Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz375

INGA 2.0: improving protein function prediction for the dark proteome

Damiano Piovesan [1,*] **and Silvio C.E. Tosatto** [1,2]

[1]Department of Biomedical Sciences, University of Padua, Padua, Italy and [2]CNR Institute of Neuroscience, Padua, Italy

Received February 13, 2019; Revised April 29, 2019; Editorial Decision April 30, 2019; Accepted April 30, 2019

ABSTRACT

Our current knowledge of complex biological systems is stored in a computable form through the Gene Ontology (GO) which provides a comprehensive description of genes function. Prediction of GO terms from the sequence remains, however, a challenging task, which is particularly critical for novel genomes. Here we present INGA 2.0, a new version of the INGA software for protein function prediction. INGA exploits homology, domain architecture, interaction networks and information from the 'dark proteome', like transmembrane and intrinsically disordered regions, to generate a consensus prediction. INGA was ranked in the top ten methods on both CAFA2 and CAFA3 blind tests. The new algorithm can process entire genomes in a few hours or even less when additional input files are provided. The new interface provides a better user experience by integrating filters and widgets to explore the graph structure of the predicted terms. The INGA web server, databases and benchmarking are available from URL: https://inga.bio.unipd.it/.

INTRODUCTION

The problem of predicting protein function from the amino acid sequence is intrinsically difficult due to the limited number of available experimentally-validated examples and the complexity of the cellular machine. The Gene Ontology (GO) (1), which provides a vocabulary of function descriptors, includes more than 45 thousand different terms. Manually annotated GO terms in UniProtKB (2) cover <1% of the entries. UniProt-GOA (3) provides automatic annotation for the rest of the database. It employs a number of different techniques exploiting sequence properties, InterPro (4) predictions and taxonomy. Yet about 40% of entries remains unannotated and the quality of predicted GO terms is unknown. The need for better methods to improve functional characterization of known proteins and to predict the function of new organisms is becoming critical. Scores of new function prediction methods are published every year, however, an objective overview of the real performance is problematic and a comparison between methods is almost impossible given the heterogeneity of adopted evaluation protocols. The Critical Assessment of protein Function Annotation (CAFA) challenge solves this problem implementing a real blind test (5) and highlights the most effective methods for automatic protein function prediction. The last CAFA results (6) show that methods using the 'transfer by homology' approaches (7–9), based on sequence similarity, compete both with machine learning (10) and integrative methods (11). BLAST F_{max}, for example is only about 10% lower than the best method. CAFA also shows that predicting biological processes (BP) is much more difficult than molecular function (MF). The Naive baseline, which assigns terms simply based on their frequency in UniProtKB to all benchmark proteins, is still a good predictor due to strong biases in annotation database, for example a very large fraction of experimentally annotated proteins are annotated with the 'protein binding' term. Eukaryotes, including simple organisms such as yeast, are much more difficult to predict than prokaryotes. Recent work shows organism complexity negatively correlated with residue level annotation (12). A large fraction of eukaryotic proteome residues, up to 50% for human, is uncharacterized and remains inaccessible to common domain detection pipelines. The so-called 'dark proteome' is thought to be composed by new folds, transmembrane regions and intrinsically disordered residues (13). Here we present a new version of INGA (11), Interaction Network GO Annotator, which combines homology, domain architecture, interaction networks and 'dark' features to predict protein function. In our previous work we already showed how protein-protein interactions can be used effectively to infer function based on the 'Guilty by Association' principle exploiting protein-protein interaction (PPI) networks (14). The fact that disordered regions, compared to globular domains, provide a repertoire of new alternative functions is becoming evident in the literature (15–17), in particular for longer regions (18). The extraction of disorder features from the sequence has been proven to be useful for function prediction methods (19,20). INGA already ranked in the top ten in

*To whom correspondence should be addressed. Tel: +39 498276269; Fax: +39 498276260; Email: damiano.piovesan@unipd.it

CAFA2 (6) for both MF and BP ontologies when considering the F_{max} in the full-evaluation / no-knowledge mode. Recently, INGA 2.0 ranked in the top ten in CAFA3 for all three ontologies and for both the F_{max} and S_{min} evaluations (manuscript in preparation). A lower S_{min}, in particular, indicates the ability of the method to predict specific and difficult rare terms, i.e. those less represented in annotation databases. The INGA server has been completely redesigned in order to improve reproducibility, reliability and usability. The INGA 2.0 algorithm can be executed in two alternative modes by providing either the protein sequence(s) or BLAST and InterProScan predictions. In the first case, different components can be excluded to speed-up the calculation at the cost of partially losing specificity. In the second case, INGA can provide maximum accuracy and predict function for entire genomes in less than one hour.

