Nucleic Acids Research, 2018 **1**
doi: 10.1093/nar/gky320

BUSCA: an integrative web server to predict subcellular localization of proteins

Castrense Savojardo[1], Pier Luigi Martelli[1,*], Piero Fariselli[2], Giuseppe Profiti[1,3] and Rita Casadio[1,3]

[1]Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna 40100, Italy, [2]Department of Comparative Biomedicine and Food Science, University of Padova, Padova 35020, Italy and [3]Institute of Biomembrane, Bioenergetics and Molecular Biotechnologies, Italian National Research Council (CNR), Bari 70126, Italy

Received January 24, 2018; Revised April 12, 2018; Editorial Decision April 14, 2018; Accepted April 17, 2018

ABSTRACT

Here, we present BUSCA (http://busca.biocomp.unibo.it), a novel web server that integrates different computational tools for predicting protein subcellular localization. BUSCA combines methods for identifying signal and transit peptides (DeepSig and TPpred3), GPI-anchors (PredGPI) and transmembrane domains (ENSEMBLE3.0 and BetAware) with tools for discriminating subcellular localization of both globular and membrane proteins (BaCelLo, MemLoci and SChloro). Outcomes from the different tools are processed and integrated for annotating subcellular localization of both eukaryotic and bacterial protein sequences. We benchmark BUSCA against protein targets derived from recent CAFA experiments and other specific data sets, reporting performance at the state-of-the-art. BUSCA scores better than all other evaluated methods on 2732 targets from CAFA2, with a F1 value equal to 0.49 and among the best methods when predicting targets from CAFA3. We propose BUSCA as an integrated and accurate resource for the annotation of protein subcellular localization.

INTRODUCTION

Subcellular localization is one of the main aspects defining protein function. Proteins have evolved to be functional in specific subcellular compartments: the biological processes directing a nascent protein sequence to its target destination, referred to as protein sorting, are still far from being completely understood and characterized. Computational methods aiming at predicting subcellular localization of proteins play a major role in large-scale functional annotation projects, whose ultimate goal is to understand the role of each protein in the context of cell complexity. For this reason, many methods have been developed in the last years (and they are reviewed in (1–3)).

The Bologna Unified Subcellular Component Annotator (BUSCA) (http://busca.biocomp.unibo.it) is a novel webserver integrating several published methods designed by the Bologna Biocomputing Group for the prediction of specific sub-cellular localization starting from protein sequence. The BUSCA annotation system relies on different machine/deep-learning approaches, devised to recognize and predict features that are relevant to determine protein subcellular localization.

More specifically, BUSCA integrates tools belonging to two different categories. The first includes methods suited to identify, along the input protein sequence, localization-related features such as signal and transit peptides, glycophosphatidylinositol (GPI) anchors and transmembrane domains (both α-helical and β-barrel). In particular, secretory signal peptides (and their respective cleavage sites) are identified using DeepSig (4), a recently developed approach based on deep-learning methods. Mitochondrial and chloroplast transit peptides are detected with the TPpred3 predictor (5). The presence and location of GPI-anchoring domains are predicted with PredGPI (6). Helical and beta stranded transmembrane domains are identified using ENSEMBLE3.0 (7) and BetAware (8), respectively.

The second category of methods comprises approaches devised to predict subcellular localization from sequence, namely BaCelLo (9) and MemLoci (10) for globular and membrane proteins, respectively, and SChloro (11), specialized on sub-chloroplast localization in plants.

These tools are organized by BUSCA in five prediction pipelines specific for animals, plants, fungi, Gram-positive and Gram-negative bacteria, respectively. Each pipeline predicts a different number of compartment classes by analyzing, processing and integrating the outcomes of the different tools.

*To whom correspondence should be addressed. Tel: +39 0512094005; Fax: +39 0512094005; Email: pierluigi.martelli@unibo.it

MATERIALS AND METHODS

Benchmarking datasets

We evaluated BUSCA and other methods against two benchmarking datasets obtained from the two most recent Critical Assessment of Function Annotation (CAFA) experiments (http://biofunctionprediction.org/cafa/). CAFA is an experiment designed to provide a large-scale assessment of computational methods for the prediction of protein function, including subcellular localization.

CAFA edition 2 (CAFA2) (12) was carried out in 2013. According to CAFA2 rules, the set of protein targets has been obtained by selecting all the sequences that acquired Gene Ontology (GO) experimental annotations for the cellular component sub-ontology, in the time-frame elapsed between releases 2013_12 and 2014_10 of the UniProtKB database. The following GO evidence codes are considered as experimental by CAFA (12): 'Inferred from experiment' (EXP), 'Inferred from direct assay' (IDA), 'Inferred from mutant phenotype' (IMP), 'Inferred from genetic interaction' (IGI), 'Inferred from expression pattern' (IEP), 'Traceable author statement' (TAS) and 'Inferred by curator' (IC). According to UniprotKB release 2014_10, CAFA2 benchmarking dataset comprises 2732 protein targets endowed with experimental GO annotation in the cellular component sub-ontology. We used this set to benchmark our BUSCA with respect to other approaches evaluated during the CAFA2 experiment (12). Proteins included in the dataset are distributed as follows: 2512 proteins are from animals, 26 from fungi, 105 from plants, 87 from Gram-negative and 2 from Gram-positive bacteria.

