LukProt: A database of eukaryotic predicted proteins designed for investigations of animal origins

Łukasz F. Sobala*[1]

[1]Laboratory of Glycobiology, Hirszfeld Institute of Immunology and Experimental Therapy, PAS, Wrocław, Poland

*Corresponding author

Author for Correspondence: Łukasz F. Sobala, Laboratory of Glycobiology, Hirszfeld Institute of Immunology and Experimental Therapy, PAS, Weigla 12, 53-114 Wrocław, Poland, +48 71 337 11 72 ext. 290, lukasz.sobala@hirszfeld.pl

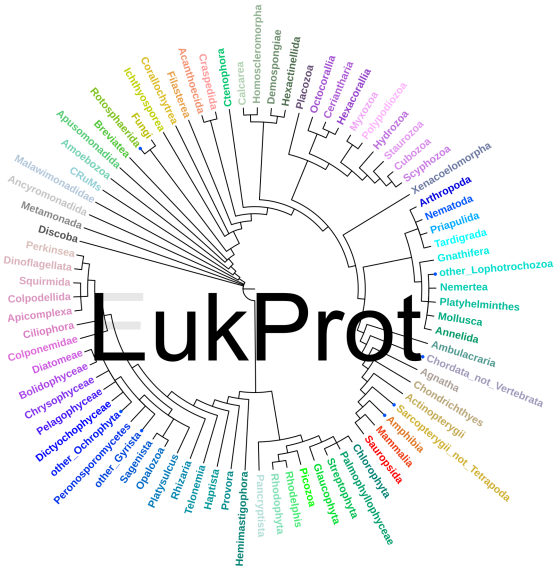
Abstract

The origins and early evolution of animals is a subject with many outstanding questions. One problem faced by researchers trying to answer them is the absence of a comprehensive database with sequences from non-bilaterians. Publicly available data is plentiful but scattered and often not associated with proper metadata. A new database presented in this paper, LukProt, is an attempt at solving this issue. The database contains protein sequences obtained mostly from genomic, transcriptomic and metagenomic studies and is an extension of EukProt (Richter et al., 2022, *Peer Community Journal*, **2**, e56). LukProt adopts

the EukProt naming conventions and includes data from 216 additional animals. The database is associated with a taxonomic grouping (taxogroup) scheme suitable for studying early animal evolution. Minor updates to the database will contain species additions or metadata corrections and major updates will synchronize LukProt to each new version of EukProt and releases are permanently stored on Zenodo. A BLAST server to search the database is available at https://lukprot.hirszfeld.pl/. Users are invited to participate in maintaining and correcting LukProt. As it can be searched without downloading locally, the database can be a convenient resource not only for evolutionary biologists, but for the broader scientific community as well.

Keywords: database, evolution, proteome

Graphical abstract



Significance statement

LukProt is a sequence database aiming to accelerate the research on the evolution of animals by cutting the time-consuming step of assembling sequences from disparate sources. Non-bilaterians are currently not well covered by general purpose databases, despite plentiful, public sequencing data. These data were integrated into a consistently curated database, presented here. It can be downloaded and used locally or used via a public BLAST search server. A clear taxonomic framework is also introduced, as well as scripts to aid local data analyses. LukProt will be publicly available on Zenodo, kept up to date and synchronized with each new version of its parent database, EukProt.

Introduction

Animals (Metazoa) are a diverse group of organisms whose deep evolutionary history has been a subject of a long debate (Ruiz-Trillo et al., 2023). Many early branching lineages of animals are extinct, for example much of the Ediacaran fauna is not similar to any living animals. Animal tissues were not often amenable to fossilization pre-Ediacaran (Murdock & Donoghue, 2011). Recent advancements in phylogenomics helped resolve fundamental questions in early history of Metazoa, for example the identity of the earliest extant animal branch (Dunn et al., 2008; Schultz et al., 2023) and the timing of the two whole genome duplications which occurred in the ancestors of all living jawed vertebrates (Simakov et al., 2022; Marlétaz et al., 2024). These developments were enabled by more advanced bioinformatics tools, broader access to processing power and broader availability of sequencing data.