MATERIALS AND METHODS

INGA derives function information from different sources to generate a consensus prediction. The method exploits homology, domain architecture and interaction networks as proxies for transferring function from annotated proteins. The new version of INGA also integrates intrinsic disorder and transmembrane region prediction to cover information from the 'dark proteome', i.e. regions poorly characterized in public databases. The consensus prediction provided by INGA 2.0 has been evaluated in the CAFA3 assessment, resulting among the top ten methods for all three ontologies and both F_{max} and S_{min}. A description of the implementation and the contribution of each component to the overall consensus accuracy follows.

Homology and protein interaction networks

Homology is based on the concept of vicinity. In the context of genetic phylogeny, homologous proteins share a common ancestor and therefore the same biological function (21). Other methods are able to distinguish paralogy from orthology because paralogous proteins often diverge too much and lose function similarity (8). However these methods are bound to the computational cost of building a phylogenetic tree and to the number of available representatives for a given protein family. Instead, INGA infers homology by simply measuring sequence similarity. In particular, it performs a BLAST search, with default parameters, against the entire UniProtKB sequence database. The default sorting based on the BLAST Bit-score is used to transfer GO terms and assign an estimated probability representing prediction precision. Different probabilities are assigned simply based on the BLAST ranking independently from input properties or alignment coverage. In contrast to the previous INGA version, hits are not filtered. INGA also exploits information from protein-protein interaction networks to predict function. This has been shown to be effective in our previous works (14), in particular for Cellular Component and partially Biological Process ontologies. The new version of INGA uses exactly the same implementation. It considers only direct interactors from the STRING database, filtered with a confidence score of at least 0.4 corresponding to the STRING default. GO terms associated to direct interactors are transferred with a probability representing their enrichment in comparison to the entire STRING database. The enrichment is calculated with a Fisher exact test, while probability is estimated considering the P-value ranking and measuring the precision for each ranking position.

Domain architecture database and the dark proteome

Proteins are organized in modular architectures (22). According to classification databases, complex architectures are provided by the repetition and rearrangements of a relatively small number of domains (23,24). Domains can be considered as functional determinants and are therefore subject to evolutionary pressure. When the three dimensional structure is conserved across different species, domain detection from the sequence is straightforward as key positions are also conserved. InterPro provides the largest collection of sequence models (signatures) of protein domains with known biological role (4). However, when processing proteomes with InterPro a large fraction of residues remains undetected, in particular for eukaryotes (12). The 'dark proteome' includes all those functional modules for which key residues are not position specific but, instead, characterized by compositionally biased regions like in disordered and transmembrane proteins (13). INGA transfers GO terms from proteins with the same domain architecture. The new version uses InterProScan (25), Phobius (26) and MobiDB-lite (27) to predict domain, transmembrane and disordered signatures (labels) respectively. In addition to 'transmembrane', Phobius also provides the 'signal peptide', 'cytoplasmic' and 'extracellular' labels. MobiDB-lite predictions are transformed into four different signatures either representing the localization in the sequence or indicating 'fully disordered' when disorder content is larger than 75%. Both InterPro and 'dark' signatures are combined to generate the INGA domain architecture database. Architectures are calculated for the entire GOA (3). GO annotations of proteins with the same architecture are grouped together and sorted inside the cluster based on their enrichment (Fisher's test) calculated in comparison to the rest of the database (background). When a target sequence matches an architecture in the INGA database, GO terms are transferred with a probability estimated on the ranking provided by the enrichment. Terms with a *P*-value lower than 0.001 are discarded. This ensures that significantly enriched terms are specific, i.e. distant from the ontology root. Table 1 shows the number of enriched terms for different architectures in the database. Notably, 57% of architectures contain 'dark' signatures (Dark), while the number of associated proteins is much higher for globular (Non-dark) architectures, indicating, on average, larger clusters. The number of enriched terms is almost the same for the two major classes but terms enriched in the 'dark' database (Dark) are slightly more specific (Average depth). The introduction of 'dark' signatures results in the split of large clusters and therefore the separation of different functional groups. On average 5 MF, 10 BP and 2 CC terms are associated to each architecture and can be safely transferred to the matching sequences.

