LUCApedia v1.0

Aaron David Goldman, adg@princeton.edu
August 12th, 2012

Contents:

I. Why use LUCApedia

Thanks to the growth of genomics, proteomics, and metabolomics, it is possible to investigate properties of the Last Universal Common Ancestor (LUCA) and its predecessors in detail. LUCApedia was established to aggregate and unify the results of studies aimed at describing early life through a variety of bioinformatics approaches and pair them with a number of enzymological characteristics predicted in previous studies to reflect catalysts important in the early evolution of life. Users may query the webserver for individual proteins to rapidly identify evidence of deep ancestry. Advanced users may download the database as a series of flat files and use it to discover trends in early enzymatic and metabolic evolution and to test hypotheses related to early life.

II. What is in LUCApedia

Underlying database framework

Datasets corresponding to studies predicting characteristics of LUCA consist of different data types: Protein structures, protein domain folds, clusters of orthologous genes, etc. In order to use these data in concert, they must be organized into a common framework. We achieve this unification by mapping these datasets to Uniprot IDs[1] (also called "entry names"), KEGG IDs[2], and Biocyc IDs[3]. These three implementations are separate and it is up to the user whether to choose one for his or her study or to compare the results of all three to achieve a greater level of confidence in his or her study. Methods of mapping each of these datasets into Uniprot, KEGG, and Biocyc IDs are described in Section V.

Early life datasets

Dataset of ribozyme functions — 32 EC codes
The RNA world hypothesis predicts that the original genetic system involved RNA genes encoding RNA enzymes (also called ribozymes)[4]. This dataset represents enzymatic functions (by Enzyme Commission[5] code) that have been observed *in vivo* or synthesized *in vitro*.

Dataset from Harris et al., 2003[6] — 80 COGs
This study attempted to identify the minimal gene set of LUCA by identifying Clusters of Orthologous Groups of genes[7] (COGs) that were present in every genome available at the time.

Dataset from Mirkin et al., 2003[8] — 571 COGs
This study attempted to use a less stringent requirement for the gene set of LUCA by adding COGs, which appear to be ancient, but do not appear in every genome because they have been replaced by functional analogs through the process of non-orthologous gene displacement. LUCApedia 1.0 uses data from this study corresponding to a gain penalty of 1.0.

Dataset from Delaye et al., 2005[9] — 115 Pfam motifs
This study attempted to model the functional repertoire of LUCA through all-against-all BLAST[10] searches of twenty taxonomically diverse organisms. The results are a series of Pfam[11] motifs that are predicted to have been present in LUCA's proteome.

Dataset from Yang et al., 2005[12] — 66 SCOP superfamilies
This study attempted to identify the minimal proteome of LUCA by creating a phylogeny of 174 taxonomically diverse organisms using a quantitative classification system based on protein domain content. This method identified universal domains, defined at the level of SCOP[13] superfamilies.

Dataset from Wang et al., 2007[14] — 165 SCOP folds
This study attempted to identify the minimal proteome of LUCA by creating a phylogeny of 185 taxonomically diverse organisms using a quantitative classification system based on genomic surveys of protein domain content. A branch of this phylogeny was identified as the point at which LUCA diverged into the three domains of life. All terminal nodes deeper than this branch are considered to represent domains present in LUCA.

Dataset from Srinivasan and Morowitz, 2009[15] — 286 EC codes
This study attempted to identify the set of metabolic reactions present in LUCA. Complete metabolomes of five autotrophic bacteria and one autotrophic archaean were compared and reactant-product pairs present in all six organismal datasets were predicted to have been present in LUCA.

Nucleotide cofactor usage
Enzyme functions that employ nucleotide-derived cofactors are predicted to reflect a prior state in which the same reaction was catalyzed by ribozymes[16]. Cofactors derived from nucleotides were identified through literature review from the complete pool of cofactors used in Uniprot annotations.

Amino acid cofactor usage
Enzyme functions that employ amino acid-derived cofactors are predicted to reflect the transition from ribozymes to protein enzymes as the primary catalytic molecule of life[16]. Cofactors derived from amino acid were identified through literature review from the complete pool of cofactors used in Uniprot annotations.

Iron-sulfur cofactor usage
Enzyme functions that employ iron-sulfur cofactors are predicted to reflect protobiological chemistry taking place on the surface of pyrite minerals[17]. Iron-sulfur cofactors were identified through literature review from the complete pool of cofactors used in Uniprot annotations.

Zinc cofactor usage
Enzyme functions that employ zinc cofactors are predicted to reflect protobiological chemistry catalyzed by zinc ions[18]. Zinc cofactors were identified through literature review from the complete pool of cofactors used in Uniprot annotations.