CAFA3 started in September 2016 and it is still ongoing. Since results are yet unreleased, we used available CAFA3 targets to simulate an in-house experiment. Among the initial 130,787 CAFA3 targets, we selected proteins that acquired GO Cellular Component (CC) annotations between January and December 2017. To this aim, we downloaded and compared the UniProt—Gene Ontology Annotation (UniProt-GOA, https://www.ebi.ac.uk/GOA) relative to releases 2016_12 and 2017_12 of UniProtKB. With this procedure, we ended-up with 3764 protein targets: 2559 from animals, 535 from fungi, 489 from plants, 165 from Gram-negative and 16 from Gram-positive bacteria. Table 1 summarizes the distribution of proteins among different cellular compartments in the CAFA 2 and 3 experiments: a protein is classified in a given compartment if it is endowed with an experimental GO term that is equal to or a descendant of the term associated to that compartment.

Furthermore, in order to cope with the low abundance of Gram-positive bacteria, we extracted from UniprotKB/SwissProt release 2018_03, all the protein sequences classified as *Firmicutes* and *Actinobacteria* and endowed with experimentally annotated subcellular localization. After homology reduction to 25% sequence identity and filtering-out protein sequences already included into BUSCA training sets, we ended up with a blind Gram-positive test set, including 1667 non-redundant protein sequences whose subcellular localizations are distributed as follows: 510 cytoplasmic, 245 extracellular and 912 plasma membrane.

BUSCA overview

Two main annotation pipelines are defined in BUSCA for processing Eukaryotic and Bacterial proteins, respectively, and their description is reported below.

Eukaryotic workflow. Eukaryotic proteins are processed using the general pipeline depicted in Figure 1. The pipeline is organized as a directed rooted computational graph where each node corresponds to the execution of a specific tool. The graph root is the query protein sequence, while leaves correspond to predicted subcellular localizations, here represented as GO terms of the cellular component ontology. A path from the root to one leaf is determined by the outcomes of the different tools. In Figure 1, GO terms and tools highlighted in green are only applied for plant proteins.

At the very first level, the query sequence is scanned for the presence of signal peptide using the DeepSig predictor (4). If the signal sequence is found (suggesting the sorting of the protein through the secretory pathway), the mature protein sequence is determined by cleaving the predicted signal peptide. The resulting mature sequence is then analyzed by the subsequent tools. Firstly, PredGPI (6) determines the presence of GPI-anchors. If an anchor is found, the sequence is classified as Membrane anchored component (GO:0046658). Otherwise, the sequence is filtered for the presence of α-helical TransMembrane (TM) domains using ENSEMBLE3.0 (7). If at least one TM domain is found, the protein is predicted as membrane protein and passed to MemLoci (10), which predicts the final membrane protein localization that includes: Endomembrane system (GO:00112505), Plasma membrane (GO:0005886) and Organelle membrane (GO:0031090). If no TM domain is found, the protein is predicted to be localized in the Extracellular space (GO:0005615).

Proteins not directed to the secretory pathway (as predicted with DeepSig) are analyzed for their potential organelle localization using TPpred3 (5), which predicts the presence of organelle-targeting peptides and distinguishes between mitochondrial and chloroplast sorting for plant proteins.

If no targeting peptide is detected with TPpred3, ENSEMBLE3.0 is used to discriminate membrane from globular proteins: MemLoci or BaCelLo (9) are hence applied to predict localization of membrane and globular protein, respectively. In particular, BaCelLo is able to distinguish among five different cellular compartments (four in case of animal or fungi proteins): Nucleus (GO:0005634), Cytoplasm (GO:0005737), Extracellular space (GO:0005615), Mitochondrion (GO:0005739) and, for plant proteins, Chloroplast (GO:0009507). Moreover, since BaCelLo adopts different optimized models for animals and fungi, information about the taxonomic origin of the input is also provided as a parameter to the predictor.

When a mitochondrial targeting signal is detected, this is cleaved-off to determine the mature protein sequence. ENSEMBLE3.0 is then used to determine whether the mature protein is localized into a Mitochondrial membrane (GO:0031966) or, more generally, into the Mitochondrion (GO:0005739).

Table 1. Distribution of proteins in CAFA2 and CAFA3 datasets among different subcellular compartments

	CAFA2					CAFA3				
Compartment	A	F	P	G-	G+	A	F	P	G-	G+
Nucleus	682	2	32	-	-	574	130	97	-	-
Extracellular	940	0	0	0	0	208	4	17	0	0
Organelle membrane	153	4	10	-	-	156	41	19	-	-
Endomembrane system	298	4	15	-	-	387	38	41	-	-
Lysosome	46	1	-	-	-	29	7	-	-	-
Cytoplasm	529	14	22	51	0	669	200	184	116	16
Mitochondrion	115	0	7	-	-	109	18	26	-	-
Peroxisome	7	1	2	-	-	5	7	9	-	-
Plasma membrane	338	5	27	29	2	368	37	44	44	0
Outer membrane	-	-	-	4	-	-	-	-	3	-
Chloroplast	-	-	26	-	-	-	-	111	-	0
Other compartments	363	6	5	11	-	590	82	32	28	0
Total	2512	26	105	87	2	2559	489	489	165	16

Labels are: A = animals, F = fungi, P = plants, G– = Gram-negative and G+ = Gram-positive.

