Decoding proteome functional information in model organisms using protein language models.

Israel Barrios-Núñez[1], Gemma I. Martínez-Redondo[3], Patricia Medina-Burgos[1], Ildefonso Cases[2]*, Rosa Fernández[3]*† and Ana M. Rojas[1]*†

[1]Computational Biology and Bioinformatics group, Andalusian Center for Developmental Biology (CABD-CSIC), Sevilla, 41013, Spain

[2]Bioinformatics Unit, Andalusian Center for Developmental Biology (CABD-CSIC), Sevilla, 41013, Spain

[3]Metazoa Phylogenomics Lab, Institute of Evolutionary Biology (CSIC-UPF), 08003 Barcelona, Spain.

* To whom correspondence should be addressed. Tel: +(34) 954 977 373; Fax: +34 954349376; Email: a.rojas.m@csic.es, icasdia@upo.es , rosa.fernandez@ibe.upf-csic.es

† Joint Authors

All the supplementary data and results are available at github: https://github.com/CBBIOtemp/Func

ABSTRACT

The gap between sequence information and experimental determination of function in proteins is currently unsolvable. The computational and automatic implementation of function prediction is not significantly improving function assignment, so there is a pressing need to find alternative methods, since standard approaches are not able to bridge the gap. Modern machine learning methods have been recently developed to predict function, being deep neural convolutional networks a popular choice. Protein language models have been also tested and proved reliable in curated datasets, but have not been applied yet to full collections of curated proteomes. We have tested how two different machine learning based methods perform when decoding the functional information from proteomes of selected model organisms. We found that the protein Language Models are more precise and informative across gene ontology categories for all the species, recovering functional information from transcriptomics experiments. These results indicate that these methods emerge as suitable alternatives for large scale annotation contexts and further downstream analyses, from which we propose a guide of use.

INTRODUCTION

As we enter the era of biodiversity genomics, thousands of sequences from multiple genomes are being released monthly, but their function is not experimentally determined. The infamous gap between experiments and available sequences has been practically unsolvable for decades now, where only experimental annotations exist for less than 0.5% of the known proteins (1).

Therefore, function annotation is mostly conducted through computational prediction, inherently based on evolutionary information (EI) reliance. Moreover, the study of the functional role of proteins is largely biased by particular interests of certain scientific communities so the comprehensive study of most genes is rarely attempted (2). For instance in human, it was found that most research is done on a small fraction of genes (~2,000 out of 19,000). With this wealth of protein-coding genes remaining understudied, there is a real limitation to understanding the functional space, and therefore it is not possible to create an unbiased and comprehensive set of annotations.

Reliance on evolutionary inference poses certain problems, the first is related to orthology identification. Despite large efforts to infer orthology relationship of novel genes, this is still a challenging problem (3). In the context of function annotation, sometimes functional analogy is the driving criteria to assign orthology (4) (even sometimes in the absence of sequence relatedness, which is essentially wrong).

On the other hand, the one sequence - one structure - one function" (SSF) paradigm (5, 6), is embedded at the core of the function prediction field, where orthologues have been proposed to have similar structures to perform the same function or a similar one - the so-called 'orthologue conjecture' (4). However, there are several scenarios when these assumptions might be violated. Indeed, the orthologue conjecture has been challenged (5), and it has been shown that at the same level of sequence divergence, orthologues' functions may differ more significantly than the paralogues.

The second problem is related to the use of sequence similarity as a proxy to assess homology. Traditional (naïve) methods to transfer functions using similarities are based on BLAST (7), reciprocal (8) scores, or profile-based models (eg, hidden Markov models) (9), being the latter the most reliable since they are robust statistical models from sequences. The rationale is that similar genes may perform similar functions, and therefore we accept to transfer functions above a certain threshold of sequence

similarity, as discussed above (10). These approaches have been used for decades, and are still in use to annotate regions of proteins, which are incorporated at the core of most modern methods. Recent methods include phylogeny-aware criteria (11) and are becoming the standards in the field. However, all these methods fall short when the sequence similarity is weak, or when the regions of the proteins exhibit certain properties (i.e., complexity, disorder, repeats, etc.).

This is especially critical as genomic and/or transcriptomics information of novel organisms are being sequenced to understand new and distinct biological questions. Non-model organisms pose a real challenge for function prediction since many of their genes are orphans or lineage-specific, as annotations are mainly transferred from a short handful of sets of model ones, which are already largely biased as explained above (2).

At the core of the problem is our understanding (or lack of) of "function", which being context-dependent can adopt diverse formulations. To circumvent this, the Gene Ontology Consortium (12) designed a protocol based on the assignment of controlled vocabularies to define functions, becoming the authoritative source that annotates the function of proteins in distinct categories. These categories (ordered as ontologies) are: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) and aim to capture the complexity of a function and standardize the terms used for function assignment.

The Critical Assessment of Functional Annotation (CAFA) monitors the state-of-the-art in automatic function prediction (13) by measuring how well computational methods predict functions. In their last report, they stated that predictions at the BP and MF categories have slightly improved over time, but overall, the term-centric prediction is still challenging.

Meanwhile, computer science methodologies were introduced years ago to assist in defining biological functions and were not showing a great impact until now, where they are gaining momentum due to the abundance of data and much improved technology in both hardware and software. Machine learning methods, for instance, were initially employed to predict secondary structure (14,15), and now, with the advancement of technology and the availability of vast amounts of data, the field is flourishing. This is neatly illustrated by the AlphaFold algorithm (16), which has been transformative in the field of predicting tridimensional structures of proteins from sequences. In the context of function predictions, a similar breakthrough is still yet to come. Machine learning methods are gaining momentum, as seen in the last CAFA3 (13) with machine learning-

based ensemble method (17) performing better than the rest of the methods across three challenges. Despite these improvements, the novel methods do not seem to have improved remarkably in predicting protein function, despite their reach (they are powerful) and the data availability.

Beyond the CAFA3 context, methods using convolutional networks (DeepGO, DeepGOPlus) have been also benchmarked against the CAFA3 datasets (18,19) in independent publications being reported to perform as good as the top performers.

Promising approaches can be adopted from orthogonal disciplines, and machine learning is definitely a vast resource for exploration (20). One is the linguistics field, where in a text, meaning can be inferred from how patterns of words are distributed (21). In analogy, as we know how constraints for words to have a meaning work, we know that the structure and function of a protein also constrain how mutations are selected or fixed through evolution. Therefore, it could be possible to infer function from sequence patterns.

In this particular context, protein function has been predicted via transfer learning (22), where proteins are represented as continuous vectors (embeddings) using language models (i.e., (23)). In Natural Language Processing (NLP), for instance the ELMo model is trained on unlabeled text-corpora to predict the most probable next word in a sentence, given all previous words in this sentence. Then, it learns a probability distribution for sentences, so the models can "learn" aspects of language such as syntax and semantics. Next, the trained vector representations (embeddings) are contextualized within its context. The main advantage is that two identical words could have different embeddings, depending on the surrounding words. These concepts have been used to train protein sequences at two levels, per-residue (word) to predict secondary structure and intrinsically disordered regions, and per-protein (sentence) to predict subcellular location (22), where they concluded that transfer learning succeeded in extracting information from unlabeled sequences to predict location.

