# OrthoGather: a local platform for orthology-based proteome and proteomics comparisons and Gene Ontology enrichment

Carlos Vivas-Rodríguez[1,2,3], David Matallanas[3,4], Colm J. Ryan[2,3], Siobhán McClean[1,2], Olivier Dennler[2,3*], Joanna Drabinska[1,2*]

1. School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland
2. UCD Conway Institute of Biomedical and Biomolecular Research, Dublin, Ireland
3. School of Medicine, University College Dublin, Dublin, Ireland
4. Systems Biology Ireland, University College Dublin, Dublin, Ireland

*** Corresponding authors:** Joanna Drabinska (joanna.drabinska.pro@gmail.com);
Olivier Dennler (olivier.dennler@gmail.com)

# Abstract

**Motivation:**

Comparative proteomic analysis may reveal common and unique pathways regulated by the same stimulus across species using data from differential protein expression studies or curated protein sets. Functional annotations are key but vary in quality, as many proteins, particularly in prokaryotes and non-model eukaryotes, are poorly or inconsistently annotated, complicating comparative studies. Orthology inference provides a robust framework to address this, but existing tools require technical expertise, command-line use, and manual processing of complex outputs, creating barriers for researchers without computational training.

**Results:**

We developed *OrthoGather*, a locally hosted web application that streamlines comparative proteomic analysis by integrating homologous protein groups across species and Gene Ontology (GO) enrichment. It leverages functional annotations from any orthogroup member to enable functional inference even when individual species lack comprehensive annotation. Its flexible design supports cross-species exploration of conserved and unique orthogroups across proteomes or user-defined protein sets, revealing functional patterns through orthogroup relationships. *OrthoGather* generates publication-ready, easy-to-interpret outputs including downloadable graphs and data files, lowering barriers for researchers without computational expertise.

**Availability and implementation:**

Source code, documentation and tutorials are available at Zenodo (https://doi.org/10.5281/zenodo.18603238) and GitHub (https://github.com/CarlosVivasR/OrthoGather). Supplementary materials, including the example dataset analysis are available online at *Bioinformatics*.

## Introduction

The incomplete functional annotation of proteins remains a major challenge in proteomics, particularly when researchers seek to interpret the biological significance of protein sets derived from experimental studies. This is especially challenging in prokaryotes or non-model eukaryotes, where incomplete or inconsistent annotations are common (Linard et al. 2021; Warren et al. 2010; Chaudhari et al. 2024; Lobb et al. 2020; Cantalapiedra et al. 2021b). This variability complicates cross-species comparisons of whole proteomes and experimental proteomic datasets, limiting the identification of functionally equivalent proteins and shared or species-specific responses to stimuli (UniProt Consortium 2025; Huerta-Cepas et al. 2019; Wattam et al. 2017). Orthology inference offers a robust strategy to address this limitation by identifying homologous, functionally related proteins across species. Tools such as OrthoFinder enable distinguishing conserved and species-specific gene sets through accurate orthology assignment and clustering of evolutionarily related proteins into orthogroups (Emms and Kelly 2015, 2019)**.** However, the use of such tools is often limited by technical barriers, including command-line execution, manual data preparation, and complex, hard-to-interpret outputs that limit accessibility for researchers without computational expertise.

To address these issues, we developed *OrthoGather*, a locally hosted web application streamlining comparative proteomic analysis. It retrieves reference proteomes from the UniProt database (UniProt Consortium, 2025), executes OrthoFinder locally, and generates browser-based, downloadable figures and summary statistics. Functional interpretation is supported by Gene Ontology (GO) enrichment, enabling analysis of user-defined protein sets in a comparative, cross-species context. Sharing GO annotations across orthogroups facilitates identifying differentially expressed pathways, even in poorly annotated species or strains. A typical application of *OrthoGather* is analysing protein sets derived from differential proteomics experiments, where changes in protein abundance are compared across conditions in distinct species. While orthogroup inference relies on reference proteomes, experiments generate subsets of proteins associated with specific responses, and their cross-species comparison enables the identification of conserved and species-specific functional patterns.

## Workflow and Output Overview

*OrthoGather* integrates comparative proteome analysis with GO enrichment, streamlining the workflow from data acquisition to interpretation (Fig. 1). An example application of this tool, based on a differential proteomics dataset derived from a study examining the response of *Mycobacterium smegmatis* (*Msm*) to sub-lethal concentrations of rifampicin (Giddey et al. 2017), is provided in the Supplementary File S1, and used throughout this manuscript.

