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

Correspondence: 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 of 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 (a collection of eukaryotic sequences from all known clades). LukProt adopts and extends EukProt naming conventions and includes data from 216 additional animals. The database is associated with a taxogroup scheme suitable for studying early animal evolution. Minor updates to the database will contain metadata corrections and major updates will synchronize LukProt to each new version of EukProt. All public LukProt releases are permanently stored on Zenodo. A BLAST server to search the database is available at https://lukprot.hirszfeld.pl/. The community is invited to participate in

maintaining and correcting LukProt. As it can be searched without downloading locally, the database can be a convenient resource for not only for evolutionary biologists, but the broader scientific community as well.

Keywords: EukProt, LukProt, database, evolution, animals, Metazoa, eukaryote, proteome, dataset

Introduction

Animals (Metazoa) are a branch of the Eukaryota – 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 the much of the Ediacaran fauna is not similar to any living animals, and 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 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 bioinformatic tools, easier access to processing power and broader availability of sequencing data.

The topology of the eukarotic tree of life has been undergoing major revisions as well. 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). It is integrated with UniEuk, a recently created reference taxonomy of eukaryotes (Berney et al., 2017). The broad taxon sampling and careful curation of EukProt makes it useful for many applications. The EukProt database, however, compromises on the number of included species in the clades with a high number of species sequenced (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 more even animal taxon sampling, AniProtDB (Barreira et al., 2021), only encompasses data from 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.

To fill this gap, I have created LukProt. It is a synthesis of the latest available EukProt database, the AniProt database and a significant number of 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. The LukProt database follows the conventions established by EukProt: each dataset has a unique identifier and each protein has a persistent ID composed of the dataset identifier, the species/strain name and protein number. In case a superior dataset from the same species replaces the previous one, the previous ID is retired in a new version of the database. A more detailed characterization of LukProt, as well as EukProt data or metadata modifications made for LukProt, is presented below.

Methods

Dataset naming and species selection

In most cases, species and strain names from EukProt were conserved. To aid further analyses, extremely long species/strain names were shortened so that no species+strain name is longer than 32 characters (including underscores). In these cases, the EukProt dataset identifier and protein ID were not changed. EukProt version 3 is used in the current LukProt revision (v1.5.1).

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

downloaded. As in EukProt IDs, spaces are replaced with underscores in individual FASTA files and additionally periods were removed from the names. After the protein ID, the sequence identifier from the original file is included.

Distribution of the database

LukProt is available as an archive containing FASTA files placed inside a directory tree which recapitulates the taxonomic relationships between the species. Additionally, an archive with BLAST+ v5 database files (one per species, one per taxogroup and one full database) is available to download. Miscellaneous files, such as the coloring scheme, the changelog, the metadata spreadsheet and helper scripts, are available in the associated Zenodo repository. Datasets that were not included or untranslated transcriptomes are not provided. The repository is designed to be easy to use and the provided files can be implemented in local workflows with minimal user modification required.

Data processing – the LukProt database

The data processing steps are detailed in the associated metadata file. Briefly, unassembled transcriptomes were assembled using Trinity v2.14 to v2.15.1 (Grabherr et al., 2011; Haas et al., 2013) or TransAbyss v2.0.1 (Robertson et al., 2010). Proteins were predicted using TransDecoder.LongOrfs (Haas, BJ. https://github.com/TransDecoder/TransDecoder, versions v5.5.0 to v5.7.1) with default options (minimum protein length: 100). Predicted peptides were then clustered (in most cases) using cd-hit v4.8.1 (Fu et al., 2012) using the command: cd-hit -s 0.5 -g 1 -d 0 -T 6 -M 16000 -c 0.99 -i input.fa -o output.fa. Strain merging was done by concatenating the protein FASTA files to a single file and running cd-hit. To extract protein sequences from a genome, gffread v0.12.7 (Pertea & Pertea, 2020) was run using the following parameters: gffread -g genome.fasta annotation.gff -y output.fa -C -J –no-pseudo. An R package taxonomizr (Sherrill-Mix, S. https://github.com/sherrillmix/taxonomizr v0.10.6) and manual curation was used to find the NCBI taxonomy IDs of each species.

In a number of cases, the dataset names were shortened with respect to EukProt, without changing the dataset ID or protein IDn number. Additional processing steps that were done to data from EukProt were: removing any asterisk (*) characters and line breaks from sequences. This was done to broaden compatibility and reduce file size. Sequence manipulation was done using SeqKit (Shen et al., 2016) and GNU coreutils.