As animals are members of Eukaryota, knowledge about eukaryotic evolution can provide context for evolution of animals. In the last decade, the topology of the eukarotic tree of life has been undergoing revisions as major as the animal tree. The discovery and characterization of Asgard archaea (Zaremba-Niedzwiedzka et al., 2017; Eme et al., 2023) helped place the trunk of the tree. Due to the recent sequencing boom, new supergroups of protists are still being defined (Tikhonenkov et al., 2022). A comprehensive database which collects predicted proteomes of eukaryotes, EukProt, plays a role in these developments (Richter et al., 2022). The broad scope and careful curation of EukProt makes it useful for many applications, but the database compromises on the number of included species from mostly multicellular clades (plants, fungi or animals).

Investigating animal evolution requires broad taxon sampling of early diverging animal clades, which EukProt, NCBI or UniProt databases do not provide. One attempt at a wide animal taxon sampling, the Animal Proteome Database – AniProtDB (Barreira et al., 2021), includes only 100 species. Another large database of non-vertebrate animal sequences focuses on protostomes (Fernández et al., 2022). A more comprehensive database, which includes data from early diverging lineages, could serve as a tool for researchers to study the origins and evolution of uniquely animal characters, such as obligatory multicellularity or neurons.

LukProt, created to fill this gap, is a synthesis of the latest available EukProt, AniProtDB and many additional collected sequences. The database focuses on holozoans (a clade of opisthokonts sister to Nucletmycea) and animal species from the clades: Ctenophora, Porifera, Placozoa and Cnidaria. Bilaterian sampling is also increased. Conventions established by EukProt are followed: each dataset (proteome) has a unique identifier and each protein has a persistent ID composed of the dataset identifier, the species/strain name and protein number. A detailed characterization of LukProt, as well as EukProt data or metadata modifications made for LukProt, is presented below.

Results and Discussion

Purpose of LukProt

The purpose of the database release is to create a resource which one could consult on whether a given domain/protein/protein family exists within a clade of eukaryotes, with a focus on animals and their closest relatives: choanoflagellates, filozoans, ichthyosporeans and corallochytreans. To this end, individual dataset quality is of lower importance. No decontamination, or contamination-based dataset exclusion was performed. It is assumed that phylogeny reconstructions will be the guide when drawing conclusions from database searches. Phylogeny-guided analysis of LukProt, using a suitable sister database, can also be the basis for studies of lateral gene transfers. LukProt is documented with less experienced users in mind.

Characterization of the LukProt database

The current version of the database (v1.5.1.rev1) contains 1281 datasets: 990 from EukProt, 41 from AniProtDB and 250 newly added. These newly added datasets were collected from various studies and publicly available repositories, or generated *in silico* from public data. Details of each dataset generation can be found in the metadata. Overall database statistics are given in **Supplementary Table 1**. Coverage of the early branching animal clades, especially ctenophores, sponges and cnidarians, is improved in comparison to EukProt. For example, 26% of the 23 currently known, potential placozoan species (Tessler et al., 2022) is included (6 datasets).

A consensus taxonomy

The database incorporates a consensus taxonomy of eukaryotes, largely based on UniEuk (Berney et al., 2017). The clades are explained in Supplementary Information and the full tree can be found in Figure 1. Each internal node was assigned a unique name, following UniEuk or recent literature where possible. Unnamed clades were assigned provisional names (Figure 1). The LukProt taxonomic framework will be adjusted to UniEuk updates and future data. It encompasses 81 groups of organisms, mostly monophyletic. The metadata also contains 59 groupings based on the consensus taxonomy, selected to aid targeted studies (e.g. Holozoa excluding Metazoa).