Then, interesting approaches such GOPredSim (1) train Language Models on sequences, and transfer annotations using embeddings similarity instead of sequence similarity. The rationale is to predict GO terms through annotation transfer learning based on the proximity of proteins in the LM embedding space rather than in sequence space. The embedding used in their original work originated from deep learned language models (LMs) for protein sequences (SeqVec, ProtT5) transferring the knowledge gained from

predicting the next amino acid in 33 million protein sequences. They benchmarked the GOPredSim method using the CAFA3 challenges and gold datasets and found very similar results to CAFA3 ten top performers (1). As these methods are agnostic to evolutionary reliance, they could be useful to annotate functions of proteins that challenge the paradigm of SSF.

Since these methods have not been benchmarked outside the CAFA3 framework, which focuses on universal datasets, particularities at the species level might have not been taken into account. Here, we have analyzed their capabilities to retrieve precise, informative, and similar functional information in four representative model organisms (*Saccharomyces cerevisiae, Mus musculus, Drosophila melanogaster,* and *Caenorhabditis elegans.*).

MATERIAL AND METHODS

Datasets used in this study

Proteomes. We collected all the proteins from the reference proteomes of: *C. elegans* (CAEEL, UP000001940), *D. melanogaster* (DROME, UP000000803), *M. musculus* (MOUSE, UP000000589), and *S. cerevisiae* (YEAST, UP000002311) from https://www.uniprot.org/.

GO annotations. GO annotations from the corresponding proteins using GOA https://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gaf.gz, were extracted (Supplementary Table ST4)

Mapping of identifiers. Entries from the proteomes were mapped to genes through https://www.uniprot.org/id-mapping (as Nov, 2023, Supplementary Table ST5).

Transcriptomics studies used in this work

Published experiments were collected, and lists of differentially expressed genes were extracted to perform further enrichment analyses (see next sections). The rationale for selection was to cover different strategies and aims of these transcriptomics analyses: systemic global changes, and temporal series changes. For each experiment, we mapped each specific differentially expressed identifier when possible (due to differences in identifiers and versions used) to UniProt annotations and predictions datasets. *S. cerevisiae*: transcriptomics data of response against heat shock during a time course (24). *C. elegans:* we used the classic experiment that addressed lifespan (25)

focusing on the results of the *Nuo-6* knocked-out organism. *D. melanogaster:* we selected a temperature-acclimation transcriptome response (26). *M. musculus:* we selected a pre-B-cell differentiation experiment, in a cell line model (27) (Supplementary Table ST9).

Prediction methods used in this work

Profile-based method. As our internal control we used profile-based searchers against the Pfam-A database, (https://www.ebi.ac.uk/interpro/download/pfam/) using the HMMER suite (9). Searches were conducted to identify protein domains. Results were conservative filtered by keeping the best domain i.evalue (smallest than i.evalue ≤ 0.001) at the longest alignment coverage to avoid overlapping domains at a minimum alignment coverage of 70%. The GO annotations associated with the identified protein domains were extracted from pfam2go (http://current.geneontology.org/ontology/external2go/pfam2go). The PFAM outputs are Supplementary File S1.

Deep Learning-based methods (DLs): two methods are selected: DeepGO (18), and its enhanced version DeepGOPlus (19). These methods are deep learning-based approaches that also use evolutionary information, as well as protein-protein interaction data in a Convolutional Neural Network framework. For DeepGO we used the provided API to retrieve the results (https://deepgo.cbrc.kaust.edu.sa/deepgo/api/create). For DeepGOPlus, the software was available by authors and it was locally installed. In the local installation, parameters were modified to enable annotations in proteins longer than 1002 residues. As these methods annotate all entries at the root of each category, we only kept those GO terms at a depth >2 from the root of each GO category to avoid annotations at the root.

Language Models-based methods (LMs). We used the GOPredSim (1) implementation of two different language models (Seqvec and ProtT5). To run these methods protein embeddings are calculated (per protein) for the test dataset. Then, they calculate distances between the protein embeddings space against the embeddings of a reference dataset (GOA2022, entries with curated GO information), followed by a "Transfer Learning" approach based on these distances (the smaller the distance, the more likely to transfer the annotations). For the LMs methods proteins longer than 5000 amino acids were excluded from the predictions since longer lengths are not accepted by the methods (these represent less than 1% of the proteins in our datasets).

In order to avoid overfitting we recompiled the lookup tables removing the entries of a particular organism prior to running predictions on that particular organism (i.e.: when predicting for *D. melanogaster*, entries and embeddings from this organism are removed from the lookup table, and the same procedure was done in the remaining organisms). This was instructed in the launching of the methods.

All methods were run on CPU AMD Ryzen 9 5950X 16-Core Processor with 128GB RAM. Embeddings for both methods were calculated on a GPU NVIDIA GeForce RTX 3090 Ti (24GB VRAM). Results can be found in Supplementary File F2.

Estimation of measures derived from GO terms

GO Term precision. Mapping to GO terms was done using the GO.db R package (https://bioconductor.org/packages/release/data/annotation/html/GO.db.html), "based on the data provided by Gene Ontology (https://current.geneontology.org/ontology/go-basic.obo) with a datestamp from the source of 2023-01-01". To check how well the methods guess the exact GO term annotated by Uniprot, we devised a 4-classification type of match to assess the precision of the match looking at the GO graph for each organism. In this schema (see Figure 2a, Supplementary File S3) a "HIT" indicates when the method guesses the exact GO term (the term as appearing in the Uniprot annotation), "Close" when the method guesses either a child or a parent of the correct term in the directed acyclic graph (DAG), "related" when the method guesses either an ancestor or descendant in the DAG, and finally "unrelated" when it identifies terms outside the branch or related terms in the DAG. We recorded the precision type by each method in all the proteins for the different organisms and computed their frequency.

GO Term specificity. The GO term structure is an acyclic graph (DAG). The straightforward way to calculate the specificity of a term is by counting the number of jumps required for a given term to reach the root node. Ideally, that would be useful if the number of children per branch is similar and the number of children per node is more or less homogenous. In that case, we could assume that each node and each level of the node would be equally informative. However, this is not the case, if we have branch A with 4 levels, and another branch B with 11 levels, the nodes of level 3 in both branches are not equally informative.

An alternative way to deal with this situation is to calculate the Information Content (IC) for each node. IC depends not only on the distance towards the root node, but also on the number of descendants the node has (the greater, the less informative), and the probability that this term would appear in an annotated set (i.e., an organism dataset a very specific term should appear with less frequency than the more general).

It was calculated using the GoSim package (28) and the formula for calculating IC is:

$$IC = -log\mathbf{2}(p(t)) \qquad (1)$$

And it is defined as the negative logarithm of the probability of observing t. Where t is a particular GO term and p is the probability of that term occurring in a data set, so p(t) is calculated as:

$$p(t) = \mathbf{1} - \frac{'Child\ count'}{'Child\ count' + 'Ancestors\ count'} \qquad (2)$$

The probability of a term occurring is the frequency with which that particular term or any of its child/ancestors occur in a data set. GO terms that occur less frequently have higher IC and are assumed to be more specific. Child and ancestor count refer to the number of proteins that are assigned to that term given a reference dataset.

In this case, the references come from the OrgDb R package (https://www.bioconductor.org/packages/release/BiocViews.html#___OrgDb). The results are in Supplementary File S4.

Similarity among GO terms. The rationale behind assessing semantic similarity relies on the assumption that genes with similar functions should have similar annotation vocabulary and have a close relationship in the ontology structure. Semantic similarity among groups of GO terms was calculated using the GoSemSim package (29). We used the Wang method (30) that determines the semantic similarity of two GO terms based on both the locations of these terms in the GO graph and their relations with their ancestor terms, enabling us to perform quantitative comparisons, by first defining the semantic value of a term A as the aggregate contribution of all terms in the "A" Directed Acyclic Graph to the semantics of term A. So terms closer to term A in the graph contribute more to its semantics.

