Wei2GO: weighted sequence similarity-based protein function prediction

Maarten J.M.F Reijnders[1,2]
Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland
Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Abstract

Motivation

Protein function prediction is an important part of bioinformatics and genomics studies. There are many different predictors available, however most of these are in the form of web-servers instead of open-source locally installable versions. Such local versions are necessary to perform large scale genomics studies due to the absence of any limitations imposed by web servers such as queues, prediction speed, and updatability of reference data.

Results

This paper describes Wei2GO: a weighted sequence similarity and python-based open-source protein function prediction software. It is compared against the Argot2 and Argot2.5 web servers, which use a similar concept. Wei2GO shows an increase in performance according to precision and recall curves, Fmax scores, and computational time. Wei2GO is further assessed on its biological relevance by annotating part of the *Chlamydomonas reinhardtii* triacylglycerol pathway, *Anopheles gambiae* Toll signaling pathway, and *Pseudomonas putida* Twin Arginine Targeting (TAT) complex. These annotations show Wei2GO is able to accurately annotate relevant data with high precision and recall.

Availability and implementation

Wei2GO can be found at https://gitlab.com/mreijnders/Wei2GO

Motivation

Predicting protein functions computationally is an integral part of genomics, due to the laborious process of experimentally determining protein functions. This is most apparent in the size difference of UniProt's TrEMBL and SwissProt databases [1], which contain predicted and manually curated proteins respectively. Because of the dependence on predicted protein functions, many software tools have been developed in an attempt to improve the accuracy of electronically annotating these proteins.

A comprehensive overview of protein function prediction methods is the Critical Assessment of Functional Annotation (CAFA) competition. Its most recent edition, CAFA3, compared the predictions of 68 participants employing a wide variety of techniques including sequence homology, machine learning, and text mining amongst others [2]. The top-performing predictors in this challenge are considered the state-of-the-art for *de novo* protein function prediction. With this exposure and proven performance, predictors participating in CAFA have the potential to act as a resource for any user wanting to predict protein functions of their organism of interest.

For the CAFA participants to be such a resource, they have multiple factors to consider: accuracy, accessibility, and speed. An in-depth look at the top performers of CAFA shows that

nine out of fourteen teams provide their method as a tool accessible through a web server, however only four out of fourteen teams provide a working locally installable option (Table 1). Web servers provide an excellent option for the annotation of a limited amount of proteins, but larger amounts are problematic due to computational limitations.

Contrary to web servers, local versions provide opportunities to speed up predictions without limitations usually imposed by web servers, such as limited protein sequence inputs or queues shared with multiple users. Local versions provide the user the means to update the software databases, which is a necessity to provide the most up-to-date and accurate predictions. As a bonus, local versions are often open-source and provide flexibility in the usage of the tool by advanced users, e.g. by changing databases or utilizing a meta-predictor [3]. As shown in Table 1, there is a need for accurate prediction tools which provide the options that come with a local installation.

Table 1: An overview of the accessibility of the top performing CAFA3 team's methodologies. Listed are the teams which placed in the top 10 performers for biological processes, molecular functions, or cellular components. Last known update is based on the most recent publication date, or tool update.

Team	Software	Web server	Working local version	Last known update
Argot25 Toppo lab	Argot2.5 [8]	✓	✗	09/1/2017
DeepMaster[1]	DeePred [11]	✗	✗	
Holm[1]	Pannzer2 [12]	✓	✓	08/05/2018
INGA Tossato	INGA [13]	✓	✗	01/10/2018
Jones-UCL[2]	FFPred [14]	✓	✓[a]	25/10/2016
Kihara lab[1]	PFP [15]	✓	✗	01/03/2019
NCCUCS	Unknown	✗	✗	
Orengo funfams	FunFams [16]	✓	✓	18/06/2019
Schoofcropbiobonn[1]	PhyloFun [17]	✓	✗	17/05/2019
Temple[2]	MS-kNN [18]	✗	✗	
Tian lab[2]	GoFDR [19]	✗	✗	
TurkuBioNLP1	Unknown	✗	✗	
Zhang-Freddolino lab[1]	i-TASSER [20]	✓	✓	17/04/2020
Zhu lab[3]	NetGO [21]	✓	✗	22/02/2020

1. Latest protein function prediction software from the authors, but uncertainty about their direct affiliation to the CAFA submission.
2. Prediction software used by the Jones-UCL team, but uncertain how much of their whole methodology it encompasses.
3. Originally used GOLabeler, but its web server is offline and author communications referred to NetGO as its continuation.

a. A local version is available and database version control is possible, but installation and customizability is non-trivial.

