Peer Community In Genomics

EukProt enables reproducible Eukaryota-wide protein sequence analyses

Gavin Douglas based on reviews by 2 anonymous reviewers

A recommendation of:

Open Access

EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes

Daniel J. Richter, Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, Colomban de Vargas *(2022), bioRxiv, 2020.06.30.180687, ver. 5 peer-reviewed and recommended by Peer Community in Genomics*

https://doi.org/10.1101/2020.06.30.180687

Published: 15 September 2022

Data used for results

- ***https://doi.org/10.6084/m9.figshare.12417881***
- ***http://evocellbio.com/eukprot/***

Scripts used to obtain or analyze results

- ***https://doi.org/10.5281/zenodo.7025267***
- ***https://github.com/beaplab/EukProt***

Submitted: 08 June 2022, Recommended: 10 September 2022

Cite this recommendation as:
Gavin Douglas (2022) EukProt enables reproducible Eukaryota-wide protein sequence analyses. *Peer Community in Genomics, 100021.* ***https://doi.org/10.24072/pci.genomics.100021***

Recommendation

Comparative genomics is a general approach for understanding how genomes differ, which can be considered from many angles. For instance, this approach can delineate how gene content varies across organisms, which can lead to novel hypotheses regarding what those organisms do. It also enables investigations into the sequence-level divergence of orthologous DNA, which can provide insight into how evolutionary forces differentially shape genome content and structure across lineages.

Such comparisons are often restricted to protein-coding genes, as these are sensible units for assessing putative function and for identifying homologous matches in divergent genomes. Although information is lost by focusing only on the protein-coding portion of genomes, this simplifies analyses and has led to crucial findings in recent years. Perhaps most dramatically, analyses based on hundreds of orthologous proteins across microbial eukaryotes are fundamentally changing our understanding of the eukaryotic tree of life (Burki et al. 2020).

These and other topics are highlighted in a new pre-print from Dr. Daniel Richter and colleagues, which describes EukProt (Richter et al. 2022): a database containing protein sets from 993 eukaryotic species. The authors provide a BLAST portal for matching custom sequences against this database (https://evocellbio.com/eukprot/) and the entire database is available for download (https://doi.org/10.6084/m9.figshare.12417881.v3). They also provide a subset of their overall dataset, 'The Comparative Set', which contains only high-quality proteomes and is meant to maximize phylogenetic diversity.

There are two major advantages of EukProt:

1. It will enable researchers to quickly compare proteomes and perform phylogenomic analyses, without needing the skills or the time commitment to aggregate and process these data. The authors make it clear that acquiring the raw protein sets was non-trivial, as they were distributed across a wide variety of online repositories (some of which are no longer accessible!).

2. Analyses based on this database will be more reproducible and easily compared across studies than those based on custom-made databases for individual studies. This is because the EukProt authors followed FAIR principles (Wilkinson et al. 2016) when building their database, which is a set of guidelines for enhancing data reusability. So, for instance, each proteome has a unique identifier in EukProt, and all species are annotated in a unified taxonomic framework, which will aid in standardizing comparisons across studies.

The authors make it clear that there is still work to be done. For example, there is an uneven representation of proteomes across different eukaryotic lineages, which can only be addressed by further characterization of poorly studied lineages. In addition, the authors note that it would ultimately be best for the EukProt database to be integrated into an existing large-scale repository, like NCBI, which would help ensure that important eukaryotic diversity was not ignored. Nonetheless, EukProt represents an excellent example of how reproducible bioinformatics resources should be designed and should prove to be an extremely useful resource for the field.

References

Burki F, Roger AJ, Brown MW, Simpson AGB (2020) The New Tree of Eukaryotes. Trends in Ecology & Evolution, 35, 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, Vargas C de (2022) EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. bioRxiv, 2020.06.30.180687, ver. 5 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2020.06.30.180687

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Reviews

Toggle reviews

Evaluation round #2

DOI or URL of the preprint: *https://doi.org/10.1101/2020.06.30.180687*

Version of the preprint: 3

Author's Reply, 09 Sep 2022

Download author's replyDownload tracked changes file

Decision by Gavin Douglas, 30 Aug 2022

To Dr. Richter and colleagues,

I think your changes have addressed most of the reviewer's comments (except for one minor comment – see below) and I think the manuscript is in excellent condition, and requires only a small tweak prior to recommendation.

One important thing to note is that I received a "timed out" error when trying to load http://evocellbio.com/eukprot/ - I'm guessing this was just a transient problem, but should be checked.

The minor comment that I think the authors perhaps missed was this partial statement from reviewer 1: "…mention the fact that they cannot technically evaluate the tools and parameters selection for the de novo transcriptome assembly paragraphs (lines 300-306) and the automated genome annotation (lines 329-338)"

Those line numbers no longer match, but the sections correspond to the paragraphs starting with "'assemble mRNA': de novotranscriptome assembly " and "'predict genes': we used EukMetaSanity", respectively. I think either a little more explanation of why these parameters were chosen (e.g., why stating why using the same parameters as Alexander et al. 2021 makes sense, in the case of the predict genes). If the options are somewhat arbitrary, which might be the case with the assembly and filtering options, then the authors could mention that these options were not evaluated but are similar to what are commonly used, which I believe would address the reviewer's point.

Last, I recommend two very minor changes:

In your title, I recommend that you change "a database" after the "Eukprot:" to be "A database". I believe that most style guides suggest the latter format, but the former is widespread in the scientific literature so I leave that choice to you.

I do however strongly think that the link to the webserver should be added to the abstract, which I think many readers comes to expect when reading about bioinformatics resource.