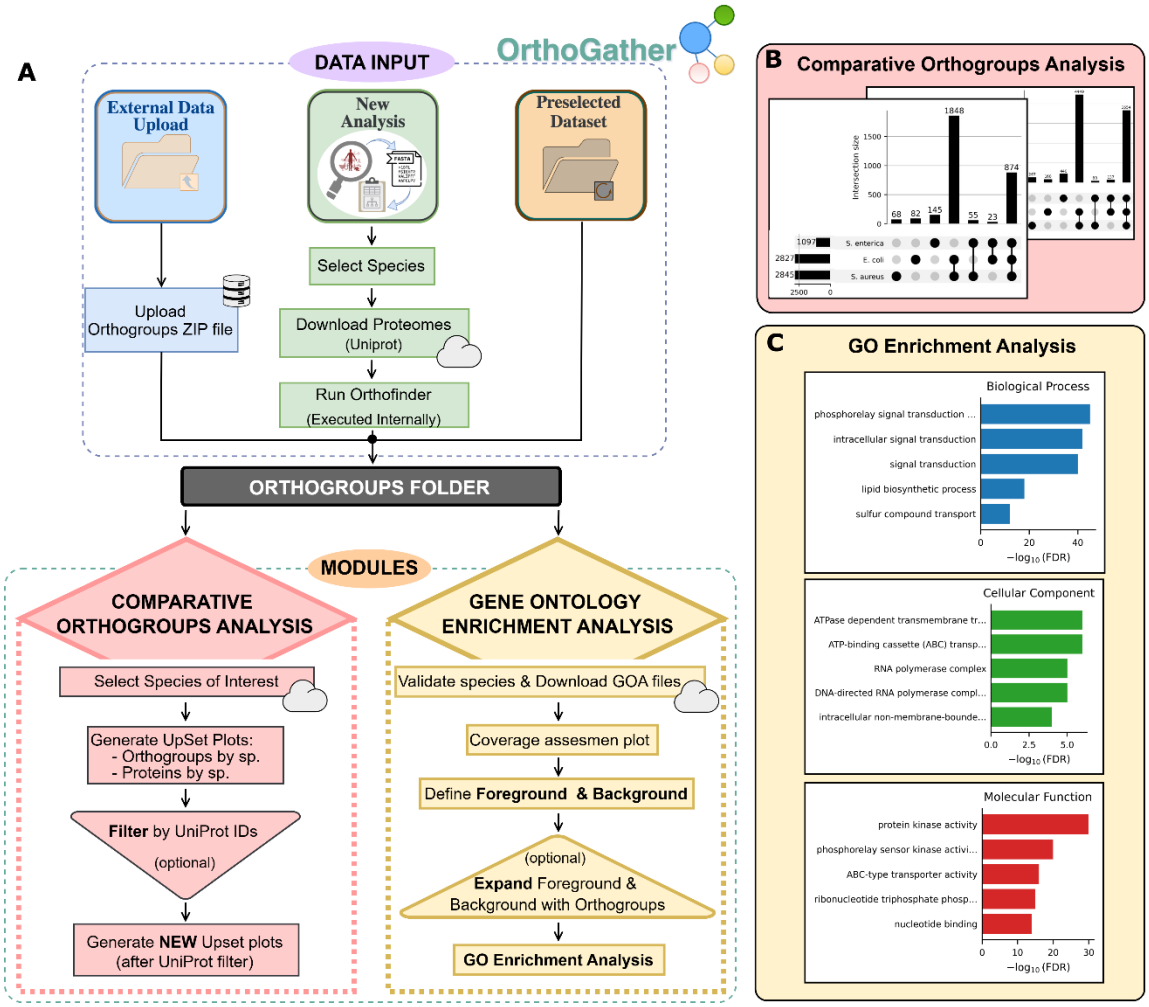


**Figure 1. Overview of the *OrthoGather* workflow and interface. (A)** *OrthoGather* workflow. All three input modes: *External Data Upload*, *New Analysis*, and *Preselected Dataset* are converged in the *Orthogroups* folder. The resulting data can be explored through two independent modules: *Comparative Orthogroup Analysis* and *Gene Ontology Enrichment*. **(B)** Representative UpSet plots summarising orthogroup overlap and protein distribution across selected species generated through the *Comparative Orthogroups Analysis* module **(C)** Representative bar plots showing enriched GO terms across Biological Process (top), Molecular Function (middle), and Cellular Component (bottom) generated through the *Gene Ontology enrichment analysis*. **Cloud**: data retrieved from external online resources; **Stacked cylinder**: external data file upload

