communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08651-2

FANTASIA leverages language models to decode the functional dark proteome across the animal tree of life

Gemma I. Martínez-Redondo[1,2] ✉, Francisco M. Perez-Canales[3], Belén Carbonetto[1], José M. Fernández[4,5], Israel Barrios-Núñez[3], Marçal Vázquez-Valls[1], Ildefonso Cases[3], Ana M. Rojas[3,6] ✉ & Rosa Fernández[1,6] ✉

Protein functional annotation is crucial in biology, but many protein-coding genes remain uncharacterized, especially in non-model organisms. FANTASIA (Functional ANnoTAtion based on embedding space SImilArity) integrates protein language models for large-scale functional annotation. Applied to ~1000 animal proteomes, FANTASIA predicts functions to virtually all proteins, including up to 50% that remained unannotated by traditional homology-based methods. This enables the discovery of novel gene functions, enhancing our understanding of molecular evolution and organismal biology. FANTASIA holds particular promise for functional discovery in non-model taxa, offering advantages over homology-based tools in sensitivity and generalizability. FANTASIA is available on GitHub at https://github.com/CBBIO/FANTASIA.

Understanding protein function is essential for deciphering genome evolution and the complexity of life. Gene Ontology (GO) provides a standardized vocabulary to describe gene functions across species[1], organizing terms into three categories: biological process, molecular function, and cellular component. This framework is crucial for interpreting genome-wide data, yet a substantial portion of protein-coding genes, particularly in non-model organisms, but also in model organisms to a lesser extent, remain unannotated, forming what we call the "dark proteome".

Experimental knowledge of protein function remains extremely limited, with fewer than 0.5% of proteins having been functionally characterized to date[2]. To bridge this gap, traditional annotation methods[3,4] primarily rely on function transference via sequence similarity, which limits their effectiveness for annotating highly divergent or fast-evolving proteins. Many genes remain undetected or uncharacterized due to rapid evolution, orphan status, or intrinsic disorder, making their functional classification particularly challenging. For example, approximately 30% of proteins in the model organism *Caenorhabditis elegans* lack a functional annotation in UniProt[5]. This problem is even more pronounced in non-model species, where 41% of genes in a tardigrade[6] and 50% in a sponge[7] remain unannotated. Moreover, orphan genes, which lack detectable homologs, can comprise up to one-third of predicted genes in eukaryote genomes[8].

The challenge posed by the dark proteome is particularly pressing, as incomplete or inaccurate functional annotations can lead to biased biological interpretations, impairing enrichment analyses, network inference, and comparative genomics. Poorly characterized genes may also be linked to important evolutionary innovations, and their omission from analyses can result in incomplete or misleading models of evolutionary change, limiting the ability to identify conserved or lineage-specific features. These limitations hinder biological discovery and slow the pace of basic and applied research. The urgency is illustrated by the rapid expansion of genomic data from large-scale initiatives such as the Earth BioGenome Project[9] and the European Reference Genome Atlas (ERGA)[10,11]. These efforts are generating unprecedented volumes of genomic information, highlighting the critical need for reliable functional annotation frameworks that go beyond conventional sequence-similarity-based approaches.

Incorporating artificial intelligence (AI) techniques has enhanced the performance of prediction methods[12]. However, despite the fast evolving pace in this area, there is still room for improvement, especially when applied to full proteomes from unconventional organisms. Further developments benchmarked in CAFA datasets[12], ranging from convolutional neural networks (CNN) deep learning models[13] to transformer-based protein language models (pLMs)[2,14], offer a promising alternative for decoding the 'hidden biology' within the dark proteome. Notably, there is little

[1]Metazoa Phylogenomics and Genome Evolution Lab, Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain. [2]Universitat de Barcelona, Barcelona, Spain. [3]Computational Biology and Bioinformatics, Andalusian Center for Developmental Biology (CABD-CSIC), Sevilla, Spain. [4]Barcelona Supercomputing Center, Plaça d'Eusebi Güell, Barcelona, Spain. [5]Spanish National Bioinformatics Institute (INB)/ ELIXIR ES Coordination Node, Barcelona Supercomputing Center (BSC), Barcelona, Spain. [6]These authors contributed equally: Ana M. Rojas, Rosa Fernández. ✉e-mail: gemma.martinez@ibe.upf-csic.es; a.rojas.m@csic.es; rosa.fernandez@ibe.upf-csic.es