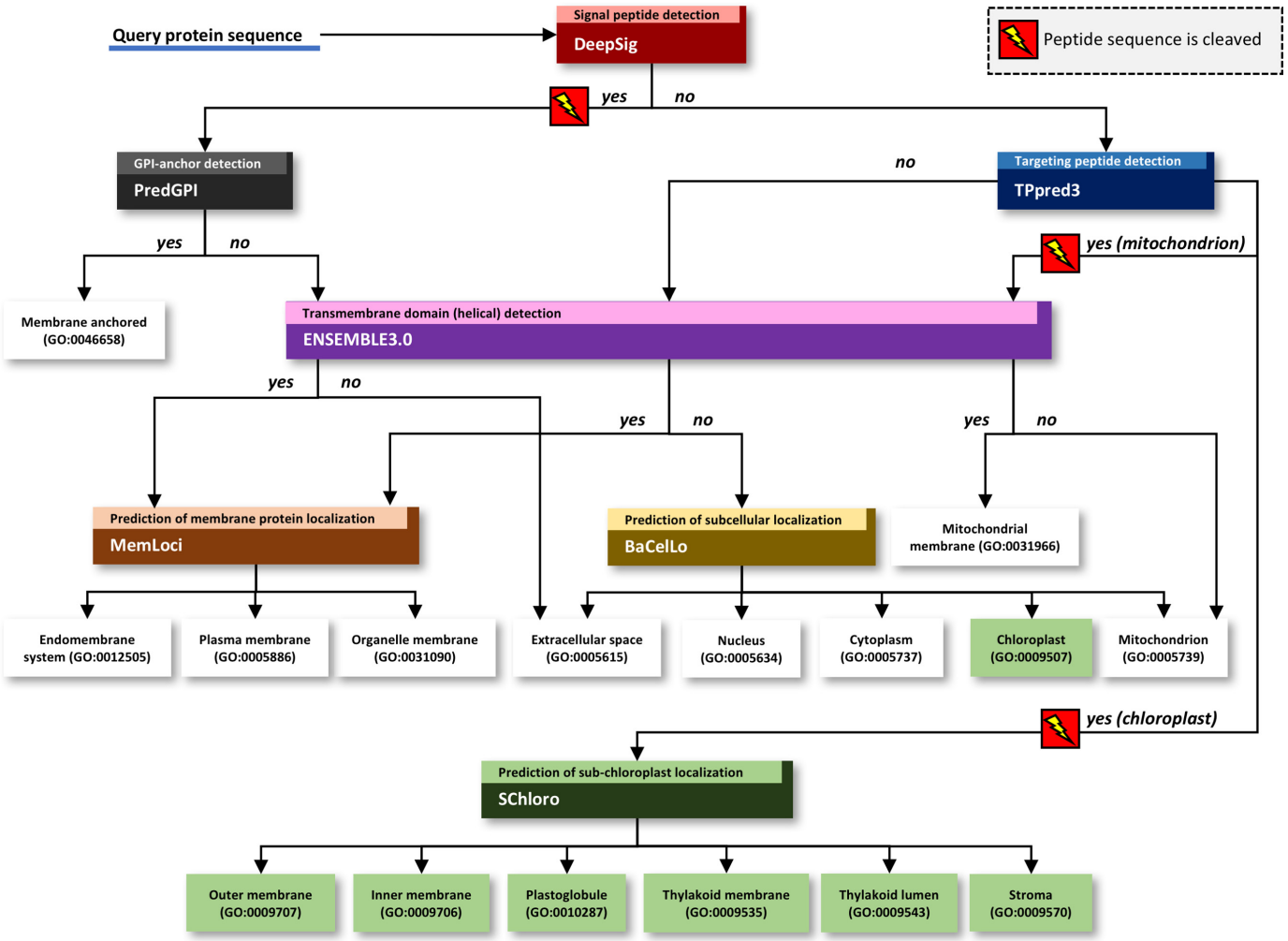

Figure 1. The Eukaryotic BUSCA workflow. The different methods are organized in a rooted computation graph. The query sequence is processed by different methods whose outputs determine the path from the root (the query sequence) to one leaf (a predicted subcellular location). Chloroplast-related localizations (and the respective tools), are highlighted in green and are relevant only for plant sequences. Overall, up to sixteen and nine different compartments are predicted for plants and other eukaryotes (i.e. animals and fungi), respectively.