Data processing software – huntingtin example

Protein sequence alignment was performed using FAMSA v2.2.2- (Deorowicz et al., 2016) (-refine-mode on). Alignment was trimmed using trimAl v1.4.rev15 (Capella-Gutierrez et al., 2009). Phylogenies were inferred using VFT (VeryFastTree) 4.0.3 (Piñeiro et al., 2020) or IQ-TREE (2.2.6). IQ-TREE branch support values were calculates 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) algorithms. 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.

Results and Discussion

General 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 of animals and their closest relatives, holozoans. To this end, individual dataset quality is of lower importance. No

decontamination procedures were applied to any datasets and none were excluded because of possible contamination. Contamination was, in fact, used to extract a new dataset – LP00177 – containing choanoflagellate-derived sequences found in EP00436 from EukProt v2 (superseded by EP00913 in EukProt v3). Many of the LukProt transcriptome-derived proteomic datasets from cnidarians and sponges are in fact holobiont transcriptomes which include bacterial transcripts and those from other associated organisms. It is assumed that phylogeny reconstructions will be the guide when drawing conclusions from database searches. Phylogeny-guided analysis can also be the basis for studies of lateral gene transfers.

The purpose of scripts associated with the database is to accelerate phylogeny inference and interpretation of the results. The documentation should guide a less experienced user and let them perform independent investigations.

Characterization of the LukProt database

The current version of the database (v1.5.1) contains 1281 datasets, with 990 coming from EukProt, 41 from AniProt and 250 newly added. The database statistics are given in **Table 1**. Taxon sampling of the early branching animal clades, especially ctenophores, sponges and cnidarians, is improved. Placozoans account for six datasets but this constitutes 26% of all currently known, potential placozoan species (23) (Tessler et al., 2022).

Table 1 – Comparison of the number of taxa in selected clades of LukProt and other databases.

Taxogroup	EukProt v3 (included in stock database)	AniProt within LukProt	LukProt v1.5.1 added/excluded	LukProt v1.5.1 total
Eukaryota	993	41	+250/-3	1281
Diaphoretickes	632	0	+28/-0	660
Holozoa (excluding Metazoa)	105 (40)	41 (0)	+221/-2 (+5/-1)	365 (44)
Metazoa (excluding Bilateria)	65 (14)	41 (3)	+216/-1 (+162/-0)	321 (179)
Ctenophora	2	1	+35/-0	38
Porifera	5	1	+41/-0	47
Placozoa	2	0	+4/-0	6
Cnidaria	5	1	+82/-0	88
Bilateria	51	38	+54/-1	142

A consensus taxonomy

Included in the database is a consensus taxonomy of eukaryotes. To aid analysis of animal evolution, some UniEuk (Berney et al., 2017) taxogroups outside Podiata were merged, and Podiata was split into more granular groups (mostly monophyletic). This was done to aid analyses that involve deep animal phylogenies. Clades are explained in Supplementary Information and the full eukaryotic tree can be found in Figure 1. Each internal node was assigned a unique name, following UniEuk or recent literature where possible. A number of unnamed clades were assigned provisional names (Figure 1). The LukProt taxonomic framework is necessarily different from UniEuk but will be adjusted followin the updates that are implemented in the latter and in future publications, as well as community suggestions. It encompasses 81 groups of organisms, most of which are monophyletic. Based on this taxonomy, 59 groups to limit the number of species were created to aid analyses focused on narrower clades. These groups can be found in the databate metadata.