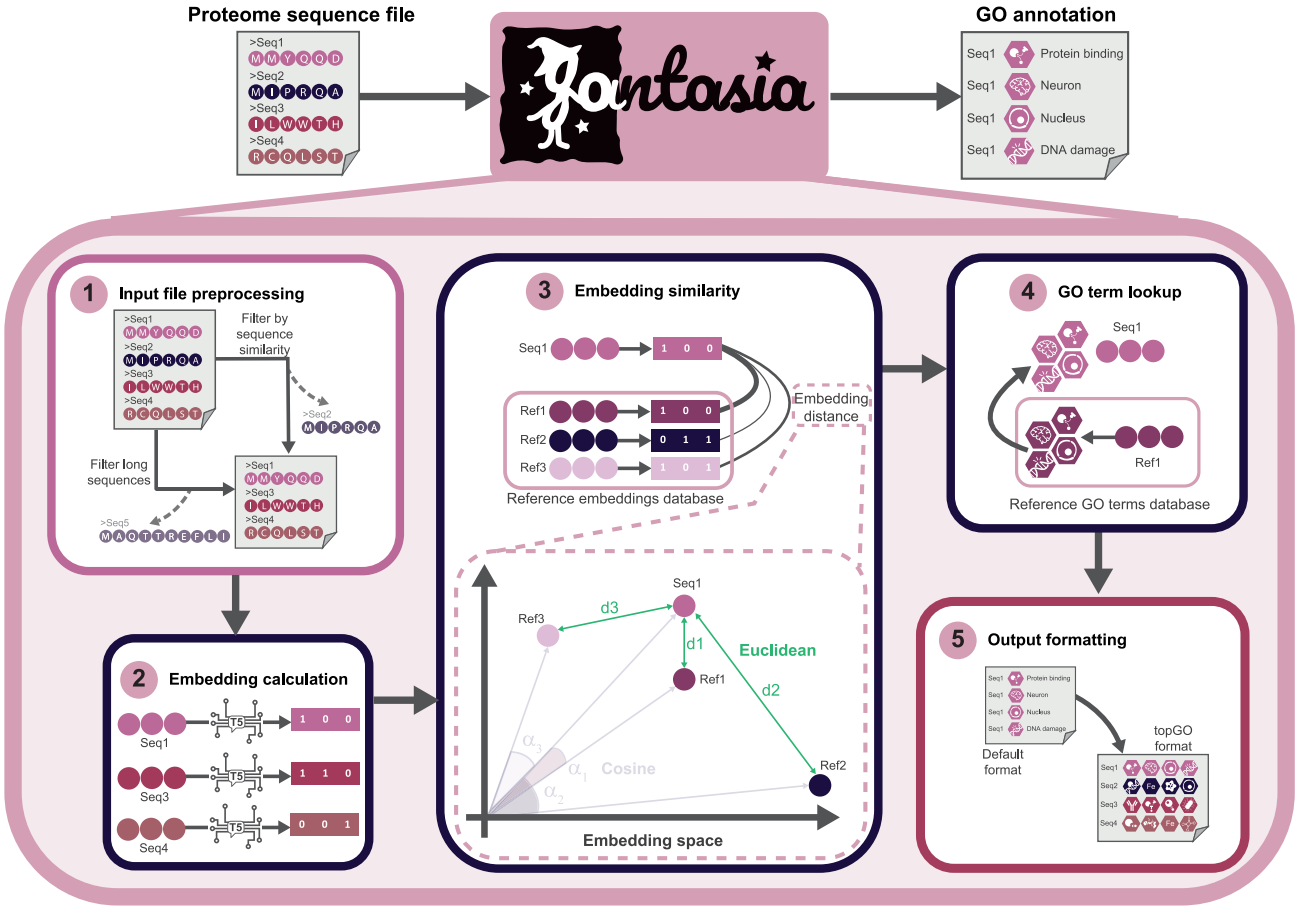

Fig. 1 | FANTASIA pipeline overview. Input proteomes are preprocessed to remove sequences based on length and sequence similarity if needed. Then, embeddings are computed, and distance embedding similarity is calculated against the reference database (using two metrics at will). Optionally, it converts the standard GOPredSim output file to the input file format for topGO[20] to facilitate its application in a wider biological workflow.

correlation between pLM-derived embeddings and raw sequence similarity, suggesting that these models capture functional and structural signals beyond traditional homology[2,14].

Results and discussion

FANTASIA: one pipeline to annotate them all

Since available tools that use pLMs are either hard to escalate to complete proteomes, install or customize for benchmarking or testing different models, here, we introduce FANTASIA (*Functional ANnoTAtion based on embedding space SImilArity*), a flexible pipeline that uses pretrained pLMs and embedding similarity searches for large-scale functional annotation (Fig. 1). FANTASIA is an improved reimplementation of the GOPredSim algorithm[2] focused on its genome-wide implementation and stability from scratch, together with necessary steps for data preprocessing. FANTASIA's zero-shot capabilities, where functional annotation can be performed without task-specific fine-tuning, represent a significant advancement. By integrating an efficient implementation, parallel computing techniques, and a streamlined workflow, it significantly improves user-friendliness while maintaining its already tested[2,5] prediction accuracy. First, given an input proteome or query proteins, it computes protein embeddings using different models, like ProtT5[14] or ESM2[15]. The list of proteins can be previously filtered by length or by sequence similarity, to remove identical sequences within the input or close sequences in the references that can influence results during benchmarking. Then, it employs a robust embedding similarity-based approach to infer GO terms, accessing a on-the-fly generated database of embeddings with functional annotations from different GOA versions. Next, it produces functional predictions using the closest hits or distance-based filtering methods to reduce noise and ensure robustness. Additionally, FANTASIA provides a flexible command-line interface, enabling seamless integration into diverse bioinformatics pipelines. See 'Methods' for more details.

Comparison with other methods

We previously benchmarked those methods on model organisms using the two available pLMs (SeqVec and ProtT5)[5] against two CNN-based models and a sequence profile-based method. We showed that transformers (SeqVec and ProtT5) predict more precise and informative GO terms than CNN-based methods (DeepGO and DeepGO+), and have a higher coverage-proportion of annotated proteins in the proteome- than traditional homology-based methods, making them suitable for large-scale annotation[5]. Our goal in this work was to assess the feasibility of zero-shot annotation transfer using broadly accessible models. We specifically showed that transformers increased annotation coverage (number of sequences annotated), term inference precision, and informativeness (or detail), while being able to recover accurate functional information from transcriptomic experiments[5]. Since prot-T5 was the best performer in our benchmark, it was selected to expand the analyses to 970 species (~23 million genes) covering the full animal diversity. As non-model organisms lack "ground" annotation standards, it is not possible to conduct a proper benchmark as the previous one.