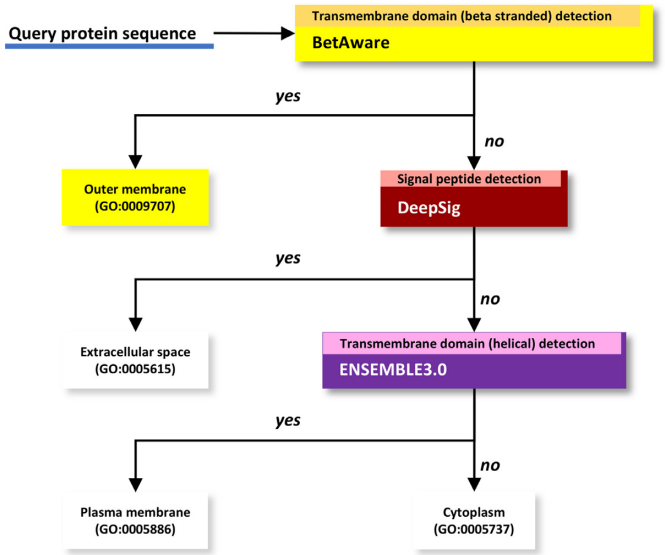

Figure 2. The Bacterial BUSCA workflow. BetAware (8) is only executed for Gram-negative bacteria. Overall, up to four and three different compartments are predicted for Gram-negative and Gram-positive bacteria, respectively.

For plant proteins, TPpred3 is also able to distinguish potential chloroplast-targeting peptides. If detected, they are cleaved and the sequence submitted to SChloro (11) that discriminates six different sub-chloroplast localizations: Outer membrane (GO:0009707), Inner membrane (GO:0009706), Plastoglobule (GO:0010287), Thylakoid lumen (GO:0009543), Thylakoid membrane (GO:0009535) and Stroma (GO:0009570).

Overall BUSCA is able to predict sixteen different compartments for plants and nine for animals and fungi.

Bacterial workflow. Bacterial proteins are processed using the pipeline depicted in Figure 2. In the first step, BetAware (8) is applied to detect a beta-barrel structure inserted into the bacterial Outer membrane (GO:0009707). Note that this step (highlighted in yellow) is performed only for proteins coming from Gram-negative bacteria, which are endowed with outer membranes. If BetAware does not recognize a beta-barrel domain (or if the protein belongs to a Gram-positive bacterium), the protein is analyzed with DeepSig (parametrized with respect to Gram-positive or negative classes, depending on the origin of the input sequence). When the signal is found, the sequence is classified as localized into the Extracellular space (GO:0005615). Otherwise, the protein is finally processed using ENSEMBLE3.0 to determine whether it is inserted into the Plasma membrane (GO:0005886) or it is localized into the Cytoplasm (GO:0005737).

Overall BUSCA predicts four different compartments for Gram-negative and three for Gram-positive bacteria.

Generation of sequence profiles

Tools included in BUSCA require, for each input sequence, the computation of a sequence profile from a multiple sequence alignment. These are computed using three runs of PSI-BLAST, with e-value threshold set to 1e–3 and using the UniprotKB/SwissProt (release 2017_11) as backend database.

Performance evaluation

Different scoring measures evaluate prediction performance of BUSCA and other methods.

For sake of comparison with other approaches scored in the context of the CAFA experiments, methods are evaluated using the $F1$ score, defined as the harmonic mean between precision and recall, both computed at the level of GO-term assignments. In particular, precision and recall for GO-term prediction are computed as follows:

$$Pr = \frac{1}{m}\sum_{i=1}^{m}\frac{\sum_{t\in P_i} 1_{\{t\in T_i\}}}{|P_i|}$$

$$Rc = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{t\in T_i} 1_{\{t\in P_i\}}}{|T_i|}$$

where m and n are the number of proteins for which at least one GO term has been predicted and the total number of proteins in a dataset, respectively; P_i and T_i are the sets of predicted and annotated GO terms for the i-th protein, respectively. The function $1_{\{c\}}$ is an indicator that equals 1 when the condition c is true, 0 otherwise, and $|\cdot|$ indicates the cardinality of the sets P_i and T_i, respectively. Complete GO annotations including all the ancestors of predicted or annotated GO terms are considered in the computation of the scores.

$F1$ is then defined as:

$$F1 = \frac{2 \times Pr \times Rc}{Pr + Rc}$$

For some method (including BUSCA), $m = n$, since a prediction is always provided for any given input protein sequence. In general, $m \leq n$, and hence we define the coverage as the ratio:

$$C = \frac{m}{n}$$

C measures the fraction of benchmark proteins for which the method provides prediction.

Methods are also scored at the level of individual subcellular compartments. To this aim, we use the Matthews Correlation Coefficient (MCC), defined as:

$$MCC = \frac{(TP_i \times TN_i - FP_i \times FN_i)}{\sqrt{(TP_i + FP_i) \times (TP_i + FN_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}}$$

where TP_i, TN_i, FP_i and FN_i are, respectively, true positives, true negatives, false positives and false negatives computed for compartment i.

RESULTS

Assessing BUSCA performance on CAFA experiments

As a first benchmark, we evaluated the performance of BUSCA on CAFA2/3 targets using the evaluation procedure adopted in CAFA experiments.

Table 2. *F*1 and *C* scores obtained by different methods on the CAFA2 benchmark dataset.