This paper describes Wei2GO: a weighted sequence similarity and python-based open-source function prediction software. Wei2GO utilizes DIAMOND and HMMER searchers against the UniProtKB and Pfam databases respectively [1, 4-6], extracts related Gene Ontology (GO) terms [7], and calculates several weighted scores to accurately associate probabilities to these GO terms. Wei2GO is similar in concept to the web-based Argot2.5 and its predecessor Argot2, which were top performers in CAFA3 and CAFA2 respectively [2, 8-10]. Due to this shared niche, this paper provides comparisons between these tools and Wei2GO using precision recall curves, Fmax scores, and a comparison of computational time for their predictions. Finally, Wei2GO is assessed on three scenarios where it annotates part of the triacylglycerol production pathway of the green algae *Chlamydomonas reinhardtii*, part of the Toll pathway of the mosquito *Anopheles gambiae*, and the Twin Arginine Targeting (TAT) complex of the bacteria *Pseudomonas putida* protein secretion system.

Methods

Wei2GO is a function prediction tool based on sequence similarity with reference proteins. It uses two of the most common similarity algorithms, DIAMOND and HMMER, against the UniProtKB and Pfam database respectively [1, 4-6]. Each respective hit is associated with Gene Ontology (GO) terms, taken from the Gene Ontology Annotation (GOA) database for UniProtKB and pfam2go for Pfam [1, 7, 22, 23]. Each GO term associated with a hit is given a weight relative to its e-value, their database of origin, and associated annotation evidence. The weights for each GO term are summed across the DIAMOND and HMMER matches. Based on their semantic similarity, the weights of similar GO terms are obtained to get a group score. Finally the GO weight, its information content, and group score are used to calculate the final score associated with the prediction. Several of the equations in these methods are adaptations of those employed by Argot2 [9].

Assigning weights to Gene Ontology terms

Wei2GO requires two input files:

1) Diamond or BLASTP hits against the UniProtKB database in tabular format.
2) HMMScan hits against the Pfam-A database with the option --tblout.

Each protein matched with a UniProtKB reference is associated with GO terms from the GOA database. Pfam domains are assigned GO terms through pfam2go. All GO terms *i* are given a weight *W* by taking the sum of the e-value *E* of all protein and domain hits associated with the GO term. GO terms associated with UniProtKB hits having the evidence tag 'Inferred from Electronic Annotation' (IEA) are weighed using equation 1 (EQ1), and GO terms with any evidence other than IEA, and Pfam associated GO terms, are weighed using EQ2. All weights for a GO term are subsequently summed.

$$W(go_i) = -log\,(Ego_i) \qquad (1)$$

$$W(go_i) = -log\,(Ego_i) * 10 \qquad (2)$$

In accordance with the GO Directed Acyclic Graph (DAG) structure, all weights associated to child terms of a GO term are added to its weight.

Internal confidence and group score

The internal confidence (InC) is calculated by taking the relative weight of the GO term compared to the weight of the root node.

$$InC(go_i) = \frac{Wgo_i}{Wgo_{root}} \qquad (3)$$

The group score (GS) of each GO term is calculated by taking the sum of the internal confidence of all GO terms having a semantic similarity of 0.7 or higher with the GO term:

$$GS(GO_i) = \sum InC[Sim(go_i, go_j) \geq 0.7] \qquad (4)$$

With the semantic similarity (Sim) calculated using Lin's formula [24] as:

$$Sim(go_i, go_j) = 2 * \frac{max(IC(go_i), IC(go_j))}{IC(go_i) + IC(go_j)} \qquad (5)$$

And the Information Content (IC) calculated as:

$$IC(GO_i) = -log\left(\frac{|\{go : go_i \in GOA\}|}{|\{go : go \in GOA\}|}\right) \qquad (6)$$