**Three input modes** are supported by *OrthoGather* (Fig. 1A; Supplementary File S1, Fig. S1):

**(i)** *New Analysis*: A predictive search interface (species name or UniProt proteome ID) allowing to select proteomes available in the UniProt database (Supplementary File S1, Fig. S2). Selected reference proteomes are automatically downloaded in FASTA format, and *OrthoFinder* infers the orthogroups that form the basis for the subsequent analyses.

**(ii)** *Preselected Dataset*: An example dataset containing 47 proteomes from 17 bacterial species associated with cystic fibrosis lung infections and/or antimicrobial resistance is distributed with *OrthoGather* (Supplementary Materials S2, Table S1). It enables familiarisation with the platform without proteome downloads or external computations.

**(iii)** *External Data Upload***:** *OrthoFinder* results can be uploaded as a compressed archive containing the *Orthogroups* directory, enabling reuse of external analyses.

Regardless of the input mode, *OrthoGather* centralises processing in the *Orthogroups* folder and temporary and auxiliary OrthoFinder files are removed to optimise disk usage.

Subsequently, users can perform:

**1. Comparative Orthogroup Analysis**

This module allows exploring orthogroup composition across species using two successive filtering phases.

**(i)** *Species filtering phase*. A minimum of two species are selected from the dataset (Supplementary File S1, Figure S3). *OrthoGather* retains only orthogroups containing at least one protein from the selected species, excluding those formed exclusively by non-selected taxa. This step defines the analysis framework for the chosen species set. Two UpSet plots are generated (Fig. 1B): the first shows orthogroup distribution across species combinations; the second displays the number of proteins in each intersection, helping interpret overlaps and species contributions. To illustrate the workflow, we used the *New Analysis* input mode selecting proteomes of *Msm*, *Acinetobacter baumannii, Burkholderia cenocepacia, Escherichia coli, Mycobacterium abscessus,* and *Pseudomonas aeruginosa* (Supplementary File S1, Fig. S4-S5)

**(ii)** *Protein filtering phase*. The analysis can be further narrowed with a custom list of UniProt IDs (Supplementary File S1, Fig. S6). *OrthoGather* identifies orthogroups with at least one provided ID and generates new UpSet plots (Fig. 1B). To highlight this, the list of 596 proteins differentially expressed following *Msm* rifampicin treatment (Supplementary File S1,

129 Table S6) was used, identifying shared and species-specific orthogroups across species
130 (Supplementary File S1, Tab. S3, Fig. S7–S8. The data used to generate the plots can be
131 exported in Excel format. It contains UniProt identifiers of proteins belonging to each
132 orthogroup in each species, enabling the identification of homologous proteins, even when
133 their names differ.

134 **2. Gene Ontology Enrichment Analysis**

135 This module enables functional interpretation of the filtered and unfiltered orthogroups,
136 provided that the included species have annotations in the Gene Ontology Annotation (GOA)
137 repository (Huntley et al. 2015). The workflow consists of three main phases:

138 **(i)** *Species validation phase. OrthoGather* automatically checks which of the analysed
139 species have corresponding GOA files by comparing species names against a reference table
140 of taxon identifiers (Supplementary File S1, Fig. S9). Only species with available annotations
141 are retained, ensuring that the subsequent analysis and avoiding statistical biases introduced
142 by species without annotation. Manual entry of a taxon ID is also possible if required. In the
143 analysed example, all selected species except *B. cenocepacia* were automatically matched to
144 their corresponding GOA files without requiring manual taxon ID entry. *B. cenocepacia*
145 proteins did not contribute GO annotations to the further enrichment analysis, although the
146 species remained included in the orthogroup structure.

147 **(ii)** *Functional coverage phase.* For each orthogroup, *OrthoGather* quantifies the
148 proportion of proteins associated with at least one GO term relative to the total number of
149 proteins in the group. This provides a measure of the GO annotation completeness across the
150 dataset, generating quantitative information on orthogroup GO annotation coverage,
151 visualised as summary distributions, with downloadable data files. The distribution of
152 coverage values is displayed as a histogram and a boxplot, allowing overall annotation level to
153 be assessed before performing enrichment analysis. Low coverage suggests that additional
154 well-annotated species could be included to improve downstream results. In the analysed
155 example, *E. coli* was included due to its high-quality GO annotation, serving as a well-
156 annotated reference and resulting in a median GO annotation coverage of 70.52% across
157 orthogroups with at least one annotated protein (Supplementary File S1, Fig. S10).