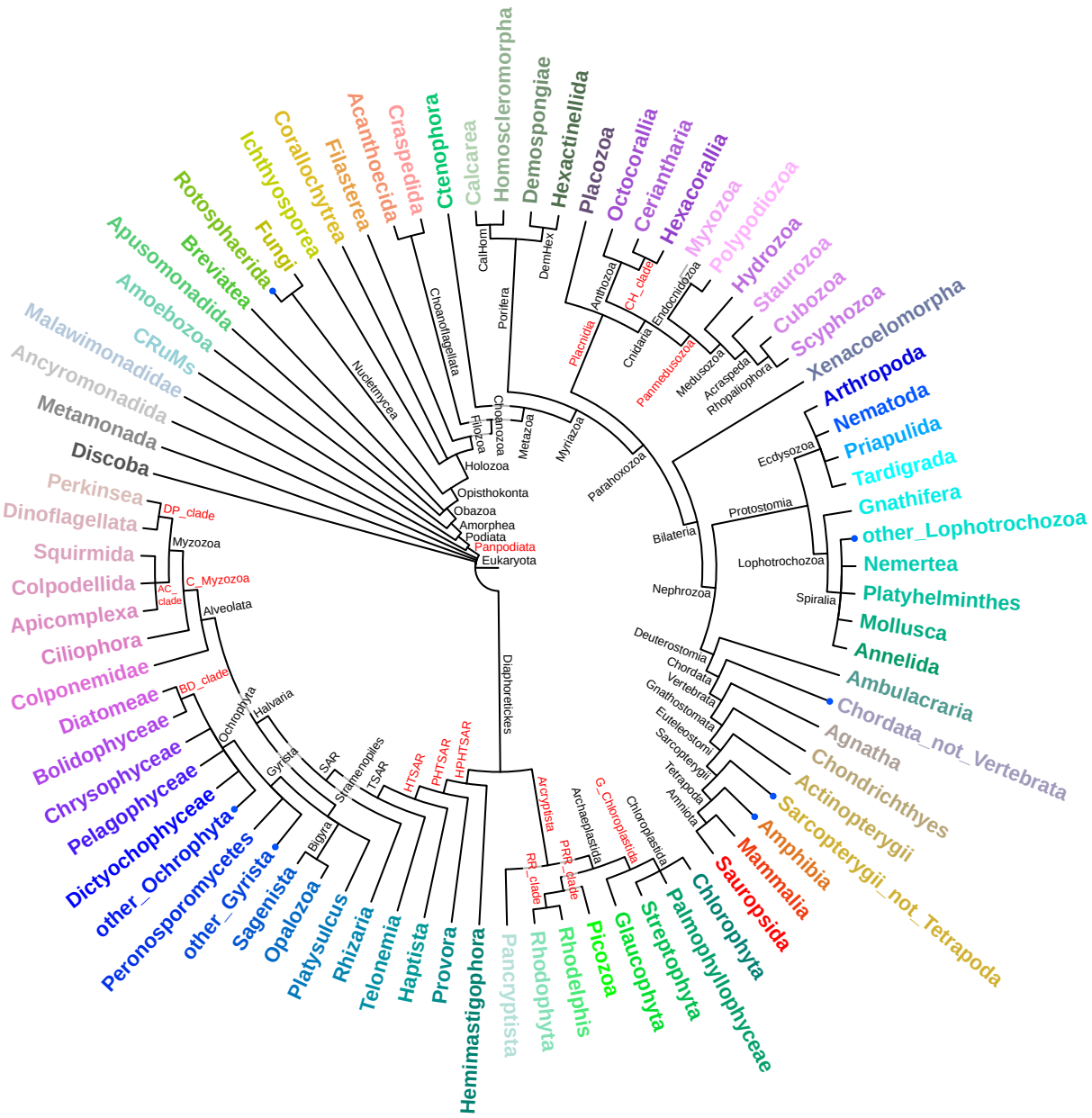


Fig. 1 – The eukaryotic tree of life used in LukProt. Each group is assigned a unique color which makes visual interpretation of phylogenies easier. Groups with blue dots are paraphyletic. Internal node names written in red were invented for the database.