Total score

Finally, the total score (TS) for each GO term is calculated as:

$$TS(go_i) = IC(go_i) * \frac{InC(go_i)}{GS(go_i)} * W(go_i) \qquad (7)$$

Test set generation

To assess the performance of Wei2GO it is compared to Argot2, Argot2.5, and the baseline DIAMOND [6, 8, 9]. All proteins in the GOA database created between 01-01-2017 and 01-03-2020 were extracted together with manually validated GO terms. These were filtered to only contain proteins having five or more GO terms with the evidence code EXP, IDA, IPI, IMP, IGI, IEP, TAS, and IC. This was done to reduce the overestimation of false positive predictions, as a result of the open world of protein function [25]. The rationale behind this is that proteins which have five or more manually validated GO terms are more complete than proteins which have fewer manually validated GO terms, reducing the chance of falsely assigning false positive predictions. Finally, this set was expanded by including all terms belonging to one of these GO terms according to their GO DAG structure using the 'is_a' and 'part_of' relationships for parent terms, and 'has_part' for child terms. The final test set consists of 553 proteins with 34,700 assigned GO terms of which 537 proteins have 29,645 biological process terms, 256 proteins have 2,165 molecular function terms, and 237 proteins have 2,886 cellular component terms.

Test set evaluation

Precision and recall are used for the assessment of Wei2GO. These are calculated as:

$$Precision = \frac{True\ positives}{True\ positives + False\ positives} \qquad (8)$$

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives} \qquad (9)$$

Where a true positive is a correct prediction above a certain probability threshold, a false positive is an incorrect prediction above a probability threshold, and a false negative is a correct prediction below a probability threshold.

Additionally, the optimal harmony between precision and recall is calculated using the maximum F1-score, by calculating the F1-score for every possible Total Score threshold:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (10)$$

DIAMOND baseline calculations

GO terms for the DIAMOND baseline are taken according to the GOA associated terms of the match with the lowest e-value. The probability score used is the percentage identity between the query and target match. Only matches with SwissProt proteins are considered.

Materials

The following datasets and software versions were used:

- UniProt Knowledgebase release 2016-11 for DIAMOND searches
- Pfam-A release 30 for HMMER searches
- DIAMOND version 0.9.11.112
- HMMER version 3.3
- GOA UniProt data release 158 for creating the Wei2GO required data files
- GOA UniProt data release of 13-03-2020 for generating the test set
- go.obo release 02-12-2016
- pfam2go generated based on interpro2go release 2016/10/31
- ec2go release 23-03-2020
- Argot2 and Argot2.5 were last accessed on 20-04-2020

Results

Two additional Wei2GO versions were created to observe the impact of different features of the algorithm. Wei2GO-unweighted removes the weight multiplier for HMMER Pfam hits and non-IEA GO terms for UniProt DIAMOND hits, which is not present in Argot. Wei2GO-enhanced enhances the Pfam hits by adding the GOA assigned GO terms belonging to member proteins of the PFAM domain, which is a feature of Argot.

Table 2 summarizes the raw predictions of each method. Wei2GO and Wei2GO-unweighted show substantially fewer predictions than Wei2GO-enhanced and Argot, due to the enhancement of Pfam hits by the latter methods. There is no notable effect on the amount of correct predictions, as shown by the marginal gain in correct predictions for Wei2GO-enhanced. There is a large effect on the correct predictions versus total predictions ratio, which is 0.43 in Wei2GO and Wei2GO-unweighted and less than half in Wei2GO-enhanced. Similarly, Argot2 and Argot2.5 have a substantially lower correct prediction ratio compared to Wei2GO.

Precision and recall assessment

Precision recall curves were generated for biological processes (BP), molecular functions (MF), and cellular components (CC) (Figure 1). These curves show an increase in precision for Wei2GO, Wei2GO-unweighted, and Wei2GO-enhanced over Argot2 and Argot2.5, most notably for the BP category. Adding extra weight to Pfam and non-IEA GO terms has a notable effect on the accuracy, as seen by the differences between Wei2GO and Wei2GO-unweighted. Wei2GO-enhanced shows a more similar curve to Argot, which is in line with the number of correct predictions relative to the total predictions displayed in Table 2.

Table 2: Summary statistics for each method and its predictions. Displayed are the total number of predictions, number of proteins with at least one GO term predicted, number of correct predictions, and the ratio of correct predictions.

