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

## Summary

The author expands a previously published eukaryotic proteomes database (EukProt), greatly increasing the taxon sampling in holozoans. This effort will definitely be useful for the field investigating this clade and its origin and in general for comparative genomics. However, I have some criticism especially regarding the longevity and reproducibility of the database itself and the lack of any quality control/statistics on the newly added genomes.

I consider the article suitable for publication after some revisions.

## Major

- I see the point the author is trying to make in line 123 but I would have expected at least a minimum analysis of the database content (or at least for the newly added proteomes), for example: number of proteins, average protein length, N50/L50 for genomes (when available). Moreover, a “comparative set” and BUSCO completeness score are listed as limitations but I consider them to be quickly and easily solvable. First, EukProt already has a “comparative set” so it would only be necessary to choose the best and most representative proteomes from the 216 new ones. Secondly, BUSCO (or other softwares such as the newly published OMArk, which also tries to assess contamination) is very quick and I do not really see why not running it. Overall, it’s very difficult this way to assess how trustable this database is, even if you assume that a single contaminated or low-quality dataset will not influence much.
- My second main point is on the Huntingtin example analysis. Firstly, no criterias are listed for manual curation of the outliers which can heavily bias the final tree topology. Further, when clustering it is not clear if the representative tip of each cluster was annotated with the taxonomic latest common ancestor of the cluster or not. In my opinion this should be done when clustering sequences. Further, the author tests many different trimming parameters but no discussion on how variable the tree topologies of the fast trees are. Is this parameter that much important? Overall, the analysis seems a bit unnecessarily complicated and I think that they make the example much less powerful than what it could actually be.
- I consider this database a good effort, especially how the author took care in homogenizing with EukProt and granting this would be kept in newer versions. However, as too often in bioinformatics, this resource may be quickly discontinued as, currently, it’s a single individual’s effort. Further, there is no code repository to try to reproduce/update what has been done. The scripts in the zenodo folder are just

utilities to parse/analyse the database. I think it would be ideal to share a version controlled repository with scripts/documentation to try to solve this.

## Minor

### Main text

line 37: phrased like this it almost seems that the debate was on Eukaryotes, not Metazoa
line 39: the much of -> much/most of
line 30: for not only for -> not only for
line 100: An R package -> The R package
line 146: followin -> following
line 150: misspelling in “databate metadata”
Fig2: caption does not explain A and B

### Metadata

- sometimes “MMETSP – be very careful” other times only “MMETSP”
- In the metadata csv file: Chlorochromonas danica notes: “also EP01039 – why are there 2?” Indeed, why?
- In general, and I consider this is something that also EukProt is lacking, it would be ideal to also link to downloadable genomes and gffs when the proteome comes from a genome. If the author is interested he may feel free to contact me as I’ve been collecting these URLs for EukProt.

Finally, just a little note that MateDB v2 was recently released (https://doi.org/10.1101/2024.02.21.581367) greatly expanding the number of proteomes which might be useful to the author for future releases of LukProt.