG database[21] to identify protein families that were both universal and vertically inherited. Vertical inheritance was determined by comparison to an rRNA-based species tree. The study identified 80 such COGs that were attributed to the LUCA.

Mirkin (97814 Uniprot accessions, 809 EggNOG accessions): Mirkin and colleagues[22] performed an analysis similar to that of Harris et al., but included a model for gene loss, which led to a more permissive set of COGs being attributed to the LUCA. While several LUCA proteome models were reported in the study based on different gene gain/loss parameters, this dataset is based on the set of 572 COGs that was determined in the study to approximate a viable organism.

Srinivasan (38262 Uniprot accessions, 1089 EggNOG accessions): Srinivasan and Morowitz[23] compared enzymatic reactions across several bacterial and archaeal metabolic networks cataloged in the KEGG database[24] to identify 286 shared reactions, deemed by this study to be universal metabolic reactions.

Weiss (6318 UniProt accessions, 3304 EggNOG accessions): Weiss and colleagues[25] identified 355 protein families predicted to have been present in the proteome of the LUCA. Protein families were defined based on clusters of homologous proteins represented in the EggNOG database. LUCA ancestry was determined based on whether members of the protein family were found in multiple phyla of both archaea

and bacteria, and also had tree topologies in which the bacterial clades and archaeal clades were respectively monophyletic.

Moody (399 Uniprot accessions, 354 EggNOG accessions): Moody and colleagues[26] identified 399 protein families predicted to have been present in the proteome of the LUCA. Protein families were defined by the KEGG Orthology (KO) database[27], and LUCA ancestry was determined by gene-tree species tree reconciliation based on the Amalgamated Likelihood Estimation algorithm67.

Using the Web Server

The web server has both a search function and a browse function. The search function allows users to search for complete or partial protein names, UniProt IDs, EggNOG IDs, or Gene Ontology IDs[28]. Either the UniProt-based version of the database or the EggNOG-based version of the database can be searched separately. While a default search shows results for all seventeen datasets, these datasets can be individually selected either on the search page or the search results page. The results of a search are shown on a separate page and can be downloaded as a TSV file.

In addition to searching the LUCApedia database, users can also browse the database by the associated function of the proteins or protein families. These functions are based on the Gene Ontology IDs linked to each LUCApedia entry. The Gene Ontology Database characterizes proteins with respect to their molecular function, cellular location, and the biological process that they participate in. Gene Ontology terms describe these features of proteins at different levels of specificity and terms that are nested within other terms. The LUCApedia web server allows users to browse Gene Ontology terms from the most general level to the most specific level in a manner similar to the AmiGO web server[29]. When a user clicks on a GO term, they are presented with LUCApedia results associated with that GO term.

The web server also features a Download page where the UniProt-based version of the database and the EggNOG-based version of the database can be downloaded as TSV files. The list of ribozymes, their associated EC numbers, and the studies that report their functions, comprising the information that was used to make the Ribozyme Function dataset, can also be downloaded from this page. The legacy version of the LUCApedia database (2013) can also be downloaded from this page.

LUCApedia Citation

When using the LUCApedia database or web server, please cite…

Goldman AD, Bernhard TM, Dolzhenko E, Landweber LF (2013) LUCApedia: a database for the study of ancient life. Nucleic Acids Res. 41(Database issue):D1079-82. doi: 10.1093/nar/gks1217.

References

1. UniProt, C. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res *51*, D523-D531. 10.1093/nar/gkac1052.
2. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res *47*, D309-D314. 10.1093/nar/gky1085.
3. Huber, C., and Wachtershauser, G. (1997). Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. Science *276*, 245-247. 10.1126/science.276.5310.245.
4. Wachtershauser, G. (1988). Before enzymes and templates: theory of surface metabolism. Microbiol Rev *52*, 452-484. 10.1128/mr.52.4.452-484.1988.
5. 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. 10.1186/1745-6150-4-26.
6. 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. 10.1186/1745-6150-4-27.
7. Goldford, J.E., Hartman, H., Smith, T.F., and Segre, D. (2017). Remnants of an Ancient Metabolism without Phosphate. Cell *168*, 1126-1134 e1129. 10.1016/j.cell.2017.02.001.
8. Gilbert, W. (1986). Origin of life: The RNA world. Nature *319*, 618-618. 10.1038/319618a0.
9. Webb, E.C. (1992). Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology and the Nomenclature and Classification of Enzymes (Elsevier Science).
10. White, H.B., 3rd (1976). Coenzymes as fossils of an earlier metabolic state. J Mol Evol *7*, 101-104. 10.1007/BF01732468.
11. Szathmary, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world. Trends Genet *15*, 223-229. 10.1016/s0168-9525(99)01730-8.
12. Blanco, C., Bayas, M., Yan, F., and Chen, I.A. (2018). Analysis of Evolutionarily Independent Protein-RNA Complexes Yields a Criterion to Evaluate the Relevance of Prebiotic Scenarios. Curr Biol *28*, 526-537 e525. 10.1016/j.cub.2018.01.014.
13. Yang, S., Doolittle, R.F., and Bourne, P.E. (2005). Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A *102*, 373-378. 10.1073/pnas.0408810102.
14. Wang, M., Yafremava, L.S., Caetano-Anolles, D., Mittenthal, J.E., and Caetano-Anolles, G. (2007). Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res *17*, 1572-1585. 10.1101/gr.6454307.
15. Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540. 10.1006/jmbi.1995.0159.

16. Delaye, L., Becerra, A., and Lazcano, A. (2005). The last common ancestor: what's in a name? Orig Life Evol Biosph *35*, 537-554. 10.1007/s11084-005-5760-3.
17. 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*, D290-301. 10.1093/nar/gkr1065.
18. Ranea, J.A., Sillero, A., Thornton, J.M., and Orengo, C.A. (2006). Protein superfamily evolution and the last universal common ancestor (LUCA). J Mol Evol *63*, 513-525. 10.1007/s00239-005-0289-7.
19. Knudsen, M., and Wiuf, C. (2010). The CATH database. Hum Genomics *4*, 207-212. 10.1186/1479-7364-4-3-207.
20. 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. 10.1101/gr.652803.
21. Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631-637. 10.1126/science.278.5338.631.
22. 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. 10.1186/1471-2148-3-2.
23. Srinivasan, V., and Morowitz, H.J. (2009). The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol Bull *216*, 126-130. 10.1086/BBLv216n2p126.
24. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30. 10.1093/nar/28.1.27.
25. Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. (2016). The physiology and habitat of the last universal common ancestor. Nat Microbiol *1*, 16116. 10.1038/nmicrobiol.2016.116.
26. Moody, E.R.R., Alvarez-Carretero, S., Mahendrarajah, T.A., Clark, J.W., Betts, H.C., Dombrowski, N., Szantho, L.L., Boyle, R.A., Daines, S., Chen, X., et al. (2024). The nature of the last universal common ancestor and its impact on the early Earth system. Nat Ecol Evol. 10.1038/s41559-024-02461-1.
27. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res *44*, D457-462. 10.1093/nar/gkv1070.
28. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29. 10.1038/75556.
29. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., Ami, G.O.H., and Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288-289. 10.1093/bioinformatics/btn615.