Method	Total predictions	Proteins with GO terms	Correct predictions	Correct prediction ratio
DIAMOND	38,379	511	24,533	0.64
Argot2	775,198	514	25,817	0.033
Argot2.5	176,616	511	13,718	0.08
Wei2GO	62,000	519	26,712	0.43
Wei2GO-unweighted	62,000	519	26,712	0.43
Wei2GO-enhanced	134,099	526	27,818	0.20

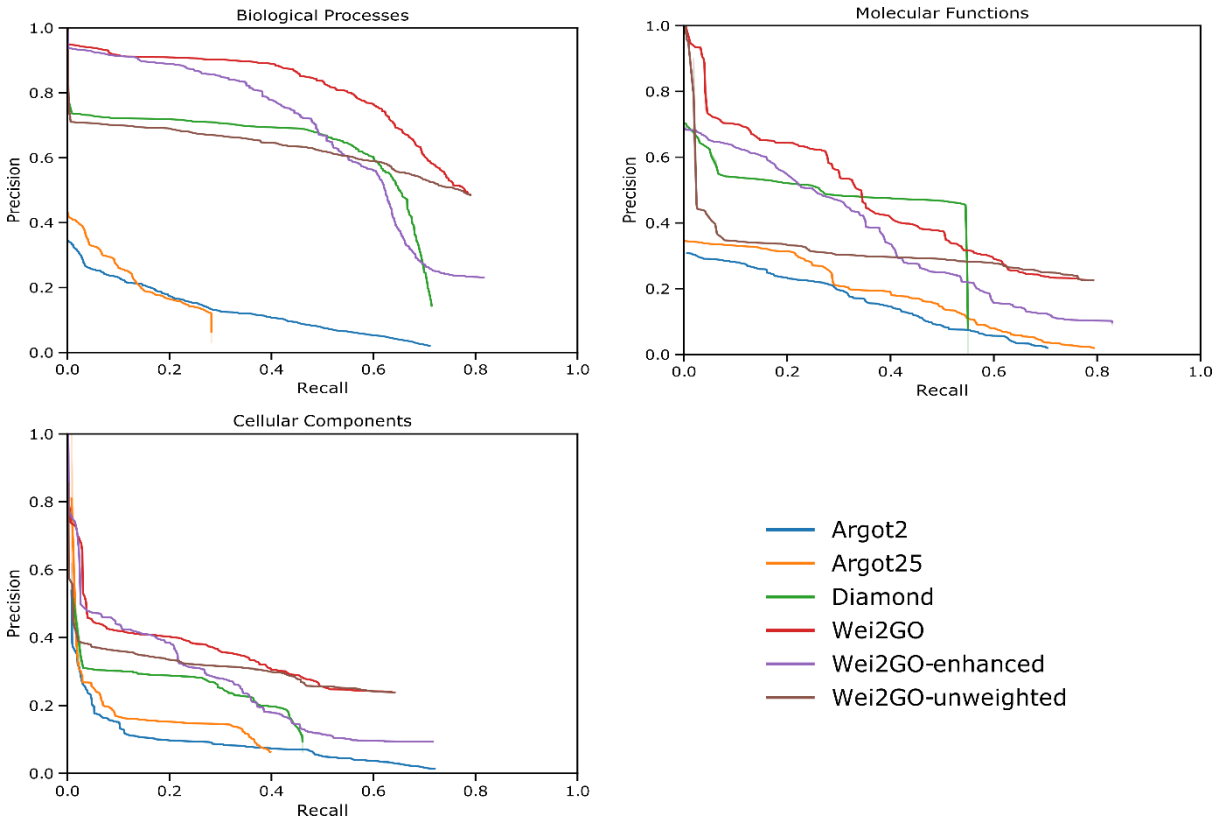


Figure 1: Precision recall curves for biological process, molecular function, and cellular component terms. These curves are sorted on precision.

Fmax scores were calculated based on the highest F1-score at any given threshold (Table 3). The Fmax consistently shows the best results in both Wei2GO and Wei2GO-unweighted.

Finally, computational times were compared in table 4. For the assessment of Argot2.5 no taxonomy code was given as an input, in order to process all proteins at once.

Table 3: Comparison of the maximum F1-scores (Fmax) between methods for biological processes (BP), molecular functions (MF), and cellular components (CC).

Method	Fmax BP	Fmax MF	Fmax CC
DIAMOND	0.54	0.31	0.27
Argot2	0.21	0.23	0.16
Argot2.5	0.22	0.31	0.26
Wei2GO	0.67	0.40	0.35
Wei2GO-unweighted	0.63	0.34	0.35
Wei2GO-enhanced	0.58	0.34	0.29

Table 4: A comparison of the computational time required for the prediction of the test set in table 3 and figure 1.