Nevertheless, repeating some comparative analyses yielded the same results in terms of informativeness (Supplementary Note 1, and Supplemetnary Fig. 1) and annotation coverage, where traditional homology-based methods failed to annotate nearly half of the genes, particularly in less-studied phyla (Fig. 2a, Supplementary Note 2, and

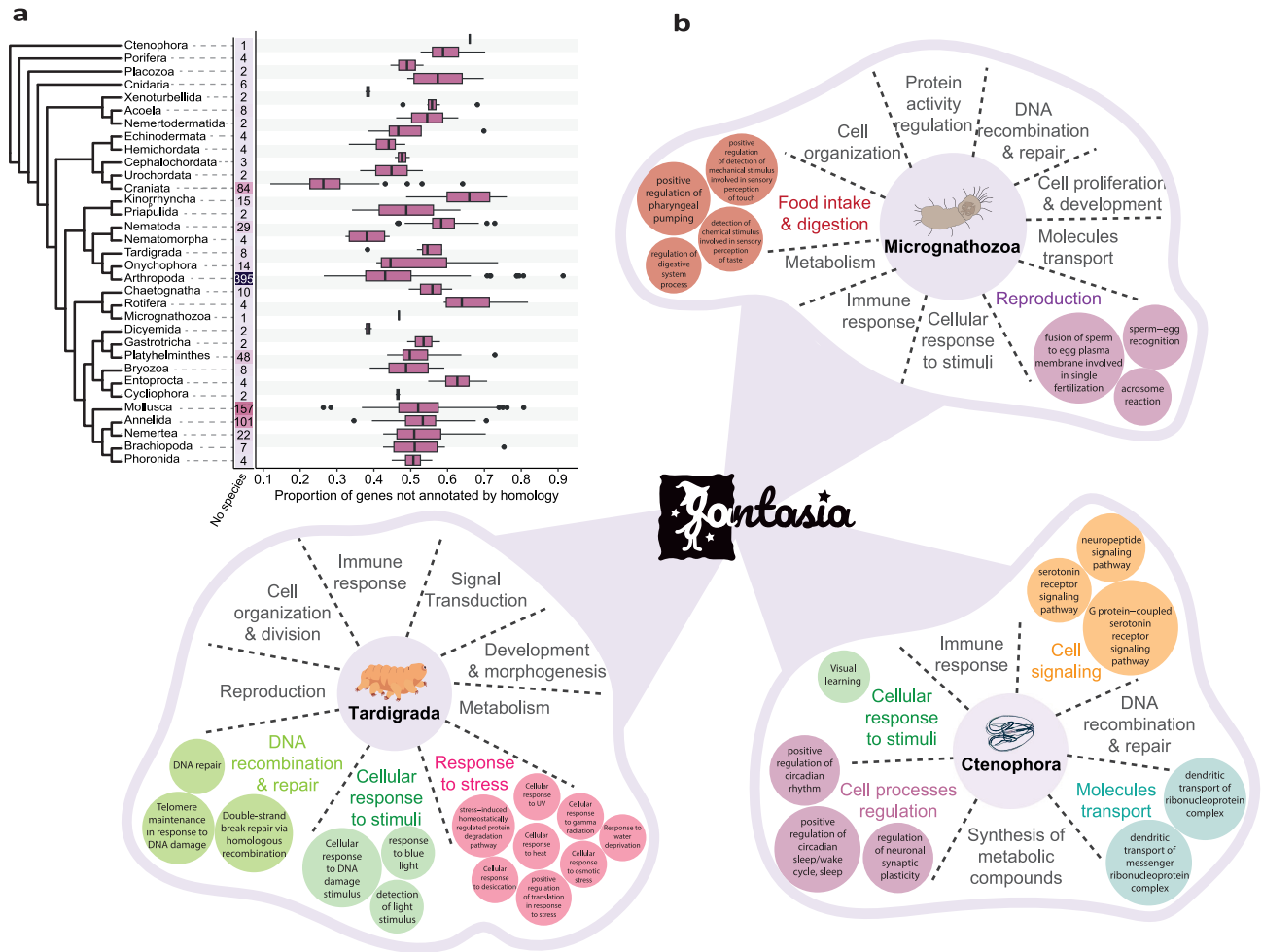

Fig. 2 | Functional annotation of the dark proteome across the Animal Tree of Life. a Number of animal species per phylum included in this study (left) and the proportion of genes not annotated by homology per phylum (right). **b** Enriched 'hidden' biological terms revealed by FANTASIA in three animal phyla in dowstream analyses. Large text summarizes key functions, while colored circles represent specific GO terms. All artwork was designed by Gemma I. Martínez-Redondo.

Supplementary Fig. 2). Moreover, we show that FANTASIA, in agreement with the GOpredsim base method, recovers less GO terms per gene (reducing the noise for downstream analyses, Supplementary Note 3, and Supplementary Fig. 3). It also recovers a moderately similar functional annotation to the homology-based one, explained by the discrepancy in the number of terms yielded by each method (Supplementary Note 4, and Supplementary Fig. 4), and exhibits minimal differences between using all isoforms (and combining all annotations) or just the longest (Supplementary Note 5, and Supplementary Fig. 5). In both methods, gene-level annotations remained highly similar, supporting the common practice of using the longest isoform in evolutionary studies while reducing computational costs.