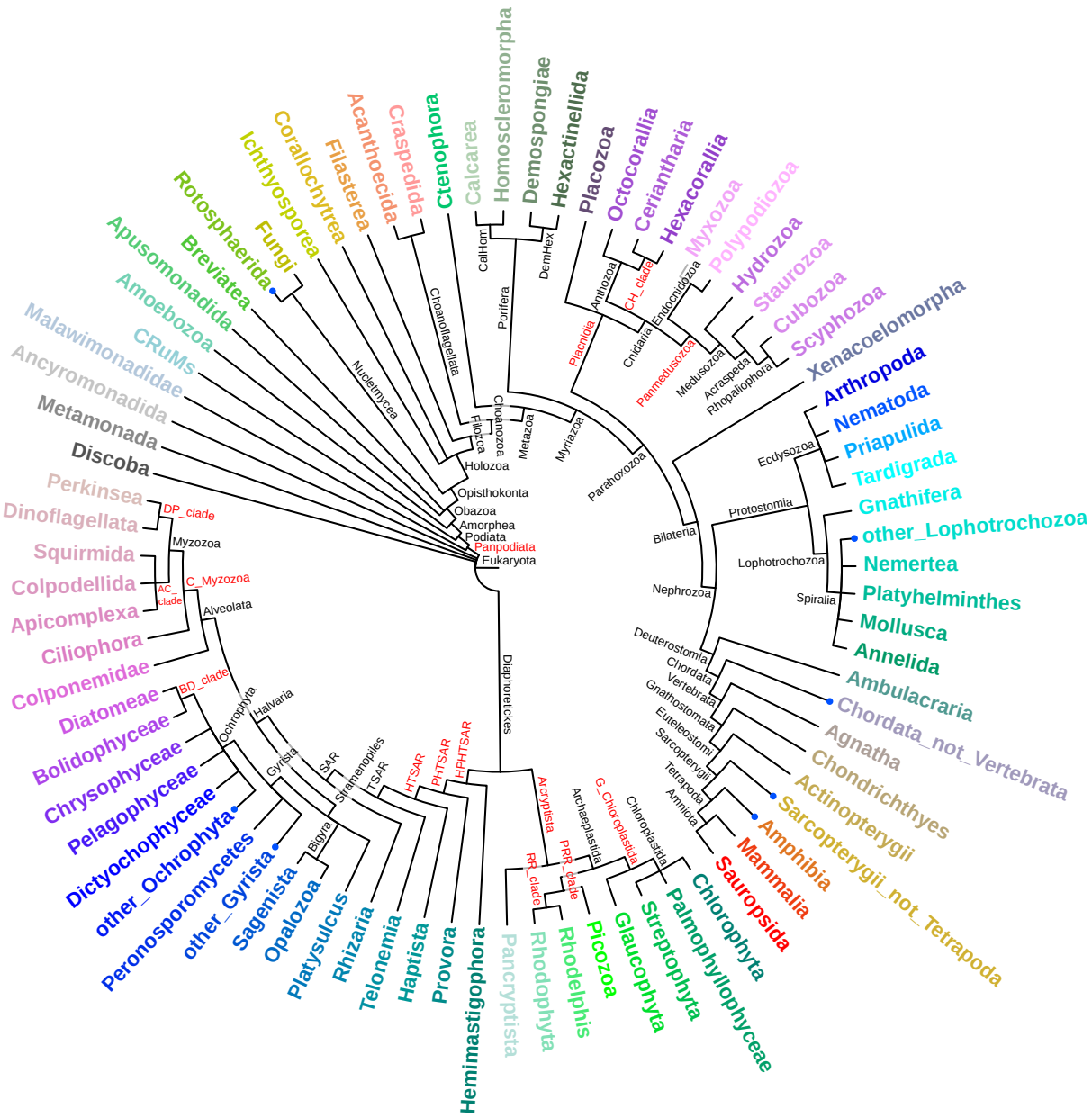

Figure 1 – the eukaryotic tree of life used in LukProt. Each group is assigned a unique color which makes visual interpretation of phylogenies easier. Groups with blue dots are paraphyletic. Internal node names written in red were invented for the database.

Source data traceability

Each dataset is associated with a URL and a DOI for the source article – if it exists – following the EukProt conventions. Considerable effort was made to apply the traceability guidelines set by EukProt to the metadata included with LukProt. As an improvement over EukProt, the metadata file now includes the NCBI taxids for the species for which they were possible to find (including EukProt-derived species).

Release schedule

Version 1.4.1 of LukProt of was made available on Zenodo in 2022 and was already downloaded 38 times (as of 29/01/2024). It was based on EukProt v2. Since then, many additions and improvements were introduced, which resulted in the publication-ready version 1.5.1, based on EukProt v3. LukProt will receive major updates to trace each new release of the EukProt database; it should be considered its

extension. Minor updates will be released as needed. In case of EukProt including any of the LukProt datasets, the LukProt ID will be retired and the EukProt ID will be used instead. Future releases will include the improvements in EukProt and contain all the additional data provided by LukProt. Changes will be documented and explained in readable form in a changelog. All public versions of the database are archived in the Zenodo repository and a unique DOI is generated for each new version (as well as a DOI which always resolves to the newest version).