Table 1. Enriched terms in the INGA domain architecture database

				Enriched terms			Average depth		
	Signature	Architectures	Proteins	MF	BP	CC	MF	BP	CC
Dark	Transmembrane	165 465	11 864 693	778 207	1 445 467	248 211	3.67	4.23	2.67
	Signal	109 529	2 953 070	433 446	884 071	133 662	3.50	4.19	2.60
	Cytoplasmic	5312	67 800	36 365	142 550	30 771	3.85	4.51	2.97
	Extracellular	3292	22 381	25 425	102 560	16 810	3.82	4.54	2.95
	C-term disorder	166 470	2 329 632	747 643	1 738 544	329 203	3.65	4.32	2.77
	N-term disorder	161 436	2 348 434	729 138	1 675 203	329 310	3.66	4.33	2.78
	Central disorder	126 412	1 134 920	519 506	1 228 743	239 888	3.67	4.37	2.81
	Fully disordered	3047	43 869	7 837	30 078	7 722	3.38	4.39	2.71
	All	488 312	18 112 467	2 181 980	4 626 913	843 145	3.61	4.26	2.71
Non-dark	All	366 108	72 418 252	2 019 833	3 943 739	650 507	3.50	4.11	2.63
Total		854 420	90 530 719	4 201 813	8 570 652	1 493 652	3.56	4.19	2.68

Number of molecular function (MF), biological process (BP) and cellular component (CC) terms statistically enriched (enriched terms) for different types of architectures in the INGA database. (Average depth) Average minimum distance from the corresponding ontology root. All architectures contain an InterPro signature, dark architectures also contain a non-globular signature (Dark). The same architecture can have multiple 'dark' signatures, partial counts are provided in separate rows (transmembrane, signal, etc.).

Consensus and training

The training set is the one provided by the CAFA organizers and published on the official web page (https://biofunctionprediction.org/cafa) as 'CAFA 3 Training Data' corresponding to all experimental GO terms available in UniProtKB, including 35 086, 50 813 and 49 328 proteins with 371 584, 2 047 227 and 582 454 terms for the MF, BP and CC ontologies respectively. The training set is used to estimate with a ten fold cross-validation the correlation between precision and ranking position for the three INGA components: Homology, Architectures and Interactions. In Figure 1 the distribution of precision in relation to the ranking is provided. When generating predictions, INGA assigns a confidence score which is the average precision of the ranking. The ranking is calculated in different ways for the three INGA components. For Homology it corresponds to the BLAST output position and ranking 1 means the hit (or set of hits) with the best Bit-score. For the Architecture and Interaction components the ranking is provided by the enrichment. Ranking 1 corresponds to all those terms with the lowest *P*-value (see methods for details). The final consensus is calculated in the same way as in the previous INGA version, i.e. calculating the joint probability for terms provided by different methods. An additional weighting parameter to balance the contribution of different methods has been trained using the same dataset and applying a simple grid search algorithm.