Revealing the 'hidden biology' in the dark proteome across the animal tree of life

To show how FANTASIA can access functionally under-characterised biological information, we performed a per phylum GO enrichment analysis of the genes that were not annotated using homology-based methods. While 34.43% of GO terms were phylum-specific, 48 were common to all phyla (Supplementary Note 6, and Supplementary Fig. 6), mainly linked to viral response and immune function, both under adaptive evolution in animals. Newly annotated phylum-specific functions are not restricted to a specific group of functions but span diverse biological roles, including essential animal processes (Supplementary Note 7 and Supplementary Fig. 7). However, per-phylum GO enrichment analyses also revealed functions previously undetected by homology-based approaches, linked to unique biological traits. For instance, tardigrades showed stress-related functions that might help us understand their resilience; Micrognathozoa displayed functions tied to the pharynx and unexpected reproductive terms, hinting at undiscovered males or reproductive genes co-option; and ctenophores had newly annotated neuronal functions, shedding light on their unique nervous system (Fig. 2b, Supplementary Note 8, and Supplementary Fig. 8).

Virtues and challenges

We have shown that FANTASIA is a powerful annotation tool, capable of uncovering meaningful biological insights from 'hidden' biology by interpreting the language of proteins. Its zero-shot prediction capabilities allow for strong generalization across diverse taxa, including those under-represented during training. By leveraging pre-trained models, FANTASIA also aligns with sustainable AI principles. A key limitation, shared by all transformer-based models, is their sensitivity to sequence length, both computationally and biologically restrictive. To mitigate this, FANTASIA v2 was designed as a modular, scalable, and extensible framework, enabling the easy integration of new models as they emerge. While FANTASIA annotates proteins independently of genome context, its performance may still be affected by low-quality or incomplete input data—an inherent challenge shared with other sequence-based methods.

In the era of genomics, pLMs offer a promising complement to homology-based approaches for inferring gene functions in both model and non-model organisms. Here, we demonstrate how these models can highlight functions that are not captured by conventional annotation methods across animals, and present FANTASIA as a reliable tool for inferring function in the coding region of newly-assembled genomes. After all, important biological functions can be found, even in the 'darkest' genes, if one has a way to turn on the light.

Methods

FANTASIA pipeline step-by-step

FANTASIA is a reimplementation of the original GOPredSim algorithm[2] for functionally annotating full proteomes based on GO term transference from protein embedding similarity. Building on the original method, our pipeline enhances usability by offering scalability to full proteomes, a more reliable installation process with minimized dependency conflicts, and an intuitive command line interface that simplifies parameter customization to facilitate usability. We provide two implementations of FANTASIA to accommodate different user needs. The first, FANTASIA v1, packaged within a Singularity container, was designed for a rapid and straightforward use when unconventional organisms (absent in GOA reference). This version presented challenges with dependency management. Based on user feedback, which also suggested increase flexibility to accomodate benchmarking approaches and more models, we developed a second version (FANTASIA v2) that uses a vector database for improved stability, traceability and ease of deployment. The full list of differences between both versions is explained in the Supplementary Note 9.

The FANTASIA pipeline consists mainly of five steps: input preprocessing, protein embeddings computation, calculation of embedding similarity against an embedding reference database that contains functional annotations (the Gene Ontology Annotation database), GO terms transference from closest reference sequences' embeddings, and output file formatting (Fig. 1).

Step 1: input file preprocessing. FANTASIA allows the removal of long sequences and the application of a sequence similarity filtering[16–18]. In FANTASIA v1, this preprocess is mandatory, ensuring that the pipeline can run smoothly by removing identical sequences in the input file and filtering out sequences longer than 5000 amino acids. In contrast, FANTASIA v2 makes these steps optional and more flexible, allowing users to define maximum length and set a minimum sequence identity threshold (maximum redundancy allowed is 50%). Unlike FANTASIA v1, the sequence similarity filter in FANTASIA v2 is conducted against the annotated reference database using mmseq2[18], and is only recommended for benchmarking purposes (e.g., to evaluate performance on mouse proteins, mouse or closely related sequences must be excluded from the look-up table). For functional annotation of non-model organisms, we do not recommend filtering the look-up table, as retaining as much reference information as possible improves annotation coverage. Input sequence redundancy (not included in FANTASIA v2) can be reduced to save computation time, but this does not impact the method, as annotations are generated per protein. If the minimum sequence identity filter is enabled, FANTASIA v2 concatenates the input dataset with the reference table before computing distances between embeddings (step 3 below). Then, for each query embedding, a comparison is performed against the entire reference table. During this process, embeddings corresponding to sequences that belong to the same cluster as the query are excluded, ensuring that no matches (hits) are made with proteins that exceed the specified sequence identity threshold.

On the other hand, long sequences pose both computational and biological challenges. Transformer models scale poorly with length, often requiring splitting, which can disrupt long-range interaction signals. Biologically, key features like domain boundaries or interaction sites may span across fragments, leading to diluted or misrepresented functional signals. Thus, while splitting long proteins may be necessary, it can compromise representation integrity and annotation accuracy. UniProtKB excludes sequences over 12,000 residues for embedding generation due to GPU limitations, affecting only a small number of proteins (https://www.uniprot.org/help/embeddings). The same is true for FANTASIA v1, where the upper limit of 5000 in length reduces the computational resources needed, while avoiding issues derived with the embedding interpretation for the SeqVec model we tested in our previous benchmark[5], which we did not finally include in any of the current versions due to its lower performance.