Release schedule and data traceability

LukProt is available as an archive containing FASTA files placed inside a directory tree which recapitulates Figure 1. Additionally, archives with BLAST+ v5 database files (one per species, one per taxogroup and one full database) are available to download. The associated Zenodo repository is designed to be easy to use and contain files ready to be implemented in local workflows with minimal user modification required. These files include: the coloring scheme, the changelog, the metadata spreadsheet and helper scripts.

The latest database version is 1.5.1.rev1, based on EukProt v3. Major LukProt updates will trace each new release of EukProt, minor updates will be released as needed. The version will be incremented only if the underlying sequence files are changed; metadata changes will be marked "revisions". All public versions of the database are archived in the Zenodo repository with a unique DOI.

Considerable effort was made to apply the traceability guidelines set by EukProt to the metadata included with LukProt. If available, the source URL and the article DOI is provided. As an improvement over EukProt, metadata also include NCBI taxids.

'LukProt Comparative Set' (LCS)

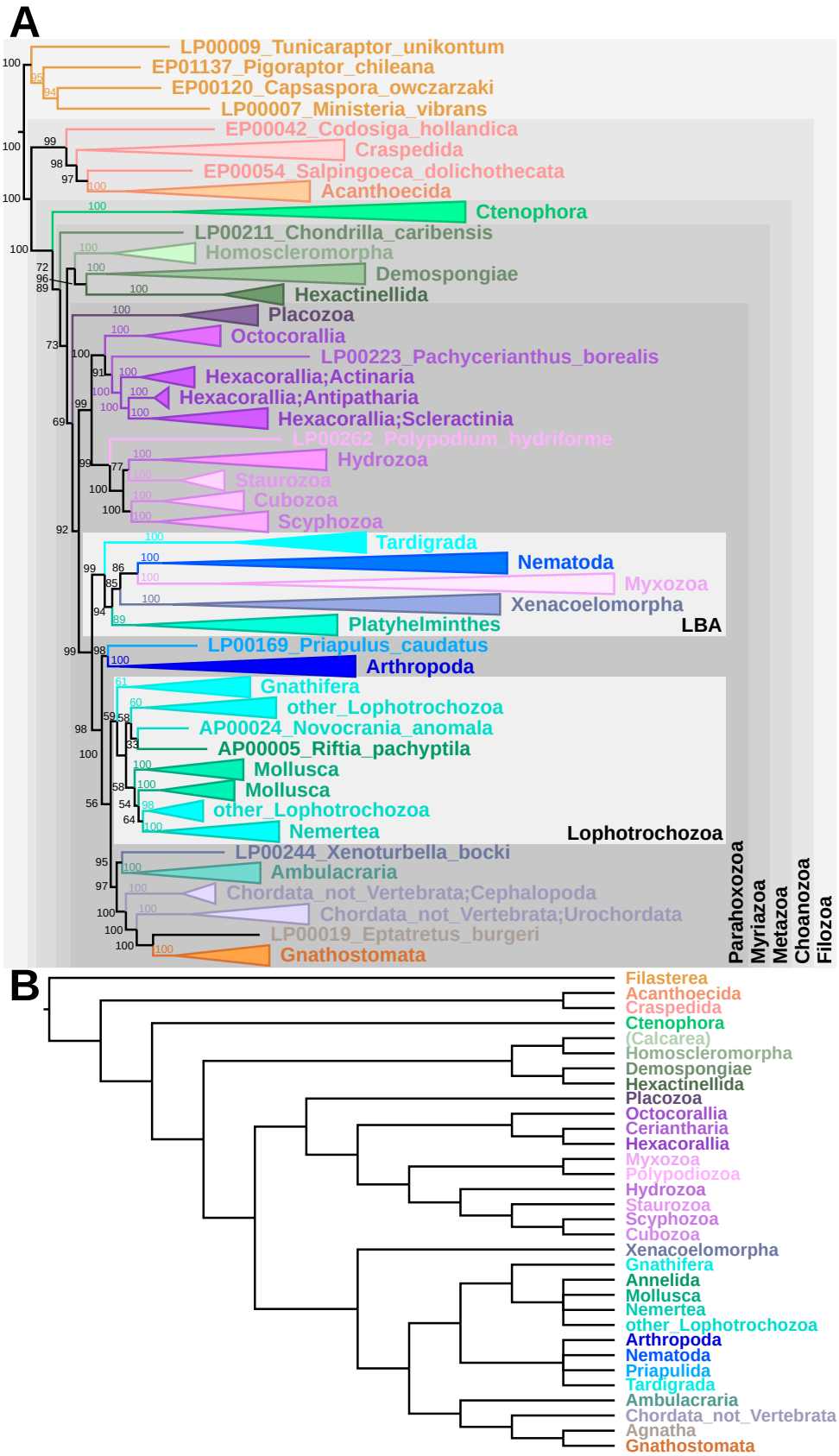
By analogy to 'The Comparative Set' of EukProt – a reduced set of species of high sequencing quality and phylogenetic importance – the LCS is proposed. The set consists of 233 datasets, 18% of the total number, and is available to search using the BLAST server.