BLAST server and convenience scripts

A publicly available BLAST+ (Camacho et al., 2009) server for LukProt is available at https://lukprot.hirszfeld.pl/ using SequenceServer 2.0.0 (Priyam et al., 2019) and NCBI BLAST 2.15.0+ as backends. The taxonomy in Figure 1 is applied to the database and broad or narrow clades can be readily selected from the tree (or alphabetically). In comparison, currently the EukProt BLAST server (https://evocellbio.com/SAGdb/EukProt/) only provides an alphabetic list of taxogroups, without tree of life topology. The LukProt BLAST server will be updated as new versions of the database are released.

The functions of the executable scripts that accompany the database include but are not limited to: extracting sequences from preselected groups of organisms (for example: Diaphoretickes, Holozoa excluding Metazoa), extracting LukProt sequences from the full BLAST database or using a list of identifiers and tree tip recoloring.

An example of LukProt application: evolution of huntingtin

One example of the utility of the database can be seen in my previous work, where an earlier version of the database (v1.4.1) was used (Sobala, 2023). For the purpose of the current publication, huntingtin was selected as a protein widely distributed among amorpheans and suitable for deep phylogenetic inference due to its length and conservation. Human huntingtin (UniProt ID: P42858) was used as a query for blastp against the full LukProt v1.5.1 database (e-value cutoff: 10^{-10}). The resulting hits (1042) were clustered using cd-hit at 80% identity, reducing their number to 672, and aligned using FAMSA. After this, trimAl was used to trim the sequences at various gap threshold values (1-10%). Gap threshold of 2% (fewest bad splits) was selected for initial phylogenetic inference. Each initial phylogenetic tree can be generated using VFT in under 2 minutes each using accessible hardware (a laptop with Ryzen 4600H processor and 32 GB RAM). The tips were colored according to the color scheme (Figure 1) and outliers were removed via manual tree inspection in FigTree, leaving 634 sequences, which were then clustered again at 65% identity.

Subsequently, the remaining 522 representative sequences were re-aligned, trimmed and pruned again and the best result of VFT inference (gap threshold 9%) was used for IQ-TREE phylogeny reconstruction. After the second round of sequence removal (506 remaining), the above procedure was repeated for final phylogeny inference with IQ-TREE (gap threshold of 6%). The best fitting model (selected by the Bayesian information criterion) was Q.pfam+F+R9. Full trees with branch support values are available in the Supplementary Information repository.