Step 2: embedding computation. FANTASIA computes the protein embeddings per model and per sequence, so it is agnostic to the proteome size. FANTASIA v1 only supports ProtT5[14] (Rostlab/prot_t5_xl_uniref50, best model previously benchmarked in ref.[5]) and uses bioembeddings, while FANTASIA v2 also supports ESM2 (facebook/esm2_t6_8M_UR50D)[15] and ProstT5 (Rostlab/ProstT5)[19] to enable addressing of different questions. Although ESM2_t6_8M_UR50D is known to capture evolutionary relationships across species, it has been trained on relatively small datasets. On the other hand, Prot-T5, being trained in the Big Fantastic Database (BFD)[14] excels in handling highly variable sequences or proteins with complex domain structures and provides zero-shot capabilities. ProstT5, as a fine-tuned model, is very useful to address protein structure information. Since bioembeddings trigger many dependencies, the calculation of embeddings in FANTASIA v2 is implemented as in (https://huggingface.co/). In both cases, embeddings are stored in HDF5 format for further use. As in the original GOPredSim method, we used the final hidden layer of each model and applied mean pooling over the sequence dimension to obtain the protein-level embeddings.

Step 3: embedding similarity. FANTASIA then computes the distance between each input sequence embedding (query) and the embeddings in the reference database (look-up table). The reference database in FANTASIA contains, for each reference protein, its pre-computed embeddings for the supported pLMs, and the GO term as in GOA, annotations that will be transferred in step 4 (see below). In FANTASIA v2, this information is stored as a reference vector database managed with PostgreSQL which also contains, for each reference protein, its sequence and metadata for further analyses.

By default, Euclidean distance (d_e) between embeddings n and m for a model with an embedding of embedding size s is computed with the following formula:

$$d_e(n,m) = \sqrt{\sum_{i=1}^{s}(n_i - m_i)^2}$$

Where s represents the number of dimensions in the embedding space, which varies depending on the selected pLM: $s = 1024$ for ProtT5 and ProstT5, and $s = 320$ for ESM2. These sizes are architectural choices.

Alternatively, FANTASIA v2 allows the selection of cosine similarity (d_c), which is calculated using the formula:

$$d_c(n,m) = \frac{\sum_{i=1}^{s} n_i m_i}{\sqrt{\sum_{i=1}^{s} {n_i}^2} \cdot \sqrt{\sum_{i=1}^{s} {n_i}^2}}$$

In FANTASIA v2, batching can be changed; however, we keep the original settings to ensure robustness and compatibility with the original use.

In FANTASIA v1, we used Euclidean distance by default to align with the original GOPredSim. In v2, we added support for cosine similarity, which better captures directional similarity and is less sensitive to vector magnitude, an advantage when comparing embeddings across models with different scaling. This allows users to explore alternative distance metrics depending on their annotation needs.

Step 4: GO transference. FANTASIA then transfers GO terms from the k closest embedding-bearing proteins in the database. By default, only the closest hit ($k = 1$) is used. In FANTASIA v2, users can define a distance threshold for each model that determines the maximum allowed distance between query and reference embeddings, which requires user optimization. As reference datasets, FANTASIA v1 uses GO terms from the March 22nd, 2022 release (GOA2022), while FANTASIA v2 uses the current UniprotKB API, (as of March 10th, 2025, UniProtKB uses version 2025 of GOA UniProt. This version was assembled using publicly available data from source databases on March 3, 2025), this guarantees that FANTASIA always use the latest GOA version, although older versions can also be used. Evidence codes are CAFA-recommended experimental evidence codes (e.g., EXP, IDA, IPI, IMP, IGI, IEP, TAS).

Step 5: output description and formatting. The standard output of FANTASIA gives the user information on the GO terms assigned to each input protein, along with the associated reliability index (see below). In FANTASIA v1, the output is a tab-separated file where each row represents a GO-input protein pair, with columns for the sequence accession, the GO term identifier, and the reliability index. In FANTASIA v2, the output is a comma-separated file (CSV) with 10 columns: 1) sequence accession (header); 2) GO term identifier; 3) GO category (F: molecular function, P: biological process, C: cellular component); 4) GOA evidence code for the reference annotation; 5) GO term description; 6) embedding distance (Euclidean or cosine) between the query and the reference; 7) protein language model used for the embedding generation; 8) UniProt ID of the reference protein bearing the closest embedding; 9) organism the target reference protein belongs to; 10) reliability index. Additional columns provide information regarding the sequence similarity of the query vs. the closest-embedding-bearing reference sequence, as calculated by EMBOSS-based parasail aligner (https://github.com/jeffdaily/parasail/blob/master/apps/README.md). It also provides information about GO term traceability (support, collapsed terms, etc). By default, FANTASIA converts this standard output to the input file format for topGO's GO enrichment[20] to facilitate its application in a wider biological workflow in functional genomics. This feature can be disabled by the user in FANTASIA v2.

The reliability index (RI) was previously established in ref.[2], and is a transformation of the distance between the query q and hit in the database (q,n_i) into a similarity scale, making it easier to compare between functional annotation results. Supported RI formulations:

For Euclidean distance (d_e), RI is defined using the inverse similarity transformation as:

$$RI = \frac{0.5}{0.5 + d_e(q, n_i)}$$

Where lower Euclidean distances yield higher confidence scores.