BLAST server and convenience scripts

A BLAST+ (Camacho et al., 2009) server for LukProt, available at https://lukprot.hirszfeld.pl/, utilizes SequenceServer 3.1.2 (Priyam et al., 2019) and NCBI BLAST+ 2.16.0+ and incorporates the tree structure from Figure 1. The accompanying scripts are documented with usage examples and can be used to: recolor tree tips, extract sequences from preselected groups of organisms (including the LCS) or extract LukProt sequences from the full BLAST database or using a list of identifiers.

133 **LukProt data sample: top single BUSCOs analysis**

134 Figure 2A presents an example of a phylogeny reconstructed from 20 most often found single
135 eukaryotic BUSCOs in the database. No post-tree reconstruction sequence removal was performed for a
136 fairly unfiltered view of data found in LukProt for the selected clade, Filozoa.


A
LP00009_Tunicaraptor_unikontum
EP01137_Pigoraptor_chileana
EP00120_Capsaspora_owczarzaki
LP00007_Ministeria_vibrans
EP00042_Codosiga_hollandica
Craspedida
EP00054_Salpingoeca_dolichothecata
Acanthoecida
Ctenophora
LP00211_Chondrilla_caribensis
Homoscleromorpha
Demospongiae
Hexactinellida
Placozoa
Octocorallia
LP00223_Pachycerianthus_borealis
Hexacorallia;Actinaria
Hexacorallia;Antipatharia
Hexacorallia;Scleractinia
LP00262_Polypodium_hydriforme
Hydrozoa
Staurozoa
Cubozoa
Scyphozoa
Tardigrada
Nematoda
Myxozoa
Xenacoelomorpha
Platyhelminthes
LBA
LP00169_Priapulus_caudatus
Arthropoda
Gnathifera
other_Lophotrochozoa
AP00024_Novocrania_anomala
AP00005_Riftia_pachyptila
Mollusca
Mollusca
other_Lophotrochozoa
Nemertea
Lophotrochozoa
LP00244_Xenoturbella_bocki
Ambulacraria
Chordata_not_Vertebrata;Cephalopoda
Chordata_not_Vertebrata;Urochordata
LP00019_Eptatretus_burgeri
Gnathostomata
Parahoxozoa
Myriazoa
Metazoa
Choanozoa
Filozoa
B
Filasterea
Acanthoecida
Craspedida
Ctenophora
(Calcarea)
Homoscleromorpha
Demospongiae
Hexactinellida
Placozoa
Octocorallia
Ceriantharia
Hexacorallia
Myxozoa
Polypodiozoa
Hydrozoa
Staurozoa
Scyphozoa
Cubozoa
Xenacoelomorpha
Gnathifera
Annelida
Mollusca
Nemertea
other_Lophotrochozoa
Arthropoda
Nematoda
Priapulida
Tardigrada
Ambulacraria
Chordata_not_Vertebrata
Agnatha
Gnathostomata

138 **Fig. 2** – (**A**) Consensus IQ-TREE phylogeny reconstruction of the 20 most prevalent
139 LukProt BUSCOs in Filozoans. UFBoot2 branch support values are indicated. LBA –
140 long branch attraction artifact. (**B**) Reference, colored cladogram of the group Filozoa.
141 Calcarea is shown in brackets because of its absence from the main phylogeny.