Method	*F*1	*C*
BUSCA	0.49	1.0
EVEX (13)	0.46	0.98
Go-FDR (21)	0.46	0.98
dcGO (20)	0.46	0.99
MS-kNN (19)	0.45	0.98
FFPred (18)	0.45	1.0
CONS (16)	0.44	1.0
LocTree3 (17)	0.44	0.96
PULP (15)	0.44	0.94
FunFams (14)	0.43	0.89
BLAST	0.35	0.98

Results for all methods except BUSCA were taken from (12). *F*1, the harmonic mean of precision and recall, and *C*, the fraction of predicted protein in the dataset, are computed as described in the "Performance evaluation" section (see text).

In Table 2 we report results obtained with BUSCA and other methods on the CAFA2 targets. Results for all the methods except BUSCA were taken from the official CAFA2 paper (12). In particular, we selected published methods among the ten top performing approaches as scored by the CAFA2 assessors: EVEX (13), FunFams (Orengo Lab) (14), PULP (15), CONS (16), LocTree3 (ROSTLab) (17), FFpred (Jones-UCL Lab) (18), MS-kNN (19), dcGO (Gough Lab) (20) and Go-FDR (Tian Lab) (21). For sake of comparison, we also report the performance of a baseline method, transferring GO terms from the highest-scoring BLAST hit.

In Table 2, all the methods outperform the baseline BLAST approach. BUSCA outperforms all other methods, scoring with *F*1 equal to 0.49, three percentage points higher than the top scoring method (EVEX, $F1 = 0.46$). Concerning coverage, being BUSCA a purely predictive approach (like FFpred and CONS), it always provides a prediction ($C = 1$). This is not true for other tools: for instance, FunFams does not provide predictions for 11% of the tested proteins ($C = 0.89$).

Since CAFA2 targets were released before some of the methods included in BUSCA were developed (in particular, DeepSig, SChloro, TPpred3 and BetAware), we evaluated the performance of BUSCA selecting only those CAFA2 proteins that were not used to train such predictors. In particular, we identified 826 CAFA2 targets that were included into training sets. After removing them from the benchmark, the BUSCA F1 measure only decrease by 1 percentage point (from 0.49 to 0.48), indicating high robustness on prediction performance.

We further investigated BUSCA performance using the more recent CAFA3 benchmark. On this dataset, BUSCA scores with *F*1 equal to 0.58, nine percentage points higher than the one obtained on CAFA2 targets. This further confirms that BUSCA performs at the state-of-the-art. Unfortunately, results of the CAFA3 have not been published yet, so we cannot directly compare with other approaches using the CAFA evaluation procedure.

As done for the CAFA2 benchmark, we scored BUSCA after removing CAFA3 targets included into training sets. In this case, we identified 690 CAFA3 targets. Filtering-out these proteins from the benchmark we obtained exactly the same *F*1 score (0.58), highlighting again a small influence on the performance evaluation.

Evaluating the prediction performance at compartment level

CAFA3 targets were used to perform a comparative benchmark among different prediction tools including BUSCA, two recently developed methods, namely the ensemble method SubCons (22,23) and DeepLoc (24), based on deep-learning, as well as LocTree3 (17) and Cello2.5 (25). All methods run using the respective web servers and their predictions are scored at the level of individual compartments using the Matthews Correlation Coefficient (MCC) (see the Performance evaluation section). Specifically, ten different protein localizations are considered: nucleus, extracellular, cytoplasm, plasma membrane, endomembrane system, mitochondrion, peroxisome, lysosome, organelle membrane and chloroplast (in plants). This choice is the one that allows to fairly compare different tools which predict different output localizations. Moreover, to ensure a sufficient number of proteins and a reasonable coverage of the different compartments, methods are scored separately on Animals + Fungi, Plants and Prokaryotes (aggregating Gram-positive and Gram-negative proteins).

Table 3 lists the results. With the exception of Cello2.5, which shows very low MCCs for some compartments (e.g. endomembrane system in Plants), all methods perform quite well and their performances are in general similar.

In the Animal+Fungi dataset, LocTree3 and SubCons are better than others for mitochondrial proteins, DeepLoc outperforms other methods in extracellular and plasma membrane compartments while BUSCA is the best-performing method for nucleus, organelle membrane, endomembrane system and cytoplasm compartments.

When scored on plant proteins, BUSCA outperforms other methods in many compartments including extracellular space, organelle membrane, endomembrane system, cytoplasm and chloroplast (in this case with same performance of LocTree3).

Overall, thanks to specialized tools such as MemLoci, TPpred3, DeepSig and SChloro, the BUSCA pipeline is more effective for proteins localized into organelle membranes and endomembrane system as well as chloroplast proteins in plants. Given the present combination of tools included in BUSCA, our pipeline is not able to predict proteins that are localized in lysosomes and peroxisomes.

Among the five methods considered, only three, BUSCA, LocTree3 and Cello2.5, explicitly support prediction on prokaryotes. In this dataset, BUSCA and LocTree3 perform equally on cytoplasmic proteins while LocTree3 outperforms on plasma membrane proteins. Interestingly, only BUSCA was able to identify outer membrane beta-barrel proteins, thanks to the BetAware tool included in the pipeline.

The prokaryotic subset only contains 16 proteins from Gram-positives. In order to benchmark BUSCA on a larger number of Gram-positives sequences, we originated a blind dataset (see the "Benchmarking datasets" section). This set includes 1667 protein sequences with experimental annotation (510 cytoplasmic, 245 extracellular and 912 plasma

Table 3. Performance comparison of different methods on the CAFA3 dataset.