For cosine similarity (d_c), RI is computed with the direct similarity measure:

$$RI = 1 - d_c(q, n_i)$$

While both formulations produce values ranging from 0 to 1, they are not directly comparable, as they capture confidence in different ways. Users should exercise caution when interpreting RI scores across different similarity metrics. Additionally, the Euclidean distance is not inherently comparable across different pLMs, as it depends on the magnitude of the embedding vectors generated by each model. In contrast, the cosine similarity metric is more suitable for cross-model comparisons, as it primarily captures the relative orientation of embeddings rather than their absolute magnitude.

Importantly, RI values do not necessarily correlate with prediction accuracy. In model organisms, we observed that true annotations (i.e., exact GO term matches) and biologically relevant functions can still be recovered within the second quantile of the RI distribution[5]. In the context of non-model organisms, where no ground truth is available, RI thresholds cannot be optimized. Therefore, we recommend using RI as a general indicator of high or low support, rather than as a strict measure of accuracy.

FANTASIA can be executed on both GPUs or CPUs, depending on the available hardware, though GPU-based execution is recommended for improved speed (Supplementary Fig. 9b). The computational complexity of the pipeline is O(n), where n is the number of sequences in the input dataset. Both versions perform very similarly on their speed (See Supplementary Note 9, and Supplementary Fig. 9).

Input datasets and processing

We obtained the proteome files for 961 animals and 9 outgroup species from MATEdb2[21] (Supplementary Data 1). This dataset comprises the longest peptide sequence (or isoform) per gene, totaling 23,184,398 sequences, derived from high-quality genomes and transcriptomes representing nearly all animal phyla. To assess the impact of annotating all isoforms versus only the longest, we created a subset of 93 species (Supplementary Data 1) to maximize representation across the Animal Tree of Life.

We annotated the functions (assigned GO terms) of the longest and all isoforms datasets using (1) FANTASIA v1 with the ProtT5 model and (2) eggNOG-mapper as a proxy for annotation based on homology. Euclidean distance was selected for consistency with previous benchmarks, where it was reported that cosine similarity yielded similar results[2]. We used eggNOG-mapper 2.1.6[3] as described in MATEdb2[21]. To ensure comparability, we used the datasets filtered by length and by removal of identical sequences, as provided by FANTASIA v1, as input for eggNOG-mapper. Homology-based results were then filtered to remove the GO terms that have a depth in the GO graph (DAG) lower than 2 (i.e. the terms are very general, close to or at the root of the DAG, and not informative). The difference in annotation coverage between the filtered and unfiltered homology-based annotations is negligible (Supplementary Fig. 2, light purple).

When exploring the effect of using all isoforms for functional annotation, we first annotated all isoform sequences and then obtained per-gene annotations by collapsing the annotation of each isoform to retrieve a list of unique GOs.

In addition, we performed an extra RI-based filtering for FANTASIA results (see above for an explanation on RI). Thresholds for filtering were obtained from the source GOPredSim publication[2] based on evaluations using CAFA3[12]: 0.35 for biological process, 0.28 for molecular function, and 0.29 for cellular component. It should be noted that these values were estimated based on the CAFA3[12] F1 max comparisons datasets which do not necessarily scale to longer and more diverse datasets, as we have previously shown in model organisms[5] where thresholds require to be optimised.

Information content

To analyse GO terms according to depth within the GO graph, we estimated their Information Content (IC) using a custom script (https://github.com/cbbio/Compute_IC) which is defined as[22]:

$$IC = -log_2(p(t))$$

Where p(t), the probability of a term occurring in a dataset, is calculated as:

$$p(t) = 1 - \frac{'Offspring\ Count'}{'Offspring\ Count' + 'Ancestors\ Count'}$$

Offspring and ancestors count refer to the number of proteins that are assigned to that term, given a reference Gene Product Association Data file (GAPD). In this case, the reference GAPD, or universe of proteins, combined all GAPD files for all species available.

Higher IC implies a more specific GO term while low IC values mean that the GO term is more general.

Semantic similarity

To assess how similar two different annotations are, we obtained the lists of genes that were commonly annotated by both methods for the representative subset of species representing all phyla ($n = 93$) and calculated for each gene their semantic similarity between the GO terms inferred independently for each of the 3 GO term categories. To be more conservative in our analysis, we used the most filtered dataset for each method (i.e., filtering by reliability index). We used pygosemsim (https://github.com/mojaie/pygosemsim) to estimate the Wang semantic similarity[23] between protein-coding gene annotations using the Best-Match Average (BMA) method for gene products annotated with multiple GO terms. Semantic similarity (sim_{BMA}) between 2 gene products ($g1$ and $g2$) is calculated as follows:

$$sim_{BMA}(g1,g2) = \frac{\sum_{1=i}^{m} \max_{1 \le j \le n} sim(go1i, go2j) + \sum_{1=j}^{n} \max_{1 \le j \le m} sim(go1i, go2j)}{m+n}$$

where m and n are the number of GO terms annotated in $g1$ and $g2$, respectively; and $go1i$ and $go2j$ are the GO term in position i (from 1 to m) from the $g1$ and the GO term in position j (from 1 to n) from the $g2$, respectively. sim_{BMA} Semantic similarity values range from 0 to 1, with 0 meaning no similarity and 1 meaning (almost) identical.

The Wang similarity method, unlike other metrics, is based on the GO DAG topology, making it more stable than others based on information content that rely on the available annotations, although it can lead to information loss. However, in the case of non-model organisms, the biological context can be highly variable. In such uncertain settings, we find that Best-Match Average (BMA) offers a balanced compromise: it reduces noise while retaining meaningful, context-relevant information by accounting for multiple matches. Although all similarity aggregation methods (such as maximum or average) have their limitations, BMA provides a practical middle ground when the precise relationship between annotations is unknown.