142

143 The phylogeny in Figure 2A mostly agrees with the current understanding of filozoan/early animal
144 evolution (Figure 2B). Ctenophores were recovered as the earliest branching animal clade with full
145 support. Most sponge sequences formed a single clade, as well as Cnidaria and Bilateria. However, the
146 long the branch attraction phenomenon (Susko & Roger, 2021) caused formation of a branch with
147 animals from clades known to have highly divergent sequences.

148 The alignment used here contained 299 species and 5229 parsimony informative sites. If sequences
149 are short, a search of LukProt may return too many for reliable maximum likelihood inference (trees may
150 be overparametrized). In such cases, sequence similarity networks (SSNpipe –
151 https://github.com/ahvdk/SSNpipe), cd-hit or MMSEQS2 (Mirdita et al., 2019) clustering can be used to
152 reduce their number.

153 **Limitations of the LukProt database and future plans**

154 The limitations of the parent database (EukProt) apply to LukProt as well. One further limitation
155 results from a decision not to predict proteins shorter than 100 amino acids (50 in EukProt). The LCS is a
156 beta version and the community is invited to suggest changes. Any missing NCBI taxids will be added in
157 due course.

158 ## Materials and Methods

159 **Dataset naming and species selection**

160 In most cases, species and strain names from EukProt were conserved. To aid further analyses,
161 extremely long species/strain names were shortened so that no species+strain name is longer than 32
162 characters (including underscores). In these cases, only the organism name was truncated – the EukProt
163 dataset identifier and the protein ID were not changed. The name mapping file is available in the Zenodo
164 repository. Sequence names from source files are included after each LukProt ID.

165 Newly downloaded datasets were assigned names after reading the associated publications, if any
166 referred to the name. Otherwise, names were adapted from the websites from which the datasets were

167 downloaded. As in EukProt IDs, spaces are replaced with underscores in individual FASTA files and
168 additionally periods were removed from the names. To broaden compatibility and reduce file size, the
169 following additional processing steps were performed on the datasets from EukProt: removing asterisk
170 (*) characters and line breaks from sequences.

171 **Data processing – LukProt**

172 The data processing steps and software versions are detailed in the associated metadata file.
173 Sequence manipulation was done using SeqKit (Shen et al., 2016) and GNU coreutils. Unassembled
174 transcriptomes were assembled using Trinity v2.12 to v2.15.1 (Grabherr et al., 2011; Haas et al., 2013) or
175 TransAbyss v2.0.1 (Robertson et al., 2010). TransAbyss was used 1 case where Trinity returned an error.
176 Proteins were predicted using TransDecoder.LongOrfs (Haas, BJ.
177 https://github.com/TransDecoder/TransDecoder, versions v5.5.0/v5.7.0). Predicted peptides were
178 clustered (in most cases) using cd-hit v4.8.1 (Fu et al., 2012). Strains were merged by running cd-hit on
179 the concatenated FASTA files. Gffread v0.12.7 (Pertea & Pertea, 2020) was used to extract proteins from
180 annotated genomes. NCBI taxonomy IDs were found manually or with the R package taxonomizr (Sherrill-
181 Mix, S. https://github.com/sherrillmix/taxonomizr v0.10.6). BUSCO 5.7.1 (Manni et al., 2021) was used
182 with dataset "eukaryota_odb10" (255 BUSCOs) created on 2024-01-08. BUSCO utilized hmmsearch
183 version 3.1 (http://hmmer.org/). The OMArk (Nevers et al., 2024) version was 2.0.3 and the OMAmer
184 version was 0.3.0, used with database LUCA.h5 (Nov 2022). The NCBI taxdb used with OMArk was
185 downloaded on 2024-05-12. BUSCO, OMAmer and OMArk were run with default parameters and the
186 species NCBI taxid was specified for OMArk where available.
187