158 **(iii)** *Enrichment phase.* Two protein sets must be defined: *foreground*, containing proteins
159 of interest to be evaluated for enrichment, and *background*, against which the foreground is

evaluated. The background can be defined either as a custom UniProt ID list or as all proteins from species with available GOA files. In both cases, the *Include Complete Orthogroups* option expands each set to include all proteins from orthogroups that contain at least one selected entry, enabling functional inference for unannotated proteins through annotated orthologs within the same orthogroup. For example, the *foreground* was defined as a set of 596 proteins differentially expressed in *Msm* following rifampicin exposure. The *background* represents the complete protein set detected in the experiment, e.g. 3,244 *Msm* proteins detected by mass spectrometry (Supplementary File S1, Fig. S11). The analysis is conducted using the GOATOOLS library (Klopfenstein et al. 2018), applying Fisher's exact test to determine whether a GO term occurs more frequently in the foreground than expected from the background. Multiple testing is performed (Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg,1995)). Significant results are displayed as bar plots (Fig. 1C), with adjustable analysis parameters including p-value threshold, ontology depth, and maximum number of displayed terms, enabling users to tailor the enrichment output to their needs. An Excel table provides detailed results for each enriched term, including background and foreground counts and the corresponding raw and FDR-adjusted p-values. In the analysed example a threshold of FDR<0.001, GO hierarchy depth level 2, and a maximum of 50 terms was used, highlighting processes associated with rifampicin resistance. (Supplementary File S1, Fig. S12–S13.

## Implementation

*OrthoGather* is developed in Python 3.7.12 and uses Flask 2.2.5 as the backend framework (Grinberg 2018)**,** with HTML and JavaScript for the frontend. Interaction is supported across all modern web browsers without additional installation. The application is organised into a modular architecture with four main components (Fig. 1):

**(i) Proteome selection**: Predictive UniProt proteome search (UniProt Consortium 2025), using a locally stored JSON index, followed by automatic FASTA download from the UniProt FTP service.

**(ii) Orthogroup inference**: OrthoFinder v3.0.1 (Emms and Kelly 2019, 2015; Emms et al. 2025) is executed internally via Python's subprocess module using default parameters; only the Orthogroups directory is retained to minimise storage usage.

**(iii) Orthogroup filtering and exploration**: Filtering by selected species and user-defined UniProt identifiers is supported, with visualisation via matplotlib 3.2.2(Hunter 2007) and UpSet plots (Lex et al. 2014), with export in CSV and Excel formats.

**(iv) Functional analysis using Gene Ontology**: GOA files are retrieved from the European Bioinformatics Institute (EBI) (Huntley et al. 2015). Foreground and background sets can be defined with optional expansion to all relevant orthogroups members. Enrichment analysis is performed using GOATOOLS 1.4.12 (Klopfenstein et al. 2018).

*OrthoGather* runs locally on macOS and Linux systems compatible with OrthoFinder and can be executed on Windows via virtualized environments or the Windows Subsystem for Linux. Resources, including installation instructions, documentation, and requirements, are available on GitHub. Internet access is required only to retrieve proteomes and GO annotations.

## Conclusion

This work introduces a novel tool that addresses a longstanding gap in the bioinformatics toolkit by providing an integrated and accessible platform for comparative and functional proteome analysis. *OrthoGather* empowers researchers with a robust, user-friendly framework that enables the complete workflow, from proteome selection to functional enrichment, to be executed locally and reproducibly without requiring advanced programming skills. Beyond the sample dataset presented, *OrthoGather* can be applied to a wide range of biological questions. This may include the analysis of differentially expressed proteins from proteomics experiments, the comparison of proteomic responses across species, and the exploration of conserved and species-specific functions in both well-annotated and poorly annotated organisms. In addition to proteomic analysis, it can be applied to transcriptomic datasets, aid whole genome sequencing analysis, and enable the identification of orthologs of specific proteins or genes across species. By leveraging orthology relationships, the platform enables functional inference in under-annotated species through the integration of information from well-annotated reference proteomes, facilitating hypothesis generation and guiding downstream experimental validation.