GO enrichment for genes not annotated by homology-based methods

We obtained the list of genes that were not annotated by homology for each species and performed a GO enrichment analysis using topGO[20] against the whole set of proteins of that species to identify which functions were being missed by homology. We used the GO terms predicted using FANTASIA. TopGO, in contrast to other GO enrichment methods, is based on the GO DAG, allowing us to take relationships between GO terms into account.

From the GO enrichment per species, we obtained a list of GO terms enriched per phylum by combining the results per species. For that, in the case of GO terms that were enriched in at least one species, we combined the p-values using the ordmeta method[24].

We then identified the list of GO terms that were commonly enriched in all animal phyla. Among all common GO terms (Supplementary Table 1), we focused on toxin activity (GO:0090729) to assess the performance of the functional annotation. Among all animals with venom, we selected the scorpion *Centruroides sculpturatus* to look into the genes with this GO term that were not annotated by homology. We used ToxinPred2[25] to further filter this list and identified CSCU1_DN0_c0_g23215_i1.p1 (XP_023242874.1, uncharacterized protein) as a putative homolog of tachylectin-5 (XP_023224331.1, 61.20% of sequence similarity). We further assessed this by predicting their structure with ColabFold[26] and comparing them with the experimentally-determined structure of the horseshoe crab *Tachypleus tridentatus* tachylectin-5A (PDB 1JC9) using the PDB pairwise structure alignment tool (TM-align)[27]. These results can be found in Supplementary Fig. 6.

Data visualization

To visualize the significantly enriched GO terms per phylum and GO category, we first calculated the Wang semantic similarity matrix for each dataset and compared several clustering algorithms using the simplifyEnrichment R package[28], and proceeded with the K-means algorithm (Supplementary Data 2). Then, to summarize the function of each cluster, we obtained their representative GO term by selecting the GO term with the lowest closeness value (the 'farthest') in the subset semantic similarity network for that cluster. Results were plotted in a TreeMap (Supplementary Data 3).

We then combined the results of all phyla by assigning each significant GO term to the GO slim AGR (Alliance of Genomics Resources) subset using the map-to-slim.py script available in GOAtools[29]. Between 10 and 60% of GO terms per phylum were not assigned to a category and were further clustered as explained above. The GO representative terms of each of the 49 clusters were later manually assigned to the GO Slim AGR classification and used to include unassigned GOs in the GO-GO category list. Final results were normalized per phylum (row) and visualized in a heatmap.

Statistics and reproducibility

Statistical significance for GO enrichment was assessed using the elim algorithm in the topGO[20] package, which accounts for the hierarchical structure of the GO graph. An alpha threshold of 0.05 was applied to identify significant terms. Because elim computes p-values conditioned on the structure of neighboring GO terms, and GO terms are, therefore, not independent, multiple testing is not applicable. Enrichment results across species were aggregated at the phylum level by combining independent p-values using the ordmeta method[24], a meta-analysis approach based on the joint distribution of ordered p-values.

GO knowledgebase is regularly updated with new releases that remove obsolete GO terms or add new ones, changing in addition the structure of the graph, which hinders comparison between analyses that use different versions[30]. Hence, to ensure consistency and comparability, we used the same Gene Ontology (GO) release version across analyses whenever possible. Specifically, we used GOA2022 (March 22, 2022) for all FANTASIA v1 annotations and GOslim subsets, and the June 11, 2023 release for information content and semantic similarity analyses.

Computational resources

FANTASIA was tested in several computing environments. The performance of FANTASIA v1 was assessed in the Finisterrae III HPC (FT3, Centro de Supercomputación de Galicia, CESGA) using CPU nodes and NVIDIA A100 GPU nodes with 32 CPUs per node and a maximum of 250 GB RAM, CUDA version 11.0. FANTASIA v2 was evaluated on (1) a desktop workstation at CBBIO lab with an AMD Ryzen 9 5950×16-core processor CPU (128 GB RAM), and a NVIDIA GeForce RTX 3090 Ti GPU (24 GB VRAM), CUDA version: 12.4; and (2) an HPC with NVIDIA A100-SXM4-80GB with CUDA 12.1, utilizing a single GPU, 50 CPU cores, and 100 GB of RAM. Animal proteomes GO annotations were performed in DragoHPC cluster (Spanish National Research Council, CSIC) using nodes with an NVIDIA Ampere A100 graphic card with 512 GB and Dual Intel Xeon Gold 6248 R 24 C 3.0 GHz processors. Only one species (*Panulirus ornatus*) was annotated using CPUs in FT3 due to the limitation in the maximum GPU node RAM memory.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All proteomes, homology-based annotations and FANTASIA.V1 (ProtT5 per-protein embeddings and annotations) are available in MATEdb2[21]. Information content, semantic similarity and other results are provided in Zenodo[31–36]. The complete list of the Zenodo files, as well as the Supplementary Data 1–4 can be found on GitHub (https://github.com/MetazoaPhylogenomicsLab/Martinez_Redondo_et_al_2024_Dark_Proteome_Animal_Tree_Of_Life).

Code availability

Both versions are functional and available. FANTASIA v1 pipeline used for the supplementary analyses is included in a Singularity (v3.8.3 + 231-g9dceb4240) container available on GitHub (https://github.com/MetazoaPhylogenomicsLab/FANTASIA). FANTASIA v2 pipeline is available on GitHub (https://github.com/CBBIO/FANTASIA). Steps followed for the creation of several intermediate files not available on Zenodo, as well as additional scripts used for the analyses and creating figures, are available on GitHub (https://github.com/MetazoaPhylogenomicsLab/Martinez_Redondo_et_al_2024_Dark_Proteome_Animal_Tree_Of_Life).