Evaluation

INGA has been evaluated in the CAFA2 (6) and CAFA3 (manuscript in preparation) blind test experiments as 'INGA-Tosatto'. In CAFA2, considering the F_{max} and the full-evaluation/no-knowledge mode, INGA ranked among the top 10 methods for MF and BP ontologies. In CAFA3, INGA (version 2.0) is in the top 10 also for CC and for both the F_{max} and S_{min} metrics. The latter takes into consideration the information content of the terms and gives an indication about prediction specificity (28). Terms with high information content are less frequent in the annotation databases and therefore more difficult to predict. A fair comparison with other methods is very difficult outside the CAFA context due to a number of variables which cannot be controlled, for example the version of training databases, ontology, etc. In Table 2, we report a comparison with the previous INGA and baseline methods as implemented in CAFA using the benchmarking data provided in CAFA2 which contains 2618 BP, 2938 CC and 1828 MF protein targets. It has to be noted that numbers in the table are not comparable with CAFA evaluations as we consider the whole reference instead of subcategories and differences in calculation details exist. For example, no- and limited-knowledge examples were not separated in order to maximize the dataset size and the source of GO terms (UniProtGOA) contains new terms not present in the benchmarking. Also, the test is not fully blind, as the training data (UniProtGOA) overlaps with test examples. Table 2 is provided just to show the contribution of the different INGA components and a comparison with baseline methods. We used the same input (when applicable), i.e. same BLAST database, UniProtGOA, Gene Ontology version, etc. in order to equally propagate the effect of possible biases. For a fair evaluation we refer to the official CAFA3 results. Table 2 also reports performance for the INGA Architectures component as it can be used for fast large-scale prediction and also for the same component without 'dark' features (INGA Arch Non-dark). The evaluation is provided as in the full CAFA evaluation, where methods with a lower coverage are penalized because recall is calculated averaging over the benchmark size. INGA 2.0 outperforms all methods and has ∼10% higher F_{max} compared to its previous version for all ontologies. The INGA Architecture component has generally an 18% lower F_{max} than the consensus but 6% higher than the one without 'dark' features. The S_{min} shows the same trend with a stronger difference between INGA 1.0 and INGA 2.0. Figure 2 shows the precision recall curves for methods reported in Table 2. The higher performance of INGA 2.0 over other methods (and INGA 2.0 Arch over INGA 2.0 Arch Non-Dark) can be explained by the higher number of considered features, as 'dark' features are expected to be extensively represented both in the training and test examples. All INGA CAFA3 predictions and benchmarking data are available for download from URL https://inga.bio.unipd.it/documentation/cafa.