		Method MCC				
Compartment	Dataset	LocTree3	SubCons	DeepLoc	Cello2.5	BUSCA
Nucleus	Animals + Fungi	0.46	0.32	0.42	0.28	**0.48**
Extracellular	Animals + Fungi	0.35	0.37	**0.46**	0.28	0.42
Organelle membrane	Animals + Fungi	0.20	-	0.23	-	**0.24**
Endomembrane system	Animals + Fungi	0.21	0.18	0.21	0.07	**0.22**
Lysosome	Animals + Fungi	-	0.09	0.11	0.04	-
Cytoplasm	Animals + Fungi	0.29	0.25	0.29	0.12	**0.38**
Mitochondrion	Animals + Fungi	**0.42**	**0.42**	0.34	0.37	0.37
Peroxisome	Animals + Fungi	0.11	0.11	0.13	0.29	-
Plasma membrane	Animals + Fungi	0.35	0.38	**0.45**	0.26	0.37
Nucleus	Plants	**0.42**	0.30	0.39	0.31	0.41
Extracellular	Plants	0.21	0.30	0.33	0.10	**0.39**
Organelle membrane	Plants	0.11	-	0.13	-	**0.20**
Endomembrane system	Plants	0.26	0.11	0.26	0.00	**0.37**
Cytoplasm	Plants	0.38	0.40	0.30	0.25	**0.43**
Mitochondrion	Plants	0.29	**0.32**	0.29	0.25	0.27
Peroxisome	Plants	**0.18**	0.08	0.14	0.00	-
Plasma membrane	Plants	0.48	0.34	**0.51**	0.30	0.44
Chloroplast	Plants	**0.37**	-	0.22	0.23	**0.37**
Cytoplasm	Prokaryotes	**0.50**	-	-	0.37	**0.50**
Plasma membrane	Prokaryotes	**0.55**	-	-	0.40	0.50
Outer membrane	Prokaryotes	0.00	-	-	0.00	**0.32**

Methods are scored using the MCC, computed for individual compartments as detailed in the "Performance evaluation" section (see text). Predictions for SubCons (22,23), DeepLoc (24), Cello2.5 (25) and LocTree3 (17) were obtained using the respective web servers. Predictions on bacterial proteins are scored only for methods providing dedicated modules for prokaryotic data (i.e. LocTree3, Cello2.5 and BUSCA). Highest MCC scores for each compartment are highlighted in bold face.

membrane). Against this dataset, and considering the experimental annotations of UniprotKB, BUSCA scores with very good MCC values in all the three compartments (0.73, 0.40 and 0.74 MCC values for Cytoplasm, Extracellular and Plasma membrane compartments, respectively; see Supplementary Table S1). This further confirms the effectiveness of our pipeline also for the annotation of Gram-positive proteins.

Scoring BUSCA on feature annotation

We also evaluated the ability of BUSCA to discriminate localization-related features. We identified within the CAFA3 dataset proteins endowed with annotations (extracted from UniprotKB) for signal and transit peptides, GPI-anchors and transmembrane regions. Then, for each feature type, we evaluated the fraction of correctly identified experimental annotations. The value ranges from 0.69 for transit peptides, to 0.91, 0.93 and 1 for Transmembrane, Signal peptides and GPI-anchors, respectively.

The BUSCA web server

The BUSCA web server accepts as input protein sequences in FASTA format. The server accepts up to 500 protein sequences per-submission. Before submitting, the user is also asked to specify the taxonomic origin of the input protein sequences, choosing among five options: Animals, Fungi, Plants, Gram-positive and Gram-negative bacteria. Depending on the user's choice, the proper prediction protocol (see "Materials and Methods" section) is applied by BUSCA to process the input sequences. Upon request submission, if input sequences pass validation checks, the server automatically redirects to the final output page which is periodically reloaded until the prediction job is completed. Alternatively, the page can be bookmarked and accessed at a later stage.

A typical BUSCA output is shown in Figure 3. Prediction results for the submitted protein sequences are displayed in tabular format. For each input sequence, basic information includes the protein accession/identifier, the predicted GO-terms, the score assigned to the prediction, an alternative localization (when available) and a summary of features that have been predicted on the sequence. Prediction scores are probabilities attached to each predicted compartment and are internally computed by the method providing the final prediction. Specifically, all the tools included in BUSCA, with the only exception of BaCelLo, already compute a prediction score. Concerning BaCelLo, which is based on a decision tree of SVMs (9), the score is computed by remapping the value of the SVM discrimination function in the range [0,1]. The mapping has been computed using a logistic function whose parameters have been calibrated on the BaCelLo training dataset (9).

An alternative localization is reported only when the final predicted compartment is computed with BaCelLo, MemLoci and SChloro. In particular, SChloro supports prediction of multiple localizations: in this case, we provide the highest-scoring prediction as primary localization and other predicted compartments (if any) as alternative localizations. BaCelLo and MemLoci do not support multi-label prediction: in these cases, the alternative localization corresponds to the second most-probable compartment with a prediction score greater than 0.1.