Once these last changes are addressed I would be pleased to recommend your article.

All the best,

Gavin Douglas

Evaluation round #1

DOI or URL of the preprint: ***https://doi.org/10.1101/2020.06.30.180687***

Version of the preprint: 1

Author's Reply, 29 Aug 2022

Download author's replyDownload tracked changes file

Decision by Gavin Douglas, 14 Jul 2022

Two reviewers have completed their assessments and have determined that the manuscript is sound and describes a useful resource for the field. They have requested only minor revisions.

I share their enthusiasm for this resource and look forward to seeing the next draft of this manuscript.

I concur with reviewer #2 that more detailed discussion regarding how their resource differs from PhyloDB is needed. I also did not find Table 1 very informative, and I would think a supplementary table listing the actual URLs used would be more relevant to interested readers. But since the reviewers did not take issue with the table, I will not require it to be changed, and leave the decision to the authors' preference.

Given the small changes requested, the authors should aim to format their manuscript for according to PCI Genomics guidelines (see: https://genomics.peercommunityin.org/help/guide_for_authors#h_327311378567161970523484 7)

Formatting issues that I noticed are

- Table and Figure should be embedded within the text.

- An email for the corresponding author should be indicated

- Rather than a database availability statement, this should be moved to the end of the abstract (for the link to the database webserver), and also mentioned in the "Data, script and code availability" section at the end of the manuscript.

- I believe moving all of the descriptions of the database to a "Results and Discussion" heading would be most appropriate for this article type (and the current headings, such as "The EukProt Database" changed to sub-headings). Based on the formatting guidelines, PCI Genomics strongly recommends separate Results and Discussion headers, but I think a combined section would be acceptable in this case, as the manuscript is very clear as it is.

- The methods section should be moved before the Results section

- Please re-format the acknowledgements section to match the recommended format. Also, is the lower-case "i" in Núria Ros i Rocher a typo? I think this is supposed to be a hyphen.

- Please add a Data, script, and code availability section at the end. Note that the authors' custom code must also be made available in this section.

- Make in-text citations square brackets when they are within parenthetic phrases (e.g., "Eisen, 2003" at L48.

Reviewed by anonymous reviewer, 11 Jul 2022

The present manuscript presents the protein based database EukProt that has been build on reference data from genome, single cells and transcriptomes. This update aligns with the FAIR principles and introduces a new high quality reference dataset that was explicitly setup for comparative genomics and that tries to meet a high taxonomic standard that aligns with UniEuk.

The manuscript is well justified and clear in its description and outlines. Thus, the only critique that I have that it missed to point the limitations of EukProt in a specific manner. For future users, however, the limitations are as important as the strenghts, in particular, when used as reference for the whole scientific community.
Therefore, I'd like to recommend to add a small paragraph that points out the limitation of the database. This could for instance highlight cases, in which the database will be of only limited use (e.g. a list of lineages that are not well covered (to balance the statement of the lineages that had more than 100 taxa) could be pointed out here and that still require joined sequencing efforts; similarily this could be pointed out for the comparative genomics), or limitations of the current gene prediction models, taxonomic paths, ... (not all maybe neccessary though, but I'd least mention/discuss the most important ones for the users)

Thank you

Reviewed by anonymous reviewer, 13 Jul 2022

This article presents the release of EukProt: a database of eukaryotic genome-scale predicted proteins. The manuscript nicely outlines the pitfalls in shared genomics data accessibility and presents EukProt as a solution for several challenges of comparative genomics analyses, which will become even stronger with the exponential increase in genomic data production. It then continues by describing the database utilities, downloadables, generic structure, abidance to FAIR principles, and community-provided update possibilities and finishes with a detailed description of the methodology.

The title and abstract are clear and straight to the point. Overall the article excellently stands out for its clarity, detailed methodology, input database specifications, comprehensiveness, and range of bioinformatics challenges that the authors address with the development of this resource. The amount of considered repositories from which the database is constructed is impressive, and so is the subsequent integration of custom processed raw data (assemblies, annotations). The authors have a clear and deep knowledge of the comparative genomics issues that the scientific community is facing and provide an elegant solution through a genomic analysis framework enriched with some of the most solid and state-of-art comparative genomics tools (examples: the UniEuk taxonomic framework, BUSCO completeness scores). It indicates particular sensitivity and integrity of the authors toward a modern (e.g. foreseeing the ocean metagenomics data integration) and virtuous (e.g. providing various downloadables such as genome annotations) way of approaching bioinformatics resource development. This sensitivity is mostly exemplified by the presentation of The Comparative Set (TCS), a selection of taxonomically fairly-distributed, highly complete predicted-protein sets, which will hopefully serve as a basis for many comparative genomics analyses in future eukaryotic biology studies.

This reviewer will only provide a few minor comments about the clarity of some sentences, as well as mention the fact that they cannot technically evaluate the tools and parameters selection for the de novo transcriptome assembly paragraphs (lines 300-306) and the automated genome annotation (lines 329-338). This reviewer particularly praises the care given to the methods producing The Comparative Set.

This reviewer would be happy to see this resource further expand and recommends it for PCI Genomics validation.

Minor comments:

Lines 68-70: The authors could better explain how EukProt differentiates from PhyloDB.

Lines 73-75: This reviewer could not find protein data files comprising protein domains, Interpro, or gene ontologies from the downloadables (genome annotations, protein fasta files). Not clear if they are provided or if they are mentioned as an example of data with difficult accessibility. Either way, it could be better explained. EDIT: found the mention of potential addition in the future at lines 205-208, this reviewer would still advise rephrasing lines 73-75 for immediate clarity.