Given that a protein can receive multiple GO terms, semantic similarity among proteins needs to be aggregated from different semantic similarity scores of multiple GO terms associated with these proteins. To combine the semantic similarity among our annotations and the gold standards, we had used the Best-Match Average (BMA) method (https://yulab-smu.top/biomedical-knowledge-mining-book/semantic-similarity-overview.html#bma), which calculates the average of all maximum similarities. (Supplementary File S5).

GO term enrichment in transcriptomics experiments

We transferred all the protein annotations to mapped genes. As methods will provide different scores for similar proteins, in case that different proteins show different scores for the same term, and these proteins are associated with the same gene, we keep the highest score for the particular term from the protein pool (Supplementary File S2).

We extracted the sets of differentially expressed genes from each experiment (genes of interest, Supplementary Table ST6) and performed GO enrichment tests (see below) on them in all GO categories, using i) the GO terms extracted from the Uniprot annotations, and ii) the GO terms predicted by the methods. Enrichments were performed against all genes from each organism's genome as the universe (Supplementary Table T7) and the results are in the Supplementary File S7.

Effect of the scores. As each method has a different scoring system, and the scores are not comparable (see results), for each method we split the universe datasets in score values quantiles for each method to estimate potential effects of the scores on IC (Supplementary File S4) and SS (Supplementary File S5), and when retrieving functional enrichment experiments using the R quantile function.

Enrichment tests: we used the TopGO (31) 10.18129/B9.bioc.topGO, R package version 2.50.0.), setting a Fisher's exact test, using the "elim" algorithm (32) for all methods in BP, MF and CC categories. We kept the classic Fisher value of $p < 0.05$ to extract significant terms (Supplementary File S6).

Semantic similarity: we calculated the semantic similarity among the group of enriched terms from each method, respect the groups of terms from the "annotations" set using the GoSemSim package (29), and using the combined BMA metric as explained above in section above. Results are available in (Supplementary Table ST9).

RESULTS

Prediction methods increase the possible annotation coverage in model organism proteomes.

Most gene annotations are given to the gene products (proteins), therefore, we have focused on proteomes (33). On the other hand, Gene Ontology (GO) (12) provides annotations formalized as controlled vocabularies structured in three directed acyclic graphs (DAG) for each category, Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

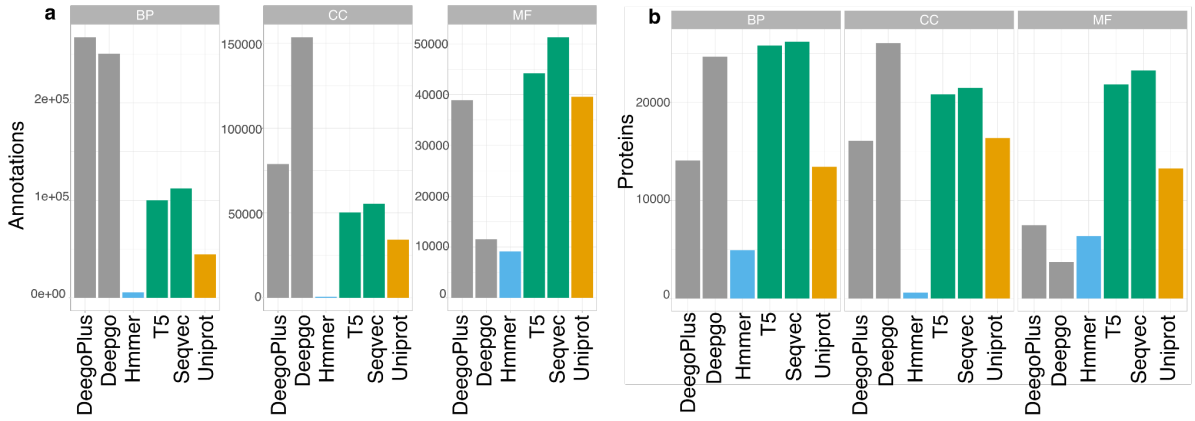

Figure 1. **Annotations and predictions in *C. elegans*. a.** Number of annotations (orange) and prediction (other colors), **b.** Proteins annotated (orange) and proteins with predictions out of a total number of 26664 proteins. BP (Biological Process), CC (Cellular Component), MF (Molecular Function). Green are Language Models based methods (LMs, T5 and Seqvec), gray are Deep Learning based methods (DLs, DeepGO and DeepGOPlus), blue is a profile based method (Hmmer, our internal control), and orange represents the annotations extracted from Uniprot. For the remaining species see Supplementary figure SF1.)

The methods provide many predictions across categories. As a general trend, the BP category receives more predictions than any other category (Figure 1a *for C.elegans,* Supplementary Figure SF1, Supplementary Table ST1). On the other hand, the annotation status for each organism is variable, being *C. elegans* the organism with less annotated proteins in all categories (Figure 1b, orange bar), and therefore it has been selected as the species to represent all the data obtained in this work through the main figures. For the remaining species, despite being better annotated, there is also a degree of proteome annotation incompleteness (Supplementary Figure SF1, Supplementary Table T2,T3, Supplementary file S2), which indicates the existence of a not negligible fraction of

proteins that remain unannotated in these extensively studied species (Supplementary Table ST8).

We used the profile-based method (HMMER) as an internal control (for high accuracy and low coverage). In general we observed that methods increased the number of proteins with predicted annotations (Figure 1b, Supplementary Figure SF1). In particular, *C. elegans*, being the worst annotated of the set, greatly benefits from these predictions in terms of coverage. LMs provide a large number of predictions in all categories (except for *S. cerevisiae* CC, Supplementary Figure SF1) as compared to the Uniprot available annotations. Interestingly, for the MF category, within DLs, DeepGOPlus performed worse than HMMER in terms of retrieving predictions, which is itself a low coverage method, however for CC, DeepGOPlus is the method providing more annotations. It is interesting to note that LMs annotate more proteins than DLs in MF and BP (except in *D. melanogaster*). So, in *C. elegans* LMs increase coverage in all categories, while in *D. melanogaster* and *M. musculus* this trend is only observed in the BP. These results may as well reflect that differences in the annotation status among proteins of different model organisms parallels the overall interests of the community annotating these organisms, as reported previously (2).

Overall, these results indicate that LMs (GOPredSim-T5 and GOPredSim-Seqvec) provide a large and more consistent coverage of predicted annotations in proteins across organisms, regardless of their annotation stage, making them promising tools to address functional studies in non-model organisms.

Language Model annotations are more precise.

As methods gave a large coverage of predictions, we needed to assess how precise they were. To this purpose we estimated the ability of the methods to guess the exact GO terms assigned to proteins (as per GOA2uniprot annotation), using a defined schema based on the DAG's localisation of the prediction in relation to the real annotation (Figure 2a). In *C.elegans*, despite the large number of predictions, BP was the worst predicted category while MF was the best (Figure 2b). The same trend was observed for the remaining species (Supplementary Figure SF2) in terms of precision.