III. Using the webserver

The LUCApedia webserver implements the Uniprot version of the database. A single Uniprot ID often corresponds to multiple synonymous protein names and a single protein name often corresponds to multiple Uniprot IDs. If a protein of interest cannot be found by protein name, the corresponding Uniprot IDs may be searched directly. Users may also browse the database by alphabetical order of protein names.

IV. Format of flat files

The MySQL dump files representing the Uniprot implementation of the database and the name lookup table used to search for database entries can be downloaded from the webserver. Users may also download text files each representing each of the individual datasets relevant to early life that were described in Section II. These data are mapped onto Uniprot, KEGG, and Biocyc IDs. The common format of these text files is...

> dataset entry
Uniprot ID 1 KEGG ID 1 Biocyc ID 1
Uniprot ID 2 KEGG ID 2 Biocyc ID 2

... and so on.

V. Methods of implementation

Underlying database framework

Each of these datasets were first mapped onto the Uniprot IDs, then extended to KEGG and Biocyc by way of the Uniprot ID mapping file available for download on the Uniprot webserver.

Early life datasets

Dataset of ribozyme functions
Ribozyme functions were collected through an exhaustive literature review and converted to EC codes. Uniprot IDs corresponding to each EC code were collected via the Uniprot webserver.

Dataset from Harris et al., 2003
COGs identified by this study were converted to GI numbers[19] via the COG database downloadable file < myva=gb> available on the COG webserver. GI numbers were converted to Uniprot IDs via the Uniprot ID mapping file available for download from the Uniprot webserver.

Dataset from Mirkin et al., 2003
COGs identified by this study were converted to GI numbers via the COG database downloadable file <myva=gb> available on the COG webserver. GI numbers were converted to Uniprot IDs via the Uniprot ID mapping file available for download from the Uniprot webserver.

Dataset from Delaye et al., 2005
Pfam codes identified by this study were converted to Uniprot codes using the file, <pfam-A.full>, available for download from the Pfam webserver.
Dataset from Yang et al., 2005
SCOP superfamilies were converted to PDB IDs using the file <dir.cla.scop.txt_1.75> downloaded from the SCOP database. PDB IDs were converted to Uniprot IDs by the Uniprot ID mapping file available for download from the Uniprot webserver.

Dataset from Wang et al., 2007
SCOP superfamilies were converted to PDB IDs using the file <dir.cla.scop.txt_1.75> downloaded from the SCOP database. PDB IDs were then converted to Uniprot IDs by the Uniprot ID mapping file available for download from the Uniprot webserver.

Dataset from Srinivasan and Morowitz, 2009
Reactant-product pairs identified by this study were manually converted to EC codes. Uniprot IDs corresponding to each EC code were collected via the Uniprot webserver.

Nucleotide cofactor usage
Cofactor usage was identified using the Uniprot annotation file, <uniprot_sprot.dat>, available for download from the Uniprot webserver.

Amino acid cofactor usage
Cofactor usage was identified using the Uniprot annotation file, <uniprot_sprot.dat>, available for download from the Uniprot webserver.

Iron-sulfur cofactor usage
Cofactor usage was identified using the Uniprot annotation file, <uniprot_sprot.dat>, available for download from the Uniprot webserver.

Zinc cofactor usage
Cofactor usage was identified using the Uniprot annotation file, <uniprot_sprot.dat>, available for download from the Uniprot webserver.

VI. Future of LUCApedia

As research on early life continues, we are dedicated to updating the database by implementing more datasets representing studies related to LUCA, regularly updating our ribozyme function dataset, and implementing new datasets related to modern enzymes that will allow users to evaluate their results. If you have questions or comments, please contact Aaron Goldman at adg@princeton.edu.

VII. References

1. The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt) Nucleic Acids Res. 40: D71-D75

2. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 28:27-30

3. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, *et al*. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res, 38:D473-D479

4. Gilbert W (1986) The RNA world. Nature, 319:618

5. Webb, Edwin C. (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press

6. Harris JK, Kelley ST, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13:407

7. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science, 278:631-637

8. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (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

9. Delaye L, Becerra A, Lazcano A (2005) The last common ancestor: what's in a name? Orig Life Evol Biosph, 35:537-554

10. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol, 215:403-410

11. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res, 32:D138-D141

12. Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content, Proc Nat Acad Sci U S A, 102:373-378

13. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol, 247:536-540

14. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res, 17:1572-1585

15. Srinvasan V and Morowitz HJ(2009) The canonical network of autotrophic intermediary metabolism: minimal metabolome of a reductive chemoautotroph. Biol Bull 216:126-130

16. White HB (1976) Coenzymes as fossils of an earlier metabolic state. J Mol Evol 7:101–104

17. Wächtershäuser G (1990) Evolution of the first metabolic cycles. Proceedings of the National Academy of Sciences USA, 87:200-204

18. Mulkidjanian AY, Galperin MY (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. BMC Biol Direct, 4:27

19. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2003) Genbank. Nucleic Acids Res, 31:23-27