Method	Computational time
Argot2	1 hours
Argot2.5	4 hours
Wei2GO	7 minutes
Wei2GO-unweighted	7 minutes
Wei2GO-enhanced	12 minutes

Pathway annotation analysis

Wei2GO was tested by annotating part of three KEGG pathways maps: glycerolipid metabolism for the green algae *Chlamydomonas reinhardtii*, Toll and Imd signalling pathway for the mosquito *Anopheles gambiae*, and the twin arginine targeting (TAT) proteins for the bacterial secretion system pathway of *Pseudomonas putida [26]*. The cut-off probability score for all annotations is 500.

Triacylglycerol metabolism of *Chlamydomonas reinhardtii*

The single-celled *Chlamydomonas reinhardtii* is considered one of the model species of green algae, due to its ease of cultivation and relatively fast duplication rate. As of writing, 340 *C. reinhardtii* proteins have a reviewed status in UniProt, with its proteome consisting of almost 20,000 proteins. One of the interesting products of green algae is the membrane lipid triacylglycerol, a biofuel and bioplastic precursor amongst others, which can be the overwhelming part of an algae's biomass composition given the right growth conditions.

Table 5 shows the ability of Wei2GO to annotate *C. reinhardtii's* triacylglycerol production pathway from glycerol, with the KEGG and UniProt protein annotations as a reference (KEGG map 00561). GO annotations are converted to Enzyme Codes (EC) using ec2go [22]. The table shows that contrary to the existing UniProt GO annotations, Wei2GO is able to

allocate at least one protein to each enzyme member. Only enzyme 3.1.1.3 seems to be a large discrepancy in comparison.

Table 15: Enzymatic annotations of a subset of the glycerolipid metabolism pathway in KEGG. Compared are the existing KEGG annotations for Chlamydomonas reinhardtii, the existing UniProt annotations, and the Wei2GO annotations. Green indicates the same number, red a lower number, and blue a higher number of EC annotations compared to KEGG.

Enzyme	KEGG	Existing	Wei2GO
2.3.1.15	2	2	2
2.3.1.158	1	0	3
2.3.1.20	6	1	11
2.3.1.51	2	2	3
2.7.1.107	1	2	4
2.7.1.30	1	1	1
3.1.1.23	2	0	3
3.1.1.3	2	10	16
3.1.3.4	1	4	3

It is important to take into account that KEGG allocates proteins to a pathway after manual review, leaving it open to under representations of allocated enzymes. Nonetheless, an 800% increase is a disproportionate increase. BLAST searches of the two KEGG proteins assigned to enzyme 3.1.1.3 against the *C. reinhardtii* proteome leads to seven significant BLAST hits, indicating that at least a proportion of the discrepancy can be explained.

Toll signaling pathway of *Anopheles gambiae*

The mosquito *Anopheles gambiae* is a major vector of the malaria parasite *Plasmodium falciparum*, and is a model system for studying mosquito biology. As of writing, 253 *A. gambiae* proteins have a reviewed status in UniProt, with its proteome consisting of over 13,000 proteins. An important component is the Toll signaling pathway that is involved in activating defense responses upon recognising molecules from broad groups of microbial pathogens.

Table 6 compares the Wei2GO annotations of several *A. gambiae* proteins involved in a subset of the Toll pathway signaling (defined by KEGG map 04624) against existing UniProt annotations. The table shows that most GO terms annotated to pathway members are retrieved by Wei2GO. GNBP3 (GNBPA2) is the only protein with multiple false positive GO terms, with Wei2GO predicting functions suggestive of extracellular RNA-based functionality.

In contrast no GO terms above the threshold are returned for Spz1. Notably however, Spz1 is unusual as all its GO terms are manually annotated. This is likely why similarity-based annotations fail to retrieve these functions, as sequence homologs are either unavailable or are not annotated with these terms.

Additionally, the failure to annotate Spz1 with GO terms indicates that Wei2GO is not reliant on sequence homology with one protein, as the *Anopheles gambiae* proteins were not removed prior to the DIAMOND and HMMER searches required for Wei2GO. Instead it indicates that Wei2GO is reliant on groups of homologous proteins.

Table 6: A comparison of functional annotations for proteins involved in a subset of the Anopheles gambiae Toll and Imd pathway as represented by KEGG. Existing annotations are taken from the UniProt database, and are compared against Wei2GO annotations. In green are Wei2GO annotated GO terms which are found back in the existing annotations (true positive), in red are Wei2GO annotated GO terms which are not found back in the existing annotations (false positive), and in yellow are existing annotations which are manually annotated.