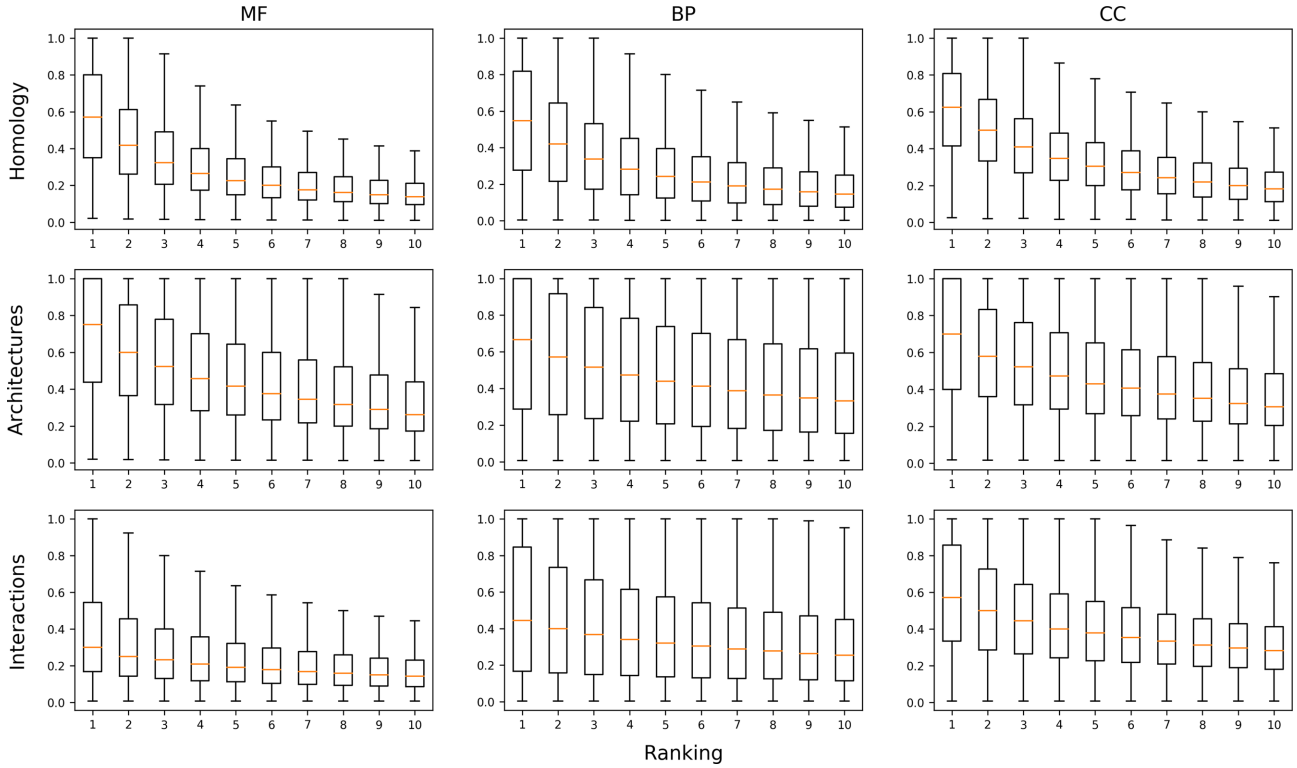

Figure 1. Estimated precision of three INGA components. Precision is reported for different ranking positions. Ranking is provided by BLAST Bit-score for Homology and by the enrichment *P*-value for Architectures and Interactions (see methods). The horizontal axes is cut at 10.

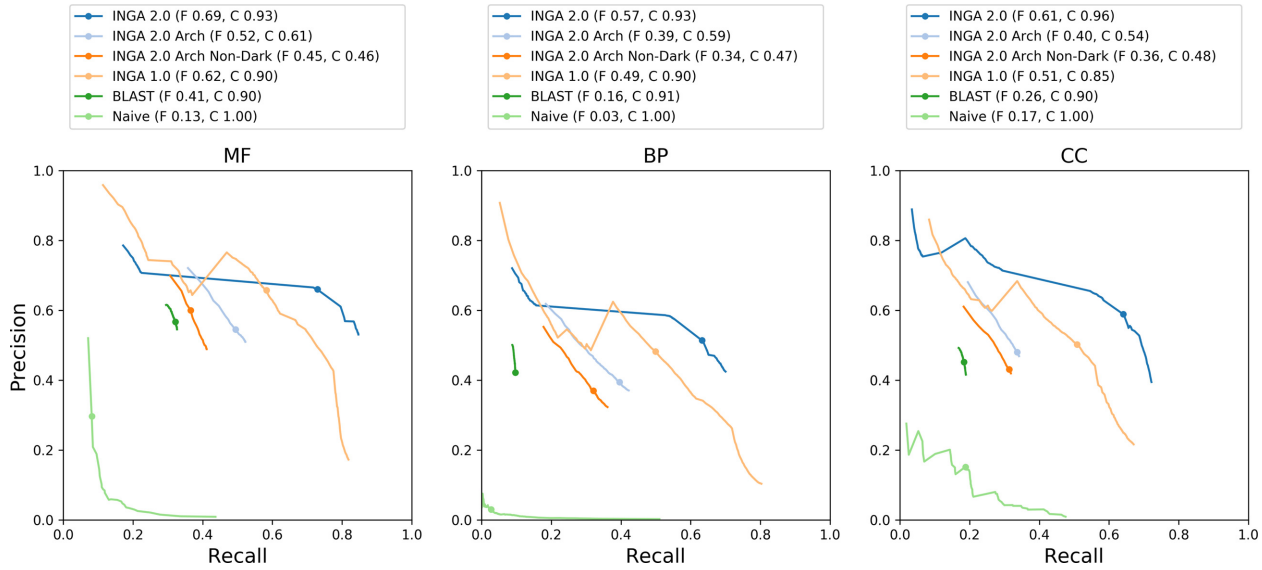

Figure 2. Precision recall curves for methods compared in Table 2 for the three GO ontologies. In the legend, (F) is the F_{max} and (C) is the coverage as the fraction of predicted targets.

Implementation

The INGA web server is implemented using the REST (Representational State Transfer) architecture. The INGA services can be accessed both from a web interface or a custom client. Submitted jobs can be retrieved at a later time by providing the session identifier or the URL to the result page. INGA guarantees to maintain job sessions for at least two weeks. Predictions are stored permanently in a database where entries are indexed by their sequence in order to speed up the service when requesting a cached protein.