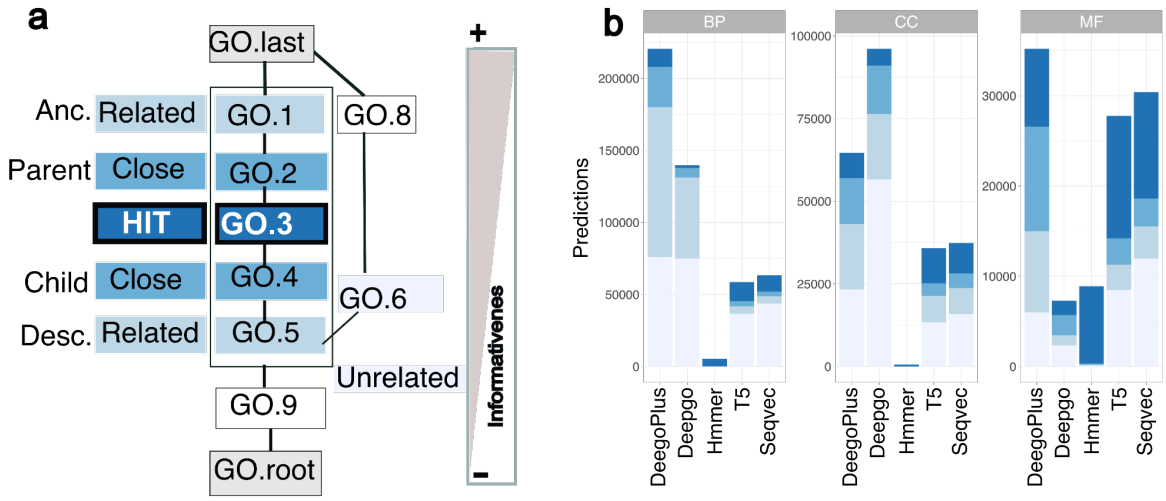

Figure 2. Precision achieved by methods in *C. elegans*. a. Definition of precision in the context of an hypothetical DAG, where "GO.root" is the root of the category (i.e. BP), while "GO.last" is the most specific term. HIT (darkest blue) matches the term, "Close" is either a parent or child, "Related" includes any descendant/ancestor, and there are unrelated predictions. The informativeness depends also on the position of the predicted term in the graph and is higher close to the end. **b. Number of predictions** (Y-axis) per method (X-axis) and per category *for C. elegans*, according to their "precision type" following the colouring scheme of a, results for the other species can be found in Supplementary figure SF2, which exhibit similar trends.

As expected, HMMER (our base-control) is very precise at the expense of a very low coverage. These results suggest that LMs, in particular GOPredSim-T5, provide a wide annotation range at a better precision.

Language Models are more informative than Deep Learning methods

Once established that methods produced a large coverage and at good precision, we estimated how informative these predictions were. One way to look at the informativeness of the terms was to identify the predicted term localisation within the DAG. The closer to the root (Figure 2a), the more general a term is, therefore, the less informative. To quantify this we calculated the information content (IC) of the predictions retrieved by the different methods, where less informative terms (closer to the root) provided low IC values (Figure 3a). Our results in *C. elegans* indicated that overall LMs were the most informative methods in BP and MF categories, providing more specific

information. Interestingly, for the CC category HMMER, the base-line method performs better in all species.

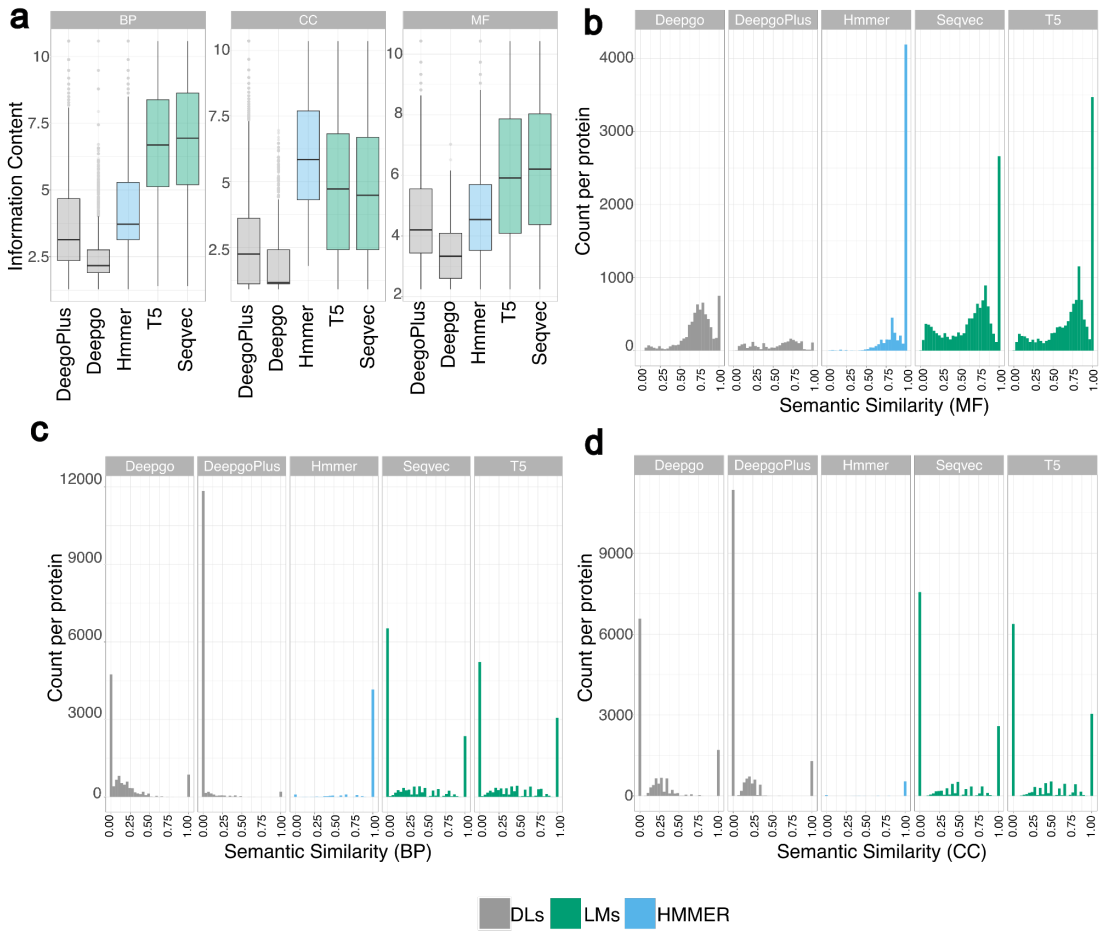

Figure 3. Informativeness provided by the methods in *C. elegans*. a. Information content. The IC (Y-axis) values for each method across categories. The larger, the more informative (more specific terms). **b.** The Semantic Similarity per protein (Y-axis) is the counts of proteins, X-axis indicates the value of SS (1 is that annotations and predictions are identical) forMF, **c for** BP, and **d** for CC. Results for the other species can be found in Supplementary figure SF3, which exhibit similar trends.

In addition to IC, we also estimated the Semantic similarity among the predicted GO terms compared to the GO terms annotated per protein, where values close to 1 indicate a high similarity among the group of terms (Figure 3b-d). We found that the MF category (Figure 3b) is the one providing more similar terms, while BP and CC (Figure 3c-d) provide many unrelated predictions per protein as compared to the standard annotations, therefore MF is the best predicted category. The same trend was observed for the other species Supplementary Figure SF3)

Language Models provide a better precision and coverage per protein

Now, in order to check the performance of these methods at protein coverage, we should consider two different but also related aspects. First, the number of annotations that are correctly predicted by a method at a given precision type (Figure 4a right), and how many of the produced predictions coincide within the per protein space of annotations (Figure 4a left), is it predicting everything that can be predicted or is it predicting differently?