Protein	Existing	Wei2GO
GNBP3 (GNBPA2)	Carbohydrate binding Carbohydrate metabolic process Hydrolase activity, hydrolyzing O-glycosyl compounds	Carbohydrate binding Carbohydrate metabolic process Hydrolase activity, hydrolyzing O-Glycosyl compounds RNA processing RNA methyltransferase activity Extracellular region Regulation of innate immune response
GNBP1 (GNBPA1)	Carbohydrate binding Carbohydrate metabolic process Hydrolase activity, hydrolyzing O-glycosyl compounds	Carbohydrate binding Carbohydrate metabolic process Hydrolase activity, hydrolyzing O-Glycosyl compounds
PGRP-SA (PGRPS1)	Innate immune response N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan binding Peptidoglycan catabolic process Zinc ion binding	N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan catabolic process Innate immune response Zinc ion binding
PGRP-SA (PGRPS2)	Innate immune response N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan binding Peptidoglycan catabolic process Zinc ion binding	Innate immune response N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan catabolic process Zinc ion binding
PGRP-SA (PGRPS3)	Innate immune response N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan binding Peptidoglycan catabolic process Zinc ion binding	Innate immune response N-acetylmuramoyl-L-alanine amidase activity Peptidoglycan catabolic process Zinc ion binding
ModSP (AGAP001798)	Serine-type endopeptidase activity	Protein binding Serine-type endopeptidase activity
PSH1 (CLIPC5)	Serine-type endopeptidase activity	Serine-type endopeptidase activity
PSH2 (CLIPC6)	Serine-type endopeptidase activity	Serine-type endopeptidase activity
Spz1 (SPZ1)	Central nervous system formation Extracellular region Growth factor activity iInate immune response Regulation of Toll signaling pathway Toll binding	
TOLL (TOLL1A)	Innate immune response Integral component of membrane Signal transduction	Innate immune response Protein binding Signal transduction Integral component of membrane

Twin Arginine Targeting (TAT) annotation of *Pseudomonas putida KT4220*

Pseudomonas putida KT4220 is a gram-negative bacteria of interest for biotechnological purposes. 725 of its proteins in UniProt have a reviewed status, and its proteome consists of over 5,500 proteins. It is able to produce Poly-3-hydroxyalkanoates (PHA), an environmental friendly polyester, from the hydrocarbon styrene, a toxic pollutant present at many industrial sites. In nature, it has a mutualistic relationship with plants through their roots. A necessary part of any mutualistic relationship is the uptake and secretion of molecules and proteins. Table 7 shows the ability of Wei2GO to annotate Twin Arginine Targeting (TAT) proteins (KEGG map 03070), which is a highly conserved complex able to transfer folded proteins across cytoplasmic membranes. These proteins are associated with two specific GO terms: the cellular component term 'TAT protein transport complex', and the biological process term 'protein transport by the TAT complex'. Wei2GO is able to retrieve these terms for all TAT proteins with the exception of the term 'TAT protein transport complex' for tatB-1. However, the KEGG associated protein for *P. putida* KT4220 is not associated with this GO term either, indicating that the absence of this term is possible not a false positive but a true negative.

Table 7: An overview of the Wei2GO annotation of the 'TAT protein transport complex' cellular component GO term, and the 'Protein transport by the TAT complex' biological process GO term for the TAT complex of Pseudomonas putida K4220.

Protein	TAT protein transport complex	Protein transport by the TAT complex
tatA-1	✓	✓
tatA-2	✓	✓
tatB-1	✗	✓
tatB-2	✓	✓
tatC-1	✓	✓
tatC-2	✓	✓

Discussion

This paper described Wei2GO: a weighted sequence similarity and python-based open source function prediction software. It utilizes a combination of DIAMOND and HMMER searches, and combines these to extract a probability score associated with Gene Ontology (GO) terms. It was compared against the web-based tools Argot2.5 and Argot2, which share a similar concept.