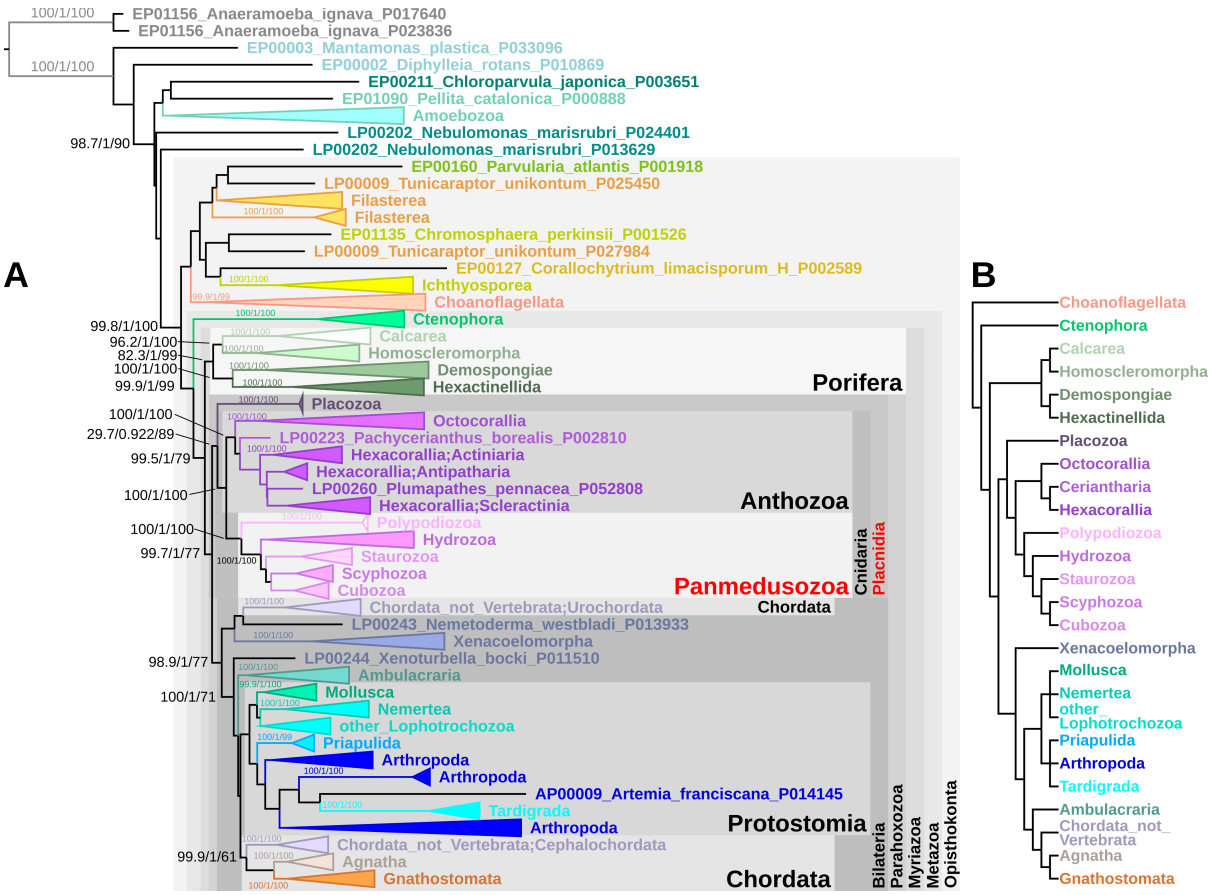

Figure 2 – IQ-TREE result of phylogenetic analysis of huntingtin proteins found in LukProt, manually rooted on the *Anaeramoeba* sequences clade and with broader clades collapsed. Selected branch support values are shown: SH-aLRT/aBayes/UFBoot2.

The obtained phylogeny (Figure 2A) broadly agrees with the current understanding of early animal evolution (Figure 2B). Ctenophore proteins were recovered as the earliest branching animal clade with high support values. All sponge sequences formed a single clade, as well as cnidarian and bilaterian sequences. Note the recovery of "Placnidia" (albeit with low support) and correct resolution of internal sponge and cnidarian phylogeny, despite only 1 ceriantharian sequence present. Non-animal holozoan sequences were placed in a separate clade, possibly suggesting an animal-specific change in huntingtin or pointing to the limitations of the single gene phylogeny approach. Urochordate sequences, which are often quite divergent, were incorrectly placed within Xenacoelomorpha due to long the branch attraction phenomenon (Susko & Roger, 2021). Some of the removed sequences were from representatives of specific animal clades (for example Nematoda) which were not present in the final phylogeny. Alternative approaches, such as sequence similarity networks (SSNpipe – https://github.com/ahvdk/SSNpipe), can help with analyses that require more complete datasets.

This example illustrates that results from LukProt searches can be readily used for downstream analyses. If only VFT is used, the phylogeny quality may be compromised but the procedure can be completed in about 15 minutes. The associated metadata can be used together with phylogenetic analyses to determine the clades in which a particular protein can be found.

Limitations of the LukProt database and future plans

As an extension of EukProt, the limitations of the parent database apply to LukProt as well. One further limitation results from an editorial decision not to predict proteins shorter than 100 amino acids (50 in EukProt). No reduced "comparative set" of species was introduced and BUSCO completeness scores were not determined. A LukProt comparative set can be introduced later with community involvement. A number of NCBI taxids are missing; these can be added in future database versions.

Acknowledgements

I would like to thank the creators and maintainers of the EukProt database as well as the teams and individuals who generated the source data. Members of the Multicellgenome Lab (IBE-CSIC, Barcelona): Marta Álvarez, Elena Casacuberta, Michelle Leger, Koryu Kin and Iñaki Ruiz-Trillo, are thanked for helpful discussions and the permission to release the protein dataset from *Corallochytrium limacisporum* (strain India). Michelle Leger and Eduard Ocaña-Pallarès are acknowledged for suggestions about the taxonomic framework. I also acknowledge Marcin Czerwiński for helpful comments about the manuscript.

Data, scripts, code, and supplementary information availability

The LukProt database and associated files are available on Zenodo: https://doi.org/10.5281/zenodo.7089120 (Sobala, 2022).

Supplementary information is available online: https://doi.org/10.5281/zenodo.10580775 (Sobala, 2024)

Conflict of interest disclosure

The author declares no financial conflicts of interest in relation to the content of the article.

Funding

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".