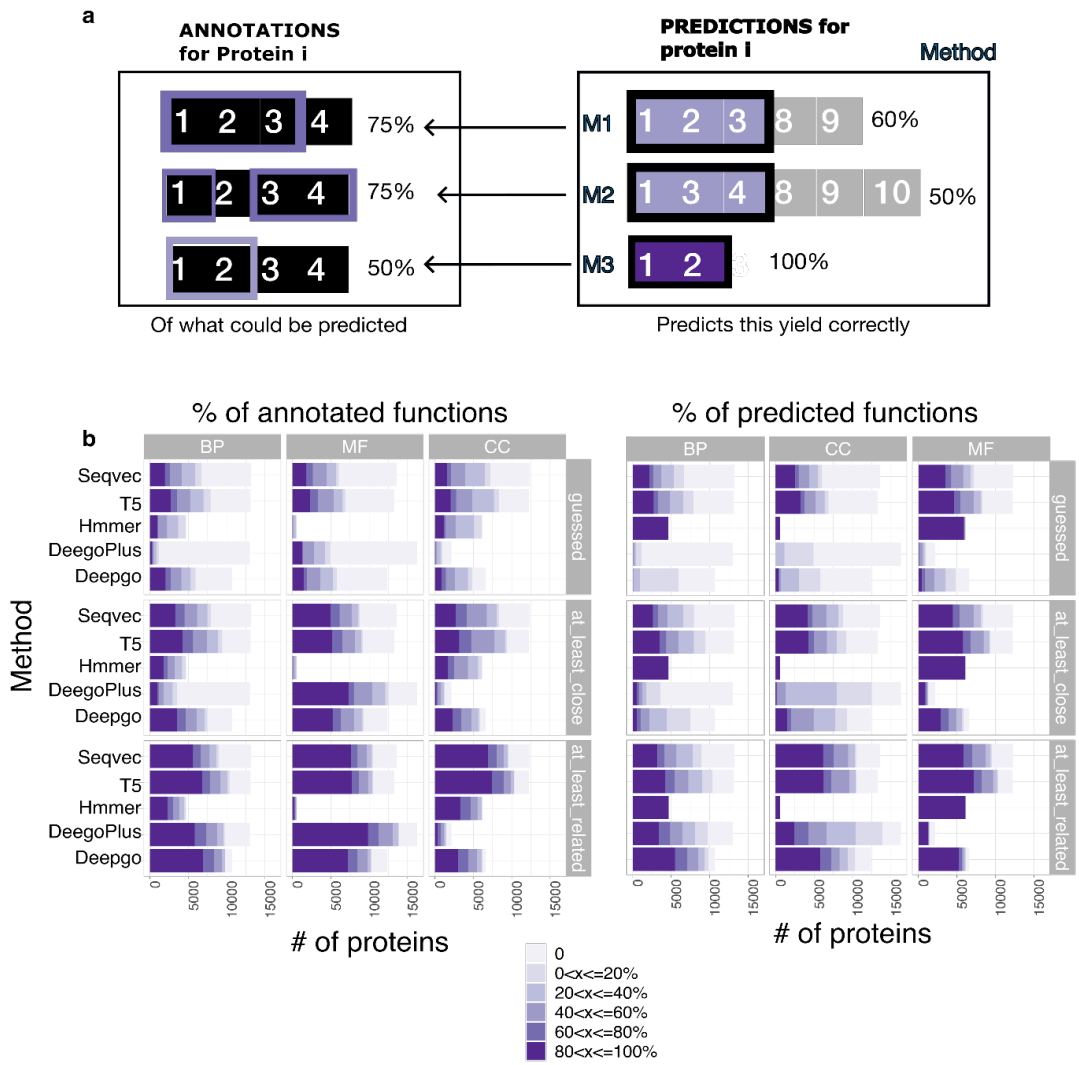

Figure 4. Per protein coverage of methods. a. Left represents the fraction of correct predictions out of what could be predicted (from the annotation space), while right represents from all the prediction's space, how many are correct (are annotated). **b. Results for *C. elegans*,** left % of annotated functions across categories, while right is the % of predicted functions. Colors indicate the ranges in percentages. Results for the other species can be found in Supplementary figure SF4, which exhibit similar trends.

For instance, For *C. elegans*, in the BP category HMMER predicts ~5000 proteins to have a GO term match as a HIT (Figure 4b, right, "guessed" panel), however it should be noted that a very short fraction of these proteins have predicted all the possible predictions that could be predicted at this level (Figure 4b, left, "guessed" panel). In other words, while most predictions of HMMER per protein are guessed perfectly (Figure 4b, right), it does not predict all the annotations that can be predicted (it misses many predictions, (Figure 4b, left)). At the highest precision level ("guessed", upper level) LMs clearly outperform DLs in precision. The same trends are observed in the remaining species (Supplementary Figure SF4).

The scores have an effect in information and coverage

With the exception of HMMER, where the selection was based on e-value and alignment length, each method analysed here produced different scores. The scores are not straightforward comparable since DLs do not fully cover the range of 0-1 (Figure 5a).

For DLs, the DeepGO method (API version) filters the scores beforehand (i.e., produces scores >0.3-1), while the range of DeepGOPlus varies from (0.1-0.5). We then distributed the scores on quantiles for each method, assuming that the last quantile will concentrate the "best" scores.

Effect on Semantic similarity. Does the scoring affect the recovery of "true" functions? The SS was used as a proxy to measure the similarity among predictions and annotations by calculating how similar their GOs are per protein. As expected, and in line with the precision results both LMs provide the highest SS values at the last quantile (Figure 5b, in *C. elegans*, same trend for the remaining organisms in Supplementary figure SF5). GOPredSim-T5 is the method providing the highest SS across categories and among species, and their SS value is increasing along quantiles. While DeepGOPlus follows the same trend, however they do not reach values >0.7). Interestingly, the SS in the BP for C. elegans is very low. In general DLs exhibit very heterogeneous trends across species and across categories. In general, DeepGOPlus is the method that provides the lowest SS from all, showing a variable effect depending on species and/or category. Respect to the species, *D. melanogaster* and *C. elegans* are the worst predicted species for BP and CC by LMs, while *M. musculus* is the best predicted species in all categories (Supplementary figure SF5, lowest panel).

Effect of scores on the Information Content. Another way to look at the quality of information is focusing on its informativeness. In this context, the IC provides a measure of the specificity of the terms (Figure 2a) . To address the effect of the scores on the IC (Figure 5c) we plotted the IC according to the quantile distribution. The plot of IC along quantiles did not show any particular trend for LMs (a slight increase on IC along quantiles was observed). However, in DLs, we found the opposite trend, with IC values decreasing along quantiles. This suggests that while increasing the score LMs will slightly produce more specific terms, for DLs we observe the opposite trend, with more general terms, therefore the "best" DLs scores are uninformative (Supplementary figure SF5 all panels c, Supplementary File S3).