The results section shows an improved precision and recall for biological processes, molecular functions, and cellular components compared to the DIAMOND baseline and Argot (Figure 1), as well as an increased Fmax score (Table 3). This difference is likely due to Wei2GO applying different weight modifiers depending on the evidence and databases of reference, as seen by a decrease in performance by Wei2GO-unweighted, and due to Wei2GO not enhancing Pfam domain matches with additional GO terms, as seen by a decrease in performance by Wei2GO-enhanced. The raw prediction numbers shown by Table 2 indicates that particularly the latter is a reason why Argot performs worse than the

DIAMOND baseline, given the low ratio of correct predictions. In fact, the purpose of enhancing Pfam hits with GO terms was originally to increase the recall of the predictions, but there is no observed difference in recall between Argot2 and Wei2GO. Argot2.5, which filters out GO terms based on taxonomic restrictions, even has a lower recall than Wei2GO. Additionally, computational times for Wei2GO predictions were substantially lower compared to the web-based Argot2.5 and Argot2 tools (Table 4). These results conclude that Wei2GO is not only an alternative open-source local version compared to existing sequence similarity based GO prediction methods, but is also an improvement by multiple accuracy performance standards.

To assess the biological relevance of Wei2GO, it annotated parts of three genomic pathways: triacylglycerol metabolism for the green algae *Chlamydomonas reinhardtii* (Table 5), Tol signaling for the mosquito *Anopheles gambiae* (Table 6), and the Twin Arginine Targeting (TAT) complex for the bacteria *Pseudomonas putida K4220* (Table 7). These re-annotations showed the ability of Wei2GO to annotate real data in a way that makes biological sense. For the enzymatic triacylglycerol pathway it was able to allocate at least one protein to all involved enzymes contrary to existing UniProt annotations, indicating that Wei2GO can successfully utilize remote homologs. Correct annotation of all but one of the Tol signalling members showed its ability to annotate non-enzymatic pathways. Correctly annotating all TAT complex members shows its ability to correctly predict proteins heavily influenced by cellular component terms.

Implementation and availability

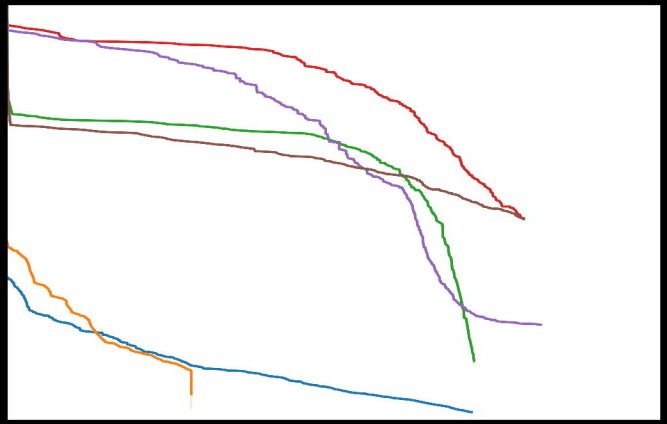
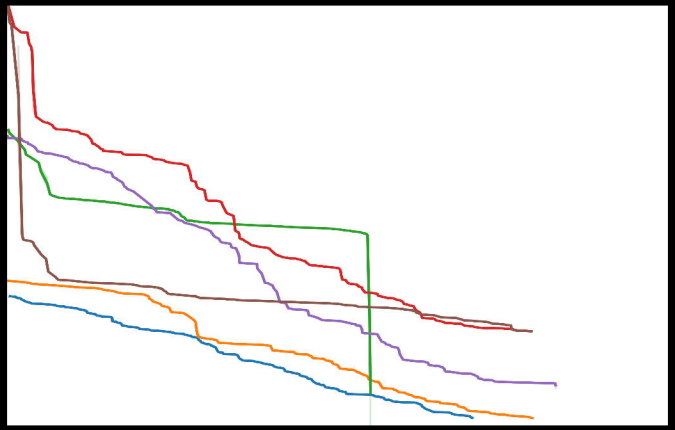
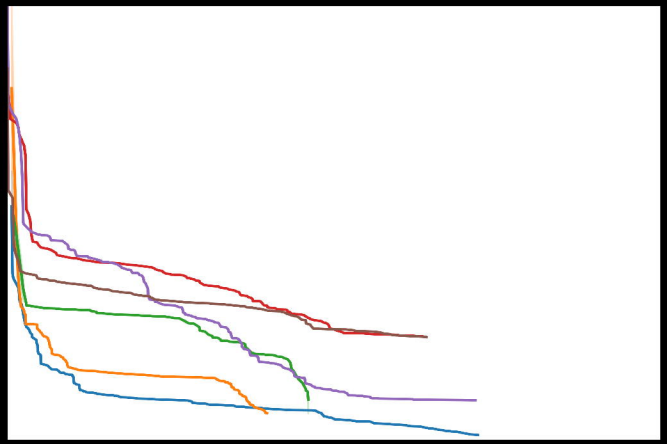
Wei2GO requires Python 3.6 or higher, and the NumPy package . Supplementary files used to create the results in this paper, instructions on how to install the software, and instructions on how to run the software can be found on https://gitlab.com/mreijnders/Wei2GO. Additionally, Snakemake scripts are provided to run the full annotation pipeline, and update the reference data to their latest versions.