Table 2. INGA performance in comparison with other methods

Ontology	Method	Th (F_{max})	Precision	Recall	F_{max}	Th (S_{min})	S_{min}	Coverage
MF	INGA 2.0	0.49	0.660	0.730	0.693	0.67	5.83	0.93
	INGA 2.0 Arch	0.28	0.545	0.495	0.519	0.60	13.25	0.61
	INGA 2.0 Arch Non-Dark	0.47	0.600	0.365	0.454	0.60	11.20	0.46
	INGA 1.0	0.78	0.658	0.583	0.618	0.95	10.33	0.90
	BLAST	0.68	0.568	0.321	0.410	1.0	19.25	0.90
	Naive	0.06	0.296	0.082	0.128	0.6	28.93	1.00
BP	INGA 2.0	0.40	0.515	0.632	0.567	0.56	29.91	0.93
	INGA 2.0 Arch	0.16	0.394	0.396	0.395	0.46	71.96	0.59
	INGA 2.0 Arch Non-Dark	0.21	0.370	0.321	0.344	0.57	62.54	0.47
	INGA 1.0	0.59	0.482	0.499	0.490	0.76	56.98	0.90
	BLAST	0.22	0.422	0.097	0.158	1.0	123.91	0.91
	Naive	0.22	0.030	0.027	0.029	0.46	150.09	1.00
CC	INGA 2.0	0.40	0.589	0.641	0.614	0.56	3.78	0.96
	INGA 2.0 Arch	0.16	0.480	0.337	0.396	0.40	12.78	0.54
	INGA 2.0 Arch Non-Dark	0.16	0.431	0.314	0.363	0.50	11.51	0.48
	INGA 1.0	0.65	0.503	0.508	0.505	0.87	10.19	0.85
	BLAST	0.79	0.452	0.184	0.262	1.0	25.77	0.90
	Naive	0.09	0.152	0.188	0.168	0.09	32.12	1.00

This evaluation corresponds to the CAFA full-evaluation with both no- and -limited-knowledge examples merged in a single benchmark. Precision and recall measures are reported for the confidence threshold which maximize the *F*-score. The coverage is the fraction of predicted targets. INGA Architecture (INGA Arch.) component includes 'dark' signatures. INGA 2.0 corresponds to the full algorithm. BLAST and Naive are implemented and trained as described in CAFA2. Table values do not correspond to a fair blind test as training and test examples overlap.

SERVER DESCRIPTION

Input

The INGA website is free and open to all users and there is no login requirement. The interface can alternatively accept either protein sequences (Sequence input tab) or BLAST and InterPro predictions (Prediction input tab). In the first case INGA outputs single or multiple predictions (up to 50 or 1000 in slow and fast mode respectively) from pasted or uploaded FASTA sequences or UniProtKB accessions (e.g. P04050). A checkbox group allows the user to choose which component to run, i.e. limiting the execution to the INGA Architectures for a faster prediction. A single job (e.g. 10 sequences) lasts around 30 min in default mode and 15 min in fast mode considering only the INGA Architectures component. The alternative Prediction input tab allows to provide intermediate files, namely InterPro output, a BLAST search against UniProtKB and another BLAST search against the STRING sequence database. In this case input sequences are not necessary and INGA generates predictions in constant time independently of the input size.

Output

The server provides a results page listing all submitted sequences. Once predictions are ready, the user can access single protein pages listing the predicted GO terms. Terms are split into three tables available in three different tabs corresponding to the different ontologies. For each GO term the score (probability) and annotation source (UniProtKB annotated entries) provided by different methods are reported in the same row. Predicted terms are sorted by INGA score and then by specificity, i.e. terms more distant from the root are shown first. A left sidebar provides filters and widgets to explore the graph structure of the predicted terms. The specificity and INGA score can be filtered on the fly. Ancestors and children of a given term can be highlighted in a single click in order to visualize specific GO branches. The protein architecture (where available) is shown on the top of the table and a feature viewer can be optionally open to visualize sequence position of the detected signatures. Both prediction and predicted features are available for download both in JSON and text formats.

CONCLUSIONS

We have presented a new version of the INGA algorithm for the prediction of Gene Ontology terms from the protein sequence. The new version integrates 'dark' proteome information to improve prediction accuracy, in particular intrinsic disorder and transmembrane region detection. INGA ranked in the top ten for both CAFA2 and CAFA3. A new option allows fast prediction of entire genomes at the cost of partially losing accuracy. The web server was completely redesigned to provide a better interpretation of the function and visualization of the different predicted GO branches. We believe that improving the characterization and classification of 'dark' features will provide a better description of protein function and quality of predictors.

ACKNOWLEDGEMENTS

The authors are grateful to Marco Necci for initial help with the generation of MobiDB-lite predictions and to members of the BioComputing UP group for insightful discussions.

FUNDING

European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement [778247].

Conflict of interest statement. None declared.