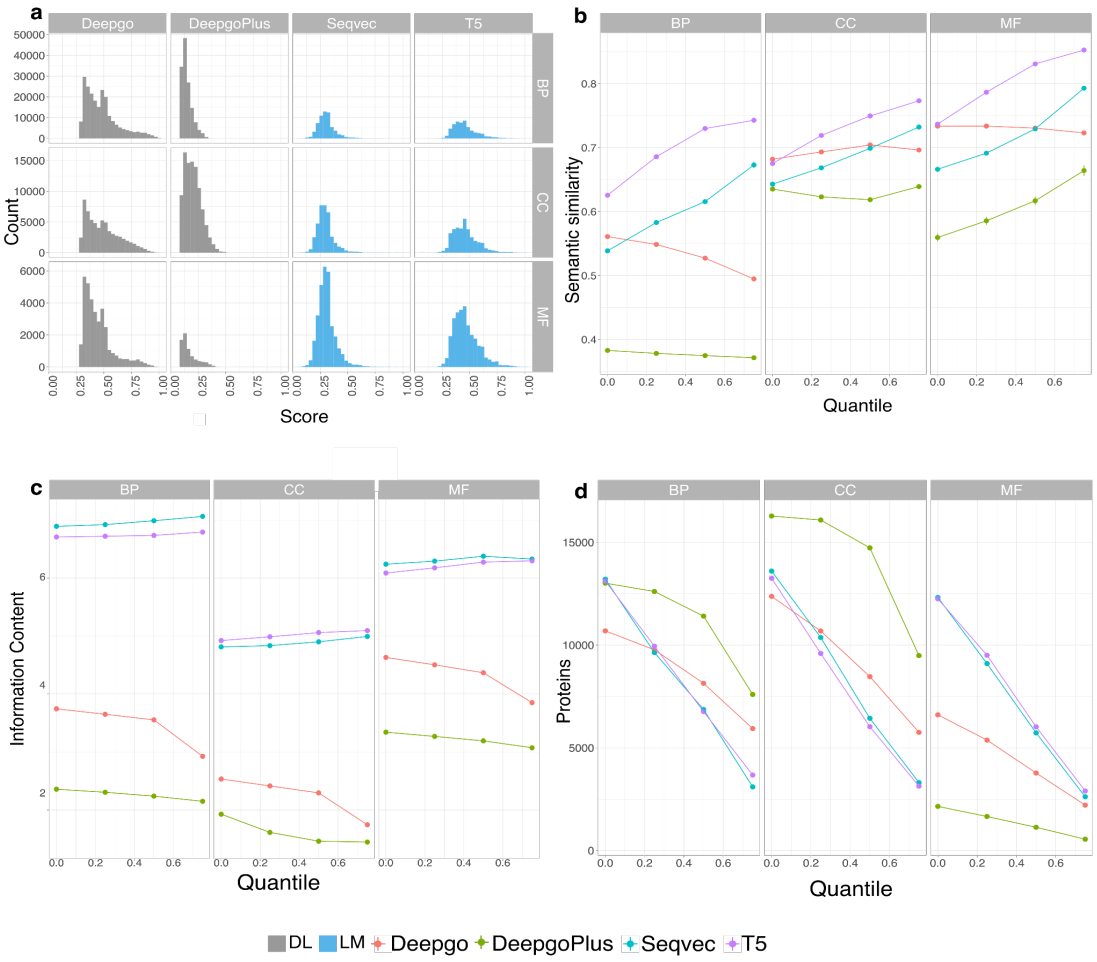

Figure 5. Score effect. a. Score distribution of methods for *C. elegans*. **a.** Effect of score on Semantic similarity, **b.** score's effect on information content, **c.** scores effect on protein coverage. Results for the other species can be found in Supplementary figure SF5, which exhibit similar trends.

Effect of scores on protein coverage. As high scores were associated with accurate predictions by means of SS and IC (as seen above, Figure 5b,c), the next question was

to ask how many proteins are dropped along quantiles. As expected for *C. elegans* (Figure 5d) there is a significant drop of proteins along quantiles for all the methods, a trend that stands for the remaining organisms (Supplementary figure SF5, all panels d). So, better scores are better choices for LMs.

Language models recover functional information from transcriptomics experiments

In functional genomics a key downstream approach is to conduct GO enrichment experiments in different conditions to identify particular enriched or depleted functions. In this line, we run GO enrichment tests on published differentially expressed genes, using first the predictions, then the real annotations, to further estimate SS of the enrichment results.

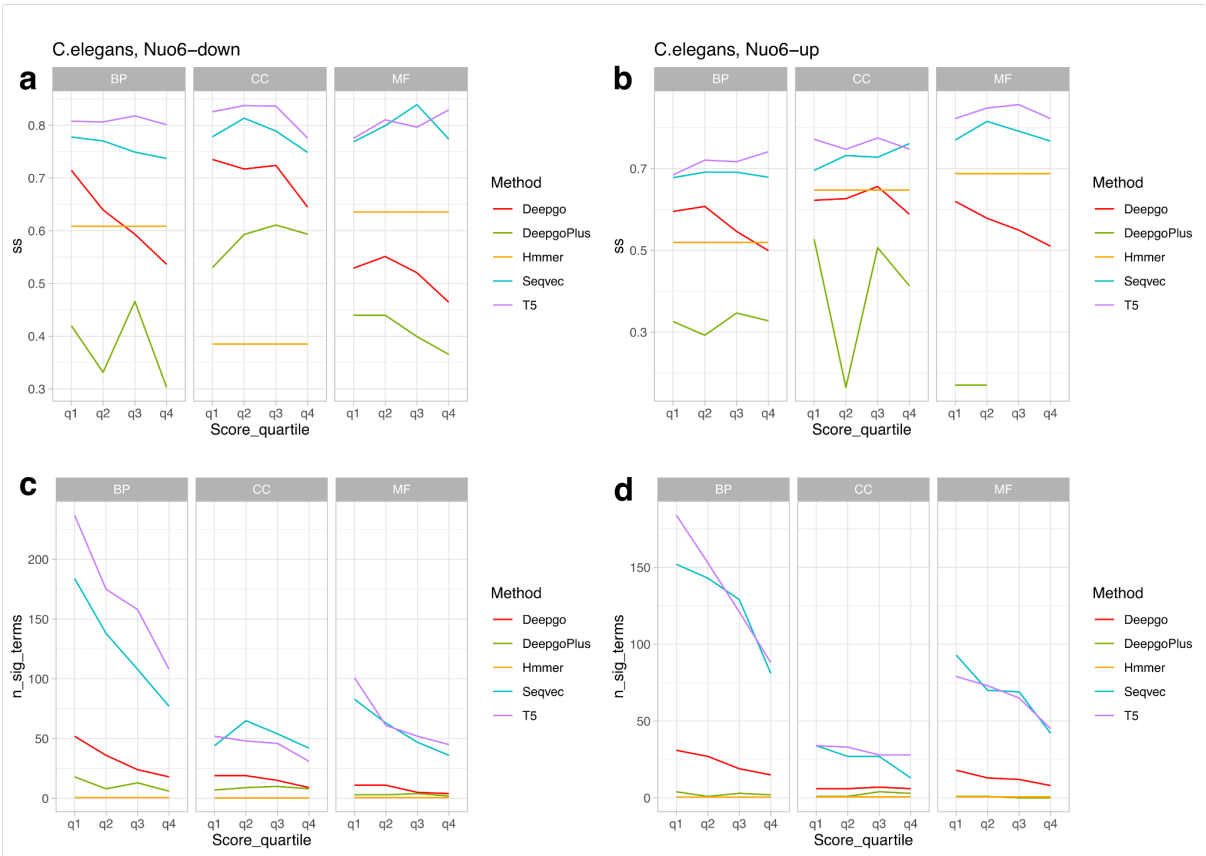

Figure 6 Recovery of functional information for *C. elegans*. Semantic similarity between GO terms from TOPGO enrichment experiments (GO from annotations in Uniprot vs GO from predictions are compared). Upper panels: SS accross score quantiles. Lower panels: number of proteins harboring significant terms across quantiles.Results for the other transcriptomics experiments can be found in Supplementary figure SF6, which exhibit similar trends.