The user can also open a detailed report for each sequence, which graphically displays the precise positions of predicted features along the sequence. Predicted local-

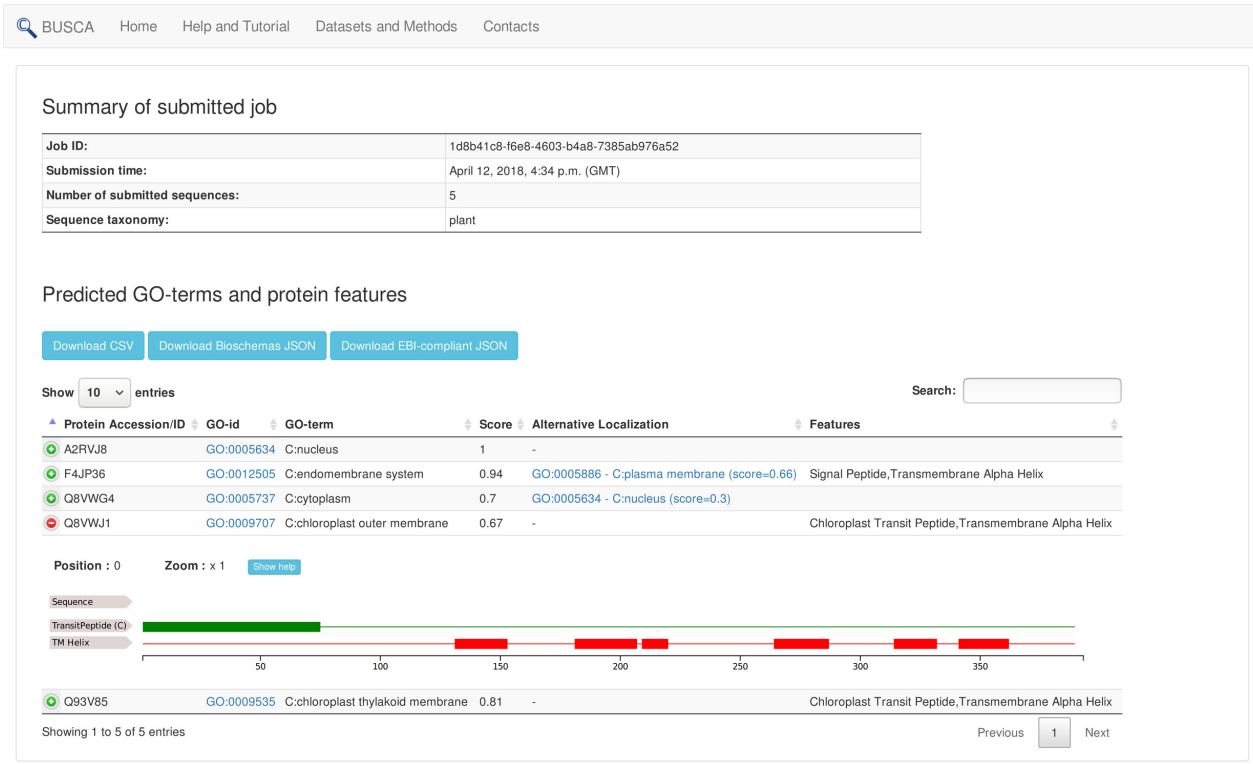

Figure 3. Example output from the BUSCA web server showing results for five different plant proteins. Protein accession/identifier, predicted GO terms, prediction score, alternative localization (if any) and protein features are reported for each input sequence. A detailed report for each protein graphically displays positions of predicted features along the sequence. .

izations and protein features can be also downloaded as Comma-Separated Values (CSVs) and/or JavaScript Object Notation (JSON) formats.

BUSCA is implemented using the Django Python Web framework (https://www.djangoproject.com). The user interface builds on top of technologies such as JQuery (https://jquery.com), Bootstrap (https://getbootstrap.com) and DataTables (https://datatables.net) to improve usability. Workflows aggregating the different tools are developed using an ad-hoc framework implemented using the Python programming language. In developing the framework, we mainly focused on reducing the effort needed in the future to update existing tools and/or to integrate new tools that will be made available.

The computation of sequence profiles represents a bottleneck that increases the computational time. In order to avoid unnecessary computations and to speed-up the prediction process, the required sequence profile is generated by a single run of PSI-BLAST whose result is subsequently shared by the different tools. Moreover, once generated, the sequence profile is cached, avoiding expensive rebuilding in case of resubmission of an identical protein sequence.

CONCLUSION

In this paper, we present BUSCA, a novel web server which centralizes several resources devised to predict protein subcellular localization, including protein feature predictors like DeepSig, TPPred3, PredGPI, BetAware and ENSEMBLE3.0, and protein localization predictors like MemLoci, BaCelLo and SChloro. Each predictor has been described and benchmarked before. Here the novelty is the rational integration of the tools into the BUSCA web server for allowing the prediction of subcellular localization in a systematic way, with the final goal of predicting the subcellular localization of the protein depending on the protein source. Furthermore, BUSCA also annotates relevant protein features such as signal/transit peptides, GPI-anchors and transmembrane domains providing a detailed characterization of query protein sequences.