REFERENCES

1. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
2. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
3. Huntley,R.P., Sawford,T., Mutowo-Meullenet,P., Shypitsyna,A., Bonilla,C., Martin,M.J. and O'Donovan,C. (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.*, **43**, D1057–D1063.
4. Mitchell,A.L., Attwood,T.K., Babbitt,P.C., Blum,M., Bork,P., Bridge,A., Brown,S.D., Chang,H.-Y., El-Gebali,S., Fraser,M.I. *et al.* (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.*, **47**, D351–D360.
5. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
6. Jiang,Y., Oron,T.R., Clark,W.T., Bankapur,A.R., D'Andrea,D., Lepore,R., Funk,C.S., Kahanda,I., Verspoor,K.M., Ben-Hur,A. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.
7. Khan,I.K., Wei,Q., Chapman,S., KC,D.B. and Kihara,D. (2015) The PFP and ESG protein function prediction methods in 2014: effect of database updates and ensemble approaches. *GigaScience*, **4**, 43.
8. Sahraeian,S.M., Luo,K.R. and Brenner,S.E. (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res.*, **43**, W141–W147.
9. Piovesan,D., Luigi Martelli,P., Fariselli,P., Zauli,A., Rossi,I. and Casadio,R. (2011) BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. *Nucleic Acids Res.*, **39**, W197–W202.
10. Minneci,F., Piovesan,D., Cozzetto,D. and Jones,D.T. (2013) FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. *PLoS One*, **8**, e63754.
11. Piovesan,D., Giollo,M., Leonardi,E., Ferrari,C. and Tosatto,S.C.E. (2015) INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res.*, **43**, 1–5.
12. Mistry,J., Coggill,P., Eberhardt,R.Y., Deiana,A., Giansanti,A., Finn,R.D., Bateman,A. and Punta,M. (2013) The challenge of increasing Pfam coverage of the human proteome. *Database*, **2013**, bat023.
13. Perdigão,N., Heinrich,J., Stolte,C., Sabir,K.S., Buckley,M.J., Tabor,B., Signal,B., Gloss,B.S., Hammang,C.J., Rost,B. *et al.* (2015) Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15898–15903.
14. Piovesan,D., Giollo,M., Ferrari,C. and Tosatto,S.C.E. (2015) Protein function prediction using guilty by association from interaction networks. *Amino Acids*, **47**, 2583–2592.
15. Piovesan,D., Tabaro,F., Mičetić,I., Necci,M., Quaglia,F., Oldfield,C.J., Aspromonte,M.C., Davey,N.E., Davidović,R., Dosztányi,Z. *et al.* (2016) DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.*, **45**, D219–D227.
16. Necci,M., Piovesan,D. and Tosatto,S.C.E. (2016) Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein Sci. Publ. Protein Soc.*, **25**, 2164–2174.
17. Radivojac,P., Iakoucheva,L.M., Oldfield,C.J., Obradovic,Z., Uversky,V.N. and Dunker,A.K. (2007) Intrinsic disorder and functional proteomics. *Biophys. J.*, **92**, 1439–1456.
18. Necci,M., Piovesan,D., Dosztanyi,Z., Tompa,P. and Tosatto,S.C.E. (2017) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*, **34**, 445–452.
19. Ofer,D. and Linial,M. (2015) ProFET: Feature engineering captures high-level protein functions. *Bioinforma. Oxf. Engl.*, **31**, 3429–3436.
20. Vidulin,V., Šmuc,T. and Supek,F. (2016) Extensive complementarity between gene function prediction methods. *Bioinformatics*, **32**, 3645–3653.
21. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
22. Söding,J. and Lupas,A.N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays News Rev. Mol. Cell. Dev. Biol.*, **25**, 837–846.
23. Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2016) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.*, **45**, D289–D295.
24. Paladin,L., Hirsh,L., Piovesan,D., Andrade-Navarro,M.A., Kajava,A.V. and Tosatto,S.C.E. (2016) RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Res.*, **45**, D308–D312.
25. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.*, **30**, 1236–1240.
26. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
27. Necci,M., Piovesan,D., Dosztányi,Z. and Tosatto,S.C.E. (2017) MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics*, **33**, 1402–1404.
28. Clark,W.T. and Radivojac,P. (2013) Information-theoretic evaluation of predicted ontological annotations. *Bioinforma. Oxf. Engl.*, **29**, i53–i61.