188 **Data processing – 20 BUSCOs example**

189 First, 20 most common single BUSCOs in the LukProt database were selected. Sequences of these 20
190 BUSCOs were collected into separate FASTA files for each dataset. Species which possessed fewer than 10
191 of these BUSCOs were filtered out. BUSCO sequences from the remaining species were then

concatenated, merged into a single multi-FASTA file and aligned using FAMSA v2.2.2- (Deorowicz et al., 2016). The alignment was trimmed using trimAl v1.4.rev15 (Capella-Gutierrez et al., 2009) (gap threshold 0.5). Phylogenies were inferred using IQ-TREE 2.3.3. IQ-TREE branch support values were calculated using SH-aLRT (Shimodaira–Hasegawa-like approximate likelihood ratio test, 5000 replicates) (Guindon et al., 2010), aBayes (approximate Bayes) (Anisimova et al., 2011) and UFBoot2 (Ultrafast bootstrap 2, 5000 replicates) (Hoang et al., 2018). IQ-TREE model was selected using ModelFinder (Kalyaanamoorthy et al., 2017). FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used for phylogeny visualization and editing.

Acknowledgements

I would like to thank the maintainers of the EukProt database. In addition, Marcin Czerwiński, Marta Álvarez, Elena Casacuberta, Michelle Leger, Koryu Kin and Iñaki Ruiz-Trillo, are thanked for helpful discussions and permission to release the *Corallochytrium limacisporum* (strain India) proteome. I further thank Michelle Leger and Eduard Ocaña-Pallarès for suggestions about the taxonomic framework.

This work was supported by National Science Centre of Poland (NCN) as part of the grant SONATINA no. 2020/36/C/NZ8/00081 entitled "The role of glycosylation in the emergence of animal multicellularity".

Data availability

The data underlying this article (the LukProt database and associated files) are available on Zenodo: https://doi.org/10.5281/zenodo.7089120 (Sobala, 2022) and https://doi.org/10.5281/zenodo.11324807 (Sobala, 2024).

References

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology*, **60**, 685–699.

Barreira SN, Nguyen A-D, Fredriksen MT, Wolfsberg TG, Moreland RT, Baxevanis AD (2021) AniProtDB: A Collection of Consistently Generated Metazoan Proteomes for Comparative Genomics Studies. *Molecular Biology and Evolution*, **38**, 4628–4633.

Berney C, Ciuprina A, Bender S, Brodie J, Edgcomb V, Kim E, Rajan J, Parfrey LW, Adl S, Audic S, Bass D, Caron DA, Cochrane G, Czech L, Dunthorn M, Geisen S, Glöckner FO, Mahé F, Quast C, Kaye JZ, Simpson AGB, Stamatakis A, Campo J del, Yilmaz P, Vargas C de (2017) UniEuk: Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology*, **64**, 407–411.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

del Campo J (2024) A protein database to study the origin of metazoans. *Peer Community in Genomics*, **1**, 100368.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.

Deorowicz S, Debudaj-Grabysz A, Gudyś A (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, **6**, 33964.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.

Eme L, Tamarit D, Caceres EF, Stairs CW, De Anda V, Schön ME, Seitz KW, Dombrowski N, Lewis WH, Homa F, Saw JH, Lombard J, Nunoura T, Li W-J, Hua Z-S, Chen L-X, Banfield JF, John ES, Reysenbach A-

238 L, Stott MB, Schramm A, Kjeldsen KU, Teske AP, Baker BJ, Ettema TJG (2023) Inference and
239 reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, **618**, 992–999.
240 Fernández R, Tonzo V, Simón Guerrero C, Lozano-Fernandez J, Martínez-Redondo GI, Balart-García P,
241 Aristide L, Eleftheriadi K, Vargas-Chávez C (2022) MATEdb, a data repository of high-quality metazoan
242 transcriptome assemblies to accelerate phylogenomic studies. *Peer Community Journal*, **2**, e58.
243 Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing
244 data. *Bioinformatics*, **28**, 3150–3152.
245 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R,
246 Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C,
247 Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data
248 without a reference genome. *Nature Biotechnology*, **29**, 644–652.
249 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New Algorithms and
250 Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.
251 *Systematic Biology*, **59**, 307–321.
252 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber
253 M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN,
254 Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from
255 RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–
256 1512.
257 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast
258 Bootstrap Approximation. *Molecular Biology and Evolution*, **35**, 518–522.
259 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model
260 selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.
261 Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) BUSCO Update: Novel and Streamlined
262 Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic,
263 Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, **38**, 4647–4654.