Acknowledgements

The author would like to thank Dr. Robert M. Waterhouse for offering suggestions and editing parts of the manuscript, and Dr. Bastian VH Hornung for proofreading the manuscript.

This study is funded by Swiss National Science Foundation grant PP00P3_170664.

References

1. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic acids research, 2019. **47**(D1): p. D506-D515.
2. Zhou, N., et al., *The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens.* bioRxiv, 2019: p. 653105.
3. Reijnders, M.J.M.F., *CrowdGO: a wisdom of the crowd-based Gene Ontology annotation tool.* BioRxiv, 2019: p. 731596.
4. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic acids research, 2018. **46**(W1): p. W200-W204.
5. El-Gebali, S., et al., *The Pfam protein families database in 2019.* Nucleic acids research, 2019. **47**(D1): p. D427-D432.
6. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. **12**(1): p. 59.

7. Gene Ontology, C., *The gene ontology resource: 20 years and still GOing strong.* Nucleic acids research, 2018. **47**(D1): p. D330-D338.
8. Lavezzo, E., et al., *Enhancing protein function prediction with taxonomic constraints--The Argot2.5 web server.* Methods, 2016. **93**: p. 15-23.
9. Falda, M., et al., *Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.* BMC Bioinformatics, 2012. **13**(Suppl 4): p. S14.
10. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.
11. Rifaioglu, A.S., et al., *DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks.* Scientific reports, 2019. **9**(1): p. 1-16.
12. Törönen, P., A. Medlar, and L. Holm, *PANNZER2: a rapid functional annotation web server.* Nucleic acids research, 2018. **46**(W1): p. W84-W88.
13. Piovesan, D. and S.C.E. Tosatto, *INGA 2.0: improving protein function prediction for the dark proteome.* Nucleic acids research, 2019. **47**(W1): p. W373-W378.
14. Cozzetto, D., et al., *FFPred 3: feature-based function prediction for all Gene Ontology domains.* Scientific reports, 2016. **6**(1): p. 1-11.
15. Hawkins, T., et al., *PFP: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data.* Proteins: Structure, Function, and Bioinformatics, 2009. **74**(3): p. 566-582.
16. Das, S., et al., *CATH FunFHMMer web server: protein functional annotations using functional family assignments.* Nucleic acids research, 2015. **43**(W1): p. W148-W153.
17. Hallab, A., *Protein Function Prediction Using Phylogenomics, Domain Architecture Analysis, Data Integration, and Lexical Scoring*. 2015, Universitäts-und Landesbibliothek Bonn.
18. Lan, L., et al. *MS-k NN: protein function prediction by integrating multiple data sources*. in *BMC bioinformatics*. 2013. BioMed Central.
19. Gong, Q., W. Ning, and W. Tian, *GoFDR: a sequence alignment based method for predicting protein functions.* Methods, 2016. **93**: p. 3-14.
20. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nature methods, 2015. **12**(1): p. 7.
21. You, R., et al., *NetGO: improving large-scale protein function prediction with massive network information.* Nucleic acids research, 2019. **47**(W1): p. W379-W387.
22. Huntley, R.P., et al., *The GOA database: gene ontology annotation updates for 2015.* Nucleic acids research, 2015. **43**(D1): p. D1057-D1063.
23. Mitchell, A., et al., *The InterPro protein families database: the classification resource after 15 years.* Nucleic acids research, 2015. **43**(D1): p. D213-D221.
24. Lin, D. *An information-theoretic definition of similarity*. in *Icml*. 1998. Citeseer.
25. Dessimoz, C., N. Škunca, and P.D. Thomas, *CAFA and the open world of protein function predictions.* Trends in Genetics, 2013. **29**(11): p. 609-610.
26. Kanehisa, M., et al., *New approach for understanding genome variations in KEGG.* Nucleic acids research, 2019. **47**(D1): p. D590-D595.