Received: 4 April 2025; Accepted: 1 August 2025;
Published online: 14 August 2025

References

1. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
2. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* **11**, 1160 (2021).
3. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
4. Thomas, P. D. et al. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).
5. Barrios-Núñez, I. et al. Decoding functional proteome information in model organisms using protein language models. *NAR Genom. Bioinform.* **6**, lqae078 (2024).
6. Hashimoto, T. et al. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat. Commun.* **7**, 12808 (2016).
7. Kenny, N. J. et al. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nat. Commun.* **11**, 3676 (2020).
8. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
9. Lewin, H. A. et al. The Earth BioGenome Project 2020: starting the clock. *Proc. Natl. Acad. Sci. USA*. **119**, e2115635118 (2022).
10. Mazzoni, C. J., Ciofi, C. & Waterhouse, R. M. Biodiversity: an atlas of European reference genomes. *Nature* **619**, 252 (2023).
11. Mc Cartney, A. M. et al. The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *npj Biodivers.* **3**, 28 (2024).
12. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).
13. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).
14. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
15. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
16. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
17. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
18. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
19. Heinzinger, M. et al. Bilingual language model for protein sequence and structure. *NAR Genomics Bioinforma.* **6**, lqae150 (2024).
20. Alexa, A. & Rahnenfuhrer, J. *topGO*. (Bioconductor, 2017). https://doi.org/10.18129/B9.BIOC.TOPGO.
21. Martínez-Redondo, G. I. et al. MATEdb2, a collection of high-quality metazoan proteomes across the animal tree of life to speed up phylogenomic studies. *Genome Biol. Evol*. **16**, evae235 (2024).
22. Louie, B., Bergen, S., Higdon, R. & Kolker, E. Quantifying protein function specificity in the gene ontology. *Stand. Genom. Sci.* **2**, 238–244 (2010).
23. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
24. Yoon, S., Baik, B., Park, T. & Nam, D. Powerful p-value combination methods to detect incomplete association. *Sci. Rep.* **11**, 6980 (2021).
25. Sharma, N., Naorem, L. D., Jain, S. & Raghava, G. P. S. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform*. **23**, bbac174 (2022).
26. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
27. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).
28. Gu, Z. & Hübschmann, D. simplifyEnrichment: a bioconductor package for clustering and visualizing functional enrichment results. *Genomics Proteom. Bioinforma.* **21**, 190–202 (2023).
29. Klopfenstein, D. V. et al. GOATOOLS: a Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).
30. Tomczak, A. et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* **8**, 5115 (2018).
31. Martínez-Redondo, G. I. Data (part 1) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10714961 (2024).
32. Martínez-Redondo, G. I. Data (part 2) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10717484 (2024).
33. Martínez-Redondo, G. I. Data (part 3) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10717774 (2024).
34. Martínez-Redondo, G. I. Data (part 4) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10717783 (2024).
35. Martínez-Redondo, G. I. Data (part 5) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10717885 (2024).
36. Martínez-Redondo, G. I. Data (part 6) from Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *Zenodo* https://doi.org/10.5281/zenodo.10717910 (2024).

Acknowledgements

G.I.M.R., A.M.R., and R.F. acknowledge funding and support from the LifeHUB/CSIC research network (PIE-202120E047). A.M.R. and R.F. acknowledge funding from the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. G.I.M.R. acknowledges the support of Secretaria d'Universitats i Recerca del Departament

d'Empresa i Coneixement de la Generalitat de Catalunya and ESF Investing in your future (grant 2021 FI_B 00476). R.F. acknowledges support from the following sources of funding: Ramón y Cajal fellowship (grant agreement no. RYC2017-22492 funded by MCIN/AEI /10.13039/501100011033 and ESF 'Investing in your future'), the Agencia Estatal de Investigación (project PID2019-108824GA-I00 funded by MCIN/AEI/10.13039/501100011033), the European Research Council (this project has received funding from the European Research Council (ERC) under the European's Union's Horizon 2020 research and innovation programme (grant agreement no. 948281)), the Human Frontier Science Program (grant no. RGY0056/2022) and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00420). A.M.R., F.P.C. and I.B. acknowledge support from MDM-2016-0687 and PID2021-127503OB-I00 funded by MCIN. We also thank Centro de Supercomputación de Galicia and CSIC for access to computer resources (CESGA and DRAGO, respectively).

Author contributions

G.I.M.R., A.M.R. and R.F. designed the study. G.I.M.R. performed the main analyses, created the FANTASIA v1 singularity container, and designed the figures. G.I.M.R. and I.B. wrote the original FANTASIA code and the pipeline to automate the methods. F.M.P.C. developed the FANTASIA v2 implementation. I.C. and J.M.F. optimized development, documentation, and extensively tested the tool and conducted statistics. B.C. and M.V.V. assisted with data analysis. G.I.M.R., A.M.R. and R.F. wrote the first version of the manuscript. R.F. and A.M.R. provided resources, the rationale and concepts behind the work, and supervised the study. All authors revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08651-2.

Correspondence and requests for materials should be addressed to Gemma I. Martínez-Redondo, Ana M. Rojas or Rosa Fernández.

Peer review information *Communications Biology* thanks Logan Hallee and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Aylin Bircan, Michele Repetto. A peer review file is available.