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. https://doi.org/10.1093/sysbio/syr041

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. https://doi.org/10.1093/molbev/msab165

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. https://doi.org/10.1111/jeu.12414

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421. https://doi.org/10.1186/1471-2105-10-421

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. https://doi.org/10.1093/bioinformatics/btp348

Deorowicz S, Debudaj-Grabysz A, Gudyś A (2016) FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, **6**, 33964. https://doi.org/10.1038/srep33964

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. https://doi.org/10.1038/nature06614

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-L, Stott MB, Schramm A, Kjeldsen KU, Teske AP, Baker BJ, Ettema TJG (2023) Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature*, 1–8. https://doi.org/10.1038/s41586-023-06186-2

Fernández R, Tonzo V, Simón Guerrero C, Lozano-Fernandez J, Martínez-Redondo GI, Balart-García P, Aristide L, Eleftheriadi K, Vargas-Chávez C (2022) MATEdb, a data repository of high-quality metazoan transcriptome assemblies to accelerate phylogenomic studies. *Peer Community Journal*, **2**. https://doi.org/10.24072/pcjournal.177

Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652. https://doi.org/10.1038/nbt.1883

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321. https://doi.org/10.1093/sysbio/syq010

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**, 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, **35**, 518–522. https://doi.org/10.1093/molbev/msx281

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589. https://doi.org/10.1038/nmeth.4285

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*, 1–3. https://doi.org/10.1038/s41586-024-07070-3

Murdock DJE, Donoghue PCJ (2011) Evolutionary Origins of Animal Skeletal Biomineralization. *Cells Tissues Organs*, **194**, 98–102. https://doi.org/10.1159/000324245

Pertea G, Pertea M (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9:304**. https://doi.org/10.12688/f1000research.23297.2

Piñeiro C, Abuín JM, Pichel JC (2020) Very Fast Tree: speeding up the estimation of phylogenies for large alignments through parallelization and vectorization strategies. *Bioinformatics*, **36**, 4658–4659. https://doi.org/10.1093/bioinformatics/btaa582

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. https://doi.org/10.1093/molbev/msz185

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**. https://doi.org/10.24072/pcjournal.173

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. https://doi.org/10.1038/nmeth.1517

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. https://doi.org/10.1146/annurev-micro-032421-120023

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. https://doi.org/10.1038/s41586-023-05936-6

Shen W, Le S, Li Y, Hu F (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, **11**, e0163962. https://doi.org/10.1371/journal.pone.0163962

Simakov O, Bredeson J, Berkoff K, Marletaz F, Mitros T, Schultz DT, O'Connell BL, Dear P, Martinez DE, Steele RE, Green RE, David CN, Rokhsar DS (2022) Deeply conserved synteny and the evolution of metazoan chromosomes. *Science Advances*, **8**, eabi5884. https://doi.org/10.1126/sciadv.abi5884

Sobala ŁF (2022) LukProt – an animal evolution-centric eukaryotic protein database. https://doi.org/10.5281/zenodo.7089120

Sobala ŁF (2023) Evolution and phylogenetic distribution of endo-α-mannosidase. *Glycobiology*, **33**, 687–699. https://doi.org/10.1093/glycob/cwad041

Sobala ŁF (2024) Data associated with the article "LukProt: A database of eukaryotic predicted proteins designed for investigations of animal origins." https://doi.org/10.5281/zenodo.10580775

Susko E, Roger AJ (2021) Long Branch Attraction Biases in Phylogenetics. *Systematic Biology*, **70**, 838–843. https://doi.org/10.1093/sysbio/syab001

Tessler M, Neumann JS, Kamm K, Osigus H-J, Eshel G, Narechania A, Burns JA, DeSalle R, Schierwater B (2022) Phylogenomics and the first higher taxonomy of Placozoa, an ancient and enigmatic animal phylum. *Frontiers in Ecology and Evolution*, **10**. https://doi.org/10.3389/fevo.2022.1016357

Tikhonenkov DV, Mikhailov KV, Gawryluk RMR, Belyaev AO, Mathur V, Karpov SA, Zagumyonnyi DG, Borodina AS, Prokina KI, Mylnikov AP, Aleoshin VV, Keeling PJ (2022) Microbial predators form a new supergroup of eukaryotes. *Nature*, **612**, 714–719. https://doi.org/10.1038/s41586-022-05511-5

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, Stott MB, Nunoura T, Banfield JF, Schramm A, Baker BJ, Spang A, Ettema TJG (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358. https://doi.org/10.1038/nature21031