We benchmark BUSCA on protein targets derived from the two most recent CAFA experiments using the same procedure described by CAFA assessors. By this, we can also compare our performances with other top-scoring available approaches. The results clearly indicate that BUSCA well compares with the state-of-the-art approaches, and it is somewhat superior in the assigned task.

We scored our method and other recently published approaches at the level of annotation of individual subcellular compartments. In these benchmarks, BUSCA performance is overall comparable to the ones achieved by other approaches.

The BUSCA web server has been implemented using modern web technologies to ensure usability and extensibility. Therefore, new retrained versions of the tools and/or new tools can easily be integrated into BUSCA, providing an updated and centralized resource for a large-scale annotation of protein subcellular localization.

DATA AVAILABILITY

BUSCA is available as web server at http://busca.biocomp.unibo.it.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

European Union RTD Framework Program [Action BM1405 to R.C.]; University of Bologna [R.F.O. 2017 to P.L.M.]. Funding for open access charge: University of Bologna [R.F.O. 2017 to R.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Casadio,R., Martelli,P.L. and Pierleoni,A. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic. Proteomic.*, **7**, 63–73.
2. Imai,K. and Nakai,K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.
3. Nielsen,H. (2017) Predicting Subcellular Localization of Proteins by Bioinformatic Algorithms. In: Bagnoli,F and Rappuoli,R (eds). *Protein and Sugar Export and Assembly in Gram-positive Bacteria*. Springer International Publishing, Cham, pp. 129–158.
4. Savojardo,C., Martelli,P.L., Fariselli,P. and Casadio,R. (2017) DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics*, doi:10.1093/bioinformatics/btx818.
5. Savojardo,C., Martelli,P.L., Fariselli,P. and Casadio,R. (2015) TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics*, **31**, 3269–3275.
6. Pierleoni,A., Martelli,P.L. and Casadio,R. (2008) PredGPI: a GPI-anchor predictor. *BMC Bioinformatics*, **9**, 392.
7. Martelli,P.L., Fariselli,P. and Casadio,R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics*, **19**(Suppl. 1), i205–i211.
8. Savojardo,C., Fariselli,P. and Casadio,R. (2013) BETAWARE: a machine-learning tool to detect and predict transmembrane beta-barrel proteins in prokaryotes. *Bioinformatics*, **29**, 504–505.
9. Pierleoni,A., Martelli,P.L., Fariselli,P. and Casadio,R. (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.
10. Pierleoni,A., Martelli,P.L. and Casadio,R. (2011) MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. *Bioinformatics*, **27**, 1224–1230.
11. Savojardo,C., Martelli,P.L., Fariselli,P. and Casadio,R. (2017) SChloro: directing Viridiplantae proteins to six chloroplastic sub-compartments. *Bioinformatics*, **33**, 347–353.
12. 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.
13. Van Landeghem,S., Hakala,K., Ronnqvist,S., Salakoski,T., Van de Peer,Y. and Ginter,F. (2012) Exploring biomolecular literature with EVEX: Connecting genes through events, homology, and indirect associations. *Adv. Bioinforma.*, **2012**, 582765.
14. Das,S., Lee,D., Sillitoe,I., Dawson,N.L., Lees,J.G. and Orengo,C.A. (2016) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. *Bioinformatics.*, **32**, 2889.
15. Youngs,N., Penfold-Brown,D., Drew,K., Shasha,D. and Bonneau,R. (2013) Parametric Bayesian priors and better choice of negative examples improve protein function prediction. *Bioinformatics*, **29**, 1190–1198.
16. 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.
17. Goldberg,T., Hecht,M., Hamp,T., Karl,T., Yachdav,G., Ahmed,N., Altermann,U., Angerer,P., Ansorge,S., Balasz,K. *et al.* (2014) LocTree3 prediction of localization. *Nucleic Acids Res.*, **42**, W350–W355.
18. Cozzetto,D., Buchan,D.W.A., Bryson,K. and Jones,D.T. (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics*, **14**(Suppl. 3), S1.
19. Lan,L., Djuric,N., Guo,Y. and Vucetic,S. (2013) MS-kNN: protein function prediction by integrating multiple data sources. *BMC Bioinformatics*, **14**(Suppl. 3), S8.
20. Fang,H. and Gough,J. (2013) A domain-centric solution to functional genomics via dcGO Predictor. *BMC Bioinformatics*, **14**(Suppl. 3), S9.
21. Gong,Q., Ning,W. and Tian,W. (2016) GoFDR: A sequence alignment based method for predicting protein functions. *Methods*, **93**, 3–14.
22. Salvatore,M., Warholm,P., Shu,N., Basile,W. and Elofsson,A. (2017) SubCons: a new ensemble method for improved human subcellular localization predictions. *Bioinformatics*, **33**, 2464–2470.
23. Salvatore,M., Shu,N. and Elofsson,A. (2018) The SubCons webserver: A user friendly web interface for state-of-the-art subcellular localization prediction. *Protein Sci.*, **27**, 195–201.
24. Almagro Armenteros,J.J., Sønderby,C.K., Sønderby,S.K., Nielsen,H. and Winther,O. (2017) DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, **33**, 3387–3395.
25. Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.