Protein predictions were transferred to their cognate genes (in case of redundant predictions from different proteins pointing to the same gene, we kept the highest score).

For each GO category, we again distributed the proteins according to score quantiles (henceforth q groups), in which the first quantile (q1) will contain all the proteins, the second (q2) will exclude the first, the third (q3) will exclude the two previous, and finally the last (q4) will contain only the most stringent scores.

Figure 6 shows the results for the *Nuo-6* ko *C. elegans* (25), where the SS decreases for all methods (Figure 6a,b, and Supplementary Figure 6 for the remaining species) when the number of genes annotated with significant terms also decrease as an effect of the score (Figure 6c,d and Supplementary Figure SF6). Despite this decrease, in LMs even at q4 the SS was >0.7 in up-regulated and downregulated genes indicating that these methods could recover functional information. Overall, the SS is higher in LMs for all categories. The best predicted category is MF, as shown before (Figure 3b) and CC, being BP the worst. In this case the number of proteins annotated with significant terms is large (Figure 6c,d), but it remarkably decreases at higher quantiles. LMs provide the highest semantic similarities (Supplementary Table ST9).

Methods complement existing proteome annotation

We found unannotated proteins with predictions (missing annotations in at least one category). Methods contributing the most to this are LMs in BP and MF categories, whereas DLs annotate more proteins in the CC category. It is interesting to observe that HMMER still provides annotations to proteins that in principle should be annotated by automatic methods, as this method is at the core of most automated predictions. Given that the scores of the methods are not comparable (see methods), we have selected examples at the top levels of each scoring range for the methods (Supplementary Table T8).

In *S. cerevisiae*, there are still 1012 proteins partially annotated by Uniprot (missing annotations in at least one category), where the methods do complement the annotations. From these, 3 proteins have no annotations at any category in Uniprot, all having evidence at a protein level. Those are P47129 (Assembly-complementing factor 4), which is receiving annotations at all categories by the methods. GO terms indicate a role in autophagy and protein transport (GOPredSim-T5), and DNA repair (GOPredSim-Seqvec). In *C. elegans t*here are 9158 proteins partially annotated in Uniprot (missing annotations in at least one category). With 86 out of them not annotated at all, methods provide a wealth of annotations for exploration. For instance, the Glutathione S-Transferase, gst-39 (Q9NAB0), receives annotations by GOPredSim-T5 in BP

(glutathione metabolic process, prostaglandin biosynthetic process), and CC (intracellular membrane-bounded organelle). Another interesting protein is Protein edg-1, (Q17436) receiving annotations by GOPredSim-T5 in BP in developmental-related processes, and CC (nucleus). In *D. melanogaster,* 6130 proteins are partially annotated, of which 35 have no annotations. For example, the "Common Dpr-interacting protein" (Q9VDD5) is a LRR protein, annotated in BP (germ cell development, molting cycle, collagen and cuticulin-based cuticle, by GOPredSim-T5) and CC (P-granule, plasma membrane, by GOPredSim-T5). This is interesting since LRRs proteins are believed to be involved in defense proteins (34). The UPF0184 protein CG14818 (O97420) also receives interesting annotations at BP (cellular heat acclimation, regulation of transcription by RNA polymerase II). The AT27166p protein (Q8T3W6), "hale gene", is involved in viral release and proteolysis (BP, by GOPredSim-T5).

Finally, in mice, 16315 proteins are partially annotated, with 78 proteins unannotated. For instance, the TBK1 binding protein 1 (Z4YKN2) is predicted at the BP to participate in I-kappaB kinase/NF-kappaB signaling, and innate immune response (BP by GOPredSim-T5 and GOPredSim-Seqvec), while other methods provide more general terms like regulation of cellular process (DeepGO). The Netrin G1 protein (F8WJ48), receives predictions at BP (substrate adhesion-dependent cell spreading, modulation of chemical synaptic transmission, nervous system development) and MF (cell-cell adhesion mediator activity), etc. The BCL2/adenovirus E1B 19 kDa-interacting protein 2 (Q91VL0), is involved in signaling and calcium binding, MF (GTPase activator activity, calcium ion binding), BP (apoptotic process, regulation of catalytic activity).

DISCUSSION

Current genomic sequencing projects produce large amounts of data, especially for non-model organisms. Traditional homology-based methods fall short in such contexts since at large divergence times it is difficult to capture the evolutionary signal. In addition, similarity cannot be used as a proxy of evolutionary relatedness in species-specific genes or it is difficult to implement when sequences have particular features such as repeat, low complexity regions, etc.

Promising approaches from the computer science field have been applied for protein annotation. However, their practical use is hindered by the trade-off between annotation precision and coverage, and most of the time by their usability, since large queries are not suitable in web-based solutions.

Regarding precision, our approach aimed to mimic the natural variance shown by expert annotators, which would annotate proteins at different levels of detail. For instance, in the MF the term GO:0016002 (sulfite reductase activity) has a child term GO:0004783 (sulfite reductase (NADPH)), and a parent term GO:0016667 (oxidoreductase activity). The three terms are pointing to a different level of detailed information. Ideally, a perfect method should match the annotation (as we have defined here as "HIT"). However, since not all annotate at the same level of detail (as human experts), it is fair then to assume that methods are performing well in this context. LMs are more precise than DLs.

Prediction methods should be accompanied by a scoring schema enabling users to either select or discard predictions. Scores are values usually ranging from 0-1 (derived from a particular scoring function), and thresholds for their usage should be recommended. In the methods tested here, either how the scores are calculated are not obvious from the publications, or explanations are partial with no clear recommendations about score usage provided (with the exception of GOPredSim for Seqvec that provides scores for each category). Although the method developers benchmarked them against dedicated functional datasets (i.e., the standard CAFA3 (13), to the best of our knowledge, not at a "within" organism level. We also analyzed the effect of each method scoring system on each species, in SS and IC. For DLs, both versions performed very differently, with DeepGOPlus providing poorer performance when compared to its predecessor.

Another relevant issue is related to the informativeness of the predicted terms. For LMs, at better scores (q3 and q4), there is a slight improvement on IC, which was already high even for the first quantile. On the contrary, increasing the score in DLs produced a decrease on IC, indicating that for these methods better scores are associated with more general terms. A reliable scoring schema should provide a higher semantic similarity at a high score. While this stands for LMs, for DLs we observed the opposite trend, indicating that for DLs the higher scores are providing different annotations.

Not all the GO categories are equally well predicted by the methods. For instance, the BP is the one receiving most annotations, but it is the one more difficult to predict in the least accurately annotated species (*C. elegans* and *D. melanogaster*), as seen by the SS results. However, GoPredsim-T5 recovered SS above 0.7 in the high scoring quantiles (q3, and q4).

Finally, we analysed the capabilities of these methods to recapitulate the functional information extracted from transcriptomic experiments. We found that these methods are capable of retrieving similar functions (measured by semantic similarity) even in depletion of significant term bearing proteins. The loss of terms according to increased score is expected since this should decrease the number of available proteins, and therefore impact the yield of significant terms in the enrichment analyses, this could be important in scenarios where few genes show differential expression in certain experiments.