Marlétaz F, Timoshevskaya N, Timoshevskiy VA, Parey E, Simakov O, Gavriouchkina D, Suzuki M, Kubokawa K, Brenner S, Smith JJ, Rokhsar DS (2024) The hagfish genome and the evolution of vertebrates. *Nature*, **627**, 811–820.

Mirdita M, Steinegger M, Söding J (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, **35**, 2856–2858.

Murdock DJE, Donoghue PCJ (2011) Evolutionary Origins of Animal Skeletal Biomineralization. *Cells Tissues Organs*, **194**, 98–102.

Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, Glover NM (2024) Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*, 1–10.

Pertea G, Pertea M (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9:304**.

Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, Chowdhary H, Pieniak I, Maynard LJ, Gibbins MA, Moon H, Davis-Richardson A, Uludag M, Watson-Haigh NS, Challis R, Nakamura H, Favreau E, Gómez EA, Pluskal T, Leonard G, Rumpf W, Wurm Y (2019) Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Molecular Biology and Evolution*, **36**, 2922–2924.

Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C (2022) EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal*, **2**, e56.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu A-L, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJM, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. *Nature Methods*, **7**, 909–912.

Ruiz-Trillo I, Kin K, Casacuberta E (2023) The Origin of Metazoan Multicellularity: A Potential Microbial Black Swan Event. *Annual Review of Microbiology*, **77**, 499–516.

Schultz DT, Haddock SHD, Bredeson JV, Green RE, Simakov O, Rokhsar DS (2023) Ancient gene linkages support ctenophores as sister to other animals. *Nature*, **618**, 110–117.

289 Shen W, Le S, Li Y, Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
290 Manipulation. *PLOS ONE*, **11**, e0163962.
291 Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O'Connell BL, Dear P, Martinez DE,
292 Steele RE, Green RE, David CN, Rokhsar DS (2022) Deeply conserved synteny and the evolution of
293 metazoan chromosomes. *Science Advances*, **8**, eabi5884.
294 [dataset]* Sobala ŁF (2022) LukProt - an animal evolution-centric eukaryotic protein database. Dataset.
295 https://doi.org/10.5281/zenodo.7089120
296 Sobala ŁF (2023) Evolution and phylogenetic distribution of endo-α-mannosidase. *Glycobiology*, **33**, 687–
297 699.
298 [dataset]* Sobala ŁF (2024) Supplementary Data associated with the article "LukProt: A database of
299 eukaryotic predicted proteins designed for investigations of animal origins." Dataset.
300 https://doi.org/10.5281/zenodo.11324807
301 Susko E, Roger AJ (2021) Long Branch Attraction Biases in Phylogenetics. *Systematic Biology*, **70**, 838–843.
302 Tessler M, Neumann JS, Kamm K, Osigus H-J, Eshel G, Narechania A, Burns JA, DeSalle R, Schierwater B
303 (2022) Phylogenomics and the first higher taxonomy of Placozoa, an ancient and enigmatic animal
304 phylum. *Frontiers in Ecology and Evolution*, **10**.
305 Tikhonenkov DV, Mikhailov KV, Gawryluk RMR, Belyaev AO, Mathur V, Karpov SA, Zagumyonnyi DG,
306 Borodina AS, Prokina KI, Mylnikov AP, Aleoshin VV, Keeling PJ (2022) Microbial predators form a new
307 supergroup of eukaryotes. *Nature*, **612**, 714–719.
308 Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW,
309 Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ,
310 Spang A, Ettema TJG (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity.
311 *Nature*, **541**, 353–358.

312

313