## Acknowledgements

The authors thank Dr Niamh Duggan and Ciarán Carey for testing *OrthoGather* with differential proteomics datasets during its development. The authors also thank Dr Metin Yazar and Narod Kebabci for their technical assistance. In accordance with the journal policy on the use of generative AI, the authors acknowledge the use of ChatGPT to assist with language editing and with debugging portions of the software code. All outputs were reviewed, tested, and validated by the authors, who take full responsibility for the content.

## Supplementary information:

Supplementary data is available at Bioinformatics online:

**S1_Example_dataset_analysis**. A full analysis of the example dataset with *OrthoGather*

**S2_Table_S1_Preselected_dataset.** List of reference proteomes available within *OrthoGather*

## Funding

This work was supported by the University College Dublin Conway Institute Director's Strategic Awards for Thematic Research 2023/2024. Colm Ryan and Olivier Dennler were funded by Research Ireland [20/FFP-P/8641] and Siobhán McClean and Joanna Drabinska were funded by Research Ireland [20/FFP-P/8717].

# References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society: Series B (Methodological) 57 (1): 289–300.

Cantalapiedra, Carlos P., Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021b. "EggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale." *Molecular Biology and Evolution* 38 (12): 5825–29.

Chaudhari, Jyoti Kant, Shubham Pant, Richa Jha, Rajesh Kumar Pathak, and Dev Bukhsh Singh. 2024. "Biological Big-Data Sources, Problems of Storage, Computational Issues, and Applications: A Comprehensive Review." *Knowledge and Information Systems* 66 (6): 3159–3209.

Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (1): 157.

Emms DM, Kelly S. 2019. "OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics." *Genome Biology* 20 (1): 238.

Emms, David M., Yi Liu, Laurence J. Belcher, Jonathan Holmes, and Steven Kelly. 2025. “OrthoFinder: Scalable Phylogenetic Orthology Inference for Comparative Genomics.” *BioRxiv*. https://doi.org/10.1101/2025.07.15.664860.

Giddey, Alexander D., Elise de Kock, Kehilwe C. Nakedi, Shaun Garnett, Andrew J. M. Nel, Nelson C. Soares, and Jonathan M. Blackburn. 2017. “A Temporal Proteome Dynamics Study Reveals the Molecular Basis of Induced Phenotypic Resistance in Mycobacterium Smegmatis at Sub-Lethal Rifampicin Concentrations.” *Scientific Reports* 7 (1). https://doi.org/10.1038/srep43858.

Grinberg, M. 2018. *Flask Web Development: Developing Web Applications Python*. O’Reilly Media.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “EggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14.

Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95.

Huntley, Rachael P., Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O’Donovan. 2015. “The GOA Database: Gene Ontology Annotation Updates for 2015.” *Nucleic Acids Research* 43 (Database issue): D1057-63.

Klopfenstein, D. V., Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, et al. 2018. “GOATOOLS: A Python Library for Gene Ontology Analyses.” *Scientific Reports* 8 (1). https://doi.org/10.1038/s41598-018-28948-z.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. “UpSet: Visualization of Intersecting Sets.” *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1983–92.

Linard, Benjamin, Ingo Ebersberger, Shawn E. McGlynn, Natasha Glover, Tomohiro Mochizuki, Mateus Patricio, Odile Lecompte, et al. 2021. “Ten Years of Collaborative Progress in the Quest for Orthologs.” *Molecular Biology and Evolution* 38 (8): 3033–45.

Lobb, Briallen, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, and Andrew C. Doxey. 2020. “An Assessment of Genome Annotation Coverage across the Bacterial Tree of Life.” *Microbial Genomics* 6 (3). https://doi.org/10.1099/mgen.0.000341.

UniProt Consortium. 2025. “UniProt: The Universal Protein Knowledgebase in 2025.” *Nucleic Acids Research* 53 (D1): D609–17.

Warren, Andrew S., Jeremy Archuleta, Wu-Chun Feng, and João Carlos Setubal. 2010. “Missing Genes in the Annotation of Prokaryotic Genomes.” *BMC Bioinformatics* 11 (1): 131.

Wattam, Alice R., James J. Davis, Rida Assaf, Sébastien Boisvert, Thomas Brettin, Christopher Bun, Neal Conrad, et al. 2017. “Improvements to PATRIC, the All-Bacterial Bioinformatics Database and Analysis Resource Center.” *Nucleic Acids Research* 45 (D1): D535–42.