Our results indicate that LMs are providing better SS than DLs, but even in the case of comparable outcomes, the informativeness of the DLs terms is too low. The DLs tested here despite providing a vast number of annotations, they are very general and do not reflect the "true" annotation state as inferred by the SS. Their inconsistent behavior could be likely due to the fact that they do contribute with very general annotations.

On the contrary, LMs are more robust to coverage loss, their predictions are more informative, and as expected, sacrificing the protein coverage has an effect on the SS.

All our combined results indicate that the LMs provide a fair coverage, at a good semantic similarity, providing more specific information. In contrast, deep learning methods offer broader coverage but yield more general annotations. This trade-off prompts a nuanced consideration of precision versus coverage in annotation methods.

Our study demonstrates that protein language models can be particularly useful in poorly annotated genomes, which is here exemplified by the model *C. elegans* both at the genome level and at the functional retrieval of differentially expressed genes.

With all this information in hand, as a recommended use of these prediction to annotate a novel genome for downstream enrichment experiments, we would suggest first to keep the standard annotations assigned by the reference repository (i.e. Uniprot), since those would likely come from homology-based methods. Then complement the annotations with the q2-q4 LMs predictions, since they have a good coverage while being informative (according to IC and SS), and last, we would complement the annotations with DLs at any score, since they provide very general terms. In this way a genome can be mostly annotated and used in downstream approaches.

DATA AVAILABILITY

Results and supplementary data are available at https://github.com/CBBIOtemp/Func

FUNDING

This work is supported by the Ministerio de Ciencia e Innovación, MCIN/AEI/10.13039/501100011033 [(PID2021-127503OB-I00) to A.M.R. and I.B., (PID2019-108824GA-I00) to R.F., (RYC2017-22492) to R.F., (MDM-2016-0687) to I.C.,P.M.,A.R.], the LifeHUB/CSIC research network [(PIE-202120E047) to G.I.M.R., A.M.R., and R.F.], the European Research Council [(grant agreement no. 948281) to R.F], the Human Frontier Science Program [(RGY0056/2022) to R.F], Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya [(AGAUR 2021-SGR00420) to G.I.M.R., (AGAUR 2021-SGR00420 to R.F.)].

Funding for open access charge: Ministerio de Ciencia e Innovación, MCIN/AEI/10.13039/501100011033

CONFLICT OF INTEREST

None declared

REFERENCES

1. Littmann,M., Heinzinger,M., Dallago,C., Olenyi,T. and Rost,B. (2021) Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep*, 11, 1160.

2. Stoeger,T., Gerlach,M., Morimoto,R.I. and Nunes Amaral,L.A. (2018) Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol*, 16, e2006643.

3. Linard,B., Ebersberger,I., McGlynn,S.E., Glover,N., Mochizuki,T., Patricio,M., Lecompte,O., Nevers,Y., Thomas,P.D., Gabaldón,T., *et al.* (2021) Ten Years of Collaborative Progress in the Quest for Orthologs. *Molecular Biology and Evolution*, 38, 3033–3045.

4. Nehrt,N.L., Clark,W.T., Radivojac,P. and Hahn,M.W. (2011) Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS Comput Biol*, 7, e1002073.

5. Studer,R.A. and Robinson-Rechavi,M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*, 25, 210–216.

6. Anfinsen,C.B., Haber,E., Sela,M. and White,F.H. (1961) THE KINETICS OF FORMATION OF NATIVE RIBONUCLEASE DURING OXIDATION OF THE REDUCED POLYPEPTIDE CHAIN. *Proceedings of the National Academy of Sciences*, 47, 1309–1314.

7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J Mol Biol*, 215, 403–410.

8. Hernández-Salmerón,J.E. and Moreno-Hagelsieb,G. (2020) Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. *BMC Genomics*, 21, 741.

9. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform*, 23, 205–211.

10. Conesa,A., Götz,S., García-Gómez,J.M., Terol,J., Talón,M. and Robles,M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21, 3674–3676.

11. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38, 5825–5829.

12. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47, D330–D338.

13. Zhou,N., Jiang,Y., Bergquist,T.R., Lee,A.J., Kacsoh,B.Z., Crocker,A.W., Lewis,K.A., Georghiou,G., Nguyen,H.N., Hamid,M.N., *et al.* (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20, 244.

14. Barton,G.J. (1995) Protein secondary structure prediction. *Curr Opin Struct Biol*, 5, 372–376.

15. Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, 90, 7558–7562.

16. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.

17. You,R., Zhang,Z., Xiong,Y., Sun,F., Mamitsuka,H. and Zhu,S. (2018) GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, 34, 2465–2473.

18. Kulmanov,M., Khan,M.A. and Hoehndorf,R. (2018) DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34, 660–668.

19. Kulmanov,M. and Hoehndorf,R. (2020) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, 36, 422–429.

20. Bepler,T. and Berger,B. (2021) Learning the protein language: Evolution, structure, and function. *Cell Syst*, 12, 654-669.e3.

21. Harris,Z.S. (1954) Distributional Structure. *WORD*, 10, 146–162.

22. Heinzinger,M., Elnaggar,A., Wang,Y., Dallago,C., Nechaev,D., Matthes,F. and Rost,B. (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20, 723.

23. Peters,M.E., Neumann,M., Iyyer,M., Gardner,M., Clark,C., Lee,K. and Zettlemoyer,L. (2018) Deep contextualized word representations. 10.48550/arXiv.1802.05365.

24. Nuño-Cabanes,C., Ugidos,M., Tarazona,S., Martín-Expósito,M., Ferrer,A., Rodríguez-Navarro,S. and Conesa,A. (2020) A multi-omics dataset of heat-shock response in the yeast RNA binding protein Mip6. *Sci Data*, 7, 69.

25. Senchuk,M.M., Dues,D.J., Schaar,C.E., Johnson,B.K., Madaj,Z.B., Bowman,M.J., Winn,M.E. and Raamsdonk,J.M.V. (2018) Activation of DAF-16/FOXO by reactive oxygen species contributes to longevity in long-lived mitochondrial mutants in Caenorhabditis elegans. *PLOS Genetics*, 14, e1007268.

26. MacMillan,H.A., Knee,J.M., Dennis,A.B., Udaka,H., Marshall,K.E., Merritt,T.J.S. and Sinclair,B.J. (2016) Cold acclimation wholly reorganizes the Drosophila melanogaster transcriptome and metabolome. *Sci Rep*, 6, 28999.

27. Gomez-Cabrero,D., Tarazona,S., Ferreirós-Vidal,I., Ramirez,R.N., Company,C., Schmidt,A., Reijmers,T., Paul,V. von S., Marabita,F., Rodríguez-Ubreva,J., *et al.* (2019) STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data*, 6, 256.

28. Fröhlich,H., Speer,N., Poustka,A. and Beißbarth,T. (2007) GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, 8, 166.

29. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26, 976–978.

30. Wang,J.Z., Du,Z., Payattakool,R., Yu,P.S. and Chen,C.-F. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 23, 1274–1281.

31. Alexa,A. and Rahnenfuhrer,J. (2023) topGO: Enrichment Analysis for Gene Ontology. 10.18129/B9.bioc.topGO.

32. Alexa,A., Rahnenführer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22, 1600–1607.

33. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49, D480–D489.

34. Proell,M., Riedl,S.J., Fritz,J.H., Rojas,A.M. and Schwarzenbacher,R. (2008) The Nod-Like Receptor (NLR) Family: A Tale of Similarities and Differences. *PLOS ONE*, 3, e2119.