Hafida Bouziane* and Abdallah Chouarfia

Use of Chou's 5-steps rule to predict the subcellular localization of gram-negative and gram-positive bacterial proteins by multi-label learning based on gene ontology annotation and profile alignment

https://doi.org/10.1515/jib-2019-0091
Received November 8, 2019; accepted April 8, 2020; published online June 29, 2020

Abstract: To date, many proteins generated by large-scale genome sequencing projects are still uncharacterized and subject to intensive investigations by both experimental and computational means. Knowledge of protein subcellular localization (SCL) is of key importance for protein function elucidation. However, it remains a challenging task, especially for multiple sites proteins known to shuttle between cell compartments to perform their proper biological functions and proteins which do not have significant homology to proteins of known subcellular locations. Due to their low-cost and reasonable accuracy, machine learning-based methods have gained much attention in this context with the availability of a plethora of biological databases and annotated proteins for analysis and benchmarking. Various predictive models have been proposed to tackle the SCL problem, using different protein sequence features pertaining to the subcellular localization, however, the overwhelming majority of them focuses on single localization and cover very limited cellular locations. The prediction was basically established on sorting signals, amino acids compositions, and homology. To improve the prediction quality, focus is actually on knowledge information extracted from annotation databases, such as protein–protein interactions and Gene Ontology (GO) functional domains annotation which has been recently a widely adopted and essential information for learning systems. To deal with such problem, in the present study, we considered SCL prediction task as a multi-label learning problem and tried to label both single site and multiple sites unannotated bacterial protein sequences by mining proteins homology relationships using both GO terms of protein homologs and PSI-BLAST profiles. The experiments using 5-fold cross-validation tests on the benchmark datasets showed a significant improvement on the results obtained by the proposed consensus multi-label prediction model which discriminates six compartments for Gram-negative and five compartments for Gram-positive bacterial proteins.

Keywords: gene ontology terms; gram-negative bacteria; gram-positive bacteria; multi-label learning; profile alignment; subcellular localization prediction.

1 Introduction

Proteins are key players in cell survival and damage and their presence in specific cell sites reflects the nature of their biological function. Protein subcellular localization (SCL) knowledge is thus valuable for protein function elucidation which is crucial for drug design, discovery, and development. Once they are synthetized

***Corresponding author: Hafida Bouziane**, Département d'Informatique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB BP 1505, El M'Naouer, 31000, Oran, Algeria, E-mail: hafida.bouziane@univ-usto.dz
Abdallah Chouarfia: Département d'Informatique, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB BP 1505, El M'Naouer, 31000, Oran, Algeria, E-mail: abdallah.chouarfia@univ-usto.dz

in the cytosol, proteins are directed to specific cell compartments called organelles to perform their proper biological functions. It is well known fact that two main phenomena are responsible for cell dysfunction or damage leading to serious diseases, protein misfolding and erroneous presence of proteins in cell compartments. Hence, to at least avoid protein subcellular mislocalization, accurate trafficking of proteins to their ultimate destinations is crucial [1]. Although the intense research efforts made to understand such mechanisms and their influence on the cell functional machinery, further investigations are still expected for full insight on the proteins behavior *in-vivo*. Such efforts depend on the development of new in silico methodologies as alternative to the costly and ardious wet-lab experiments which are sometimes impractical due to the nature of certain proteins. Today, with the rapid development of structural bioinformatics and sequential bioinformatics, computational methods have become essential in genomic and proteomic analyses. However, the major machine learning-based methods are facing two challenges, consisting in: (i) the presence of multi-location proteins which may be located in more than one organelle simultaneously and assigning proteins to a single location is a drastic simplification of reality [2], (ii) the imbalanced nature of learning datasets of annotated protein sequences as most learner systems exhibit bias towards the majority class while the minority class is generally of greatest interest. In order to make up the shortfalls and achieve desirable results, computational methods development for SCL prediction focuses on powerful individual learning models, ensemble models to take advantage of their combined strengths, and heterogeneous data integration. The pioneering methods predicted the SCL solely from the amino acid sequence such as the rule-based expert system PSORT-I developed by Nakai and Kanehisa [3, 4] and the probabilistic model proposed by Horton and Nakai [5]. Thereafter, different classification algorithms have been used to further improve the performance. Among these algorithms, k-Nearest Neighbor (k-NN) [6, 7], binary Decision Tree (DT) [8], Naive Bayesian (NB) classifier [9–11], Artificial Neural Networks (ANNs) [12–15], Support Vector Machines (SVMs) [16–23], Hidden Markov Models (HMMs) [24], and Bayesian networks [9, 25]. Since these works, many systems using a variety of machine learning techniques have been proposed achieving varying degrees of success. They were specialized for specific organisms and certain localization sites, but no significant improvements over the k-NN algorithm were reported until the burt of the new generation methods based on hybrid models and fusion approach [26–33] taking into account both protein sequence and structure charachteristics. They are categorized as sorting signals-based, composition-based and homology-based methods. The first category includes MitoProt [33], PSORT-II [6], ChloroP [34], TargetP [13], iPSORT [35], PSORT-B [25], and NucPred [36]. The second category includes Sub-Loc [16], Esub8 [24], ESLpred [37], pSLIP [38], AAIndexLoc [39], ngLoc [112], YLoc [11], and BaCelLo [41]. The third category includes phylogenetic profiling based methods [40, 42] and sequence homology-based methods such as GOASVM [23] and SCLpredT [43]. Recently, protein–protein interaction [44–46], gene expression levels [23, 26, 47–49] and textual information [40] has been also integrated to infer the subcellular locations exploiting the available databases of proteins with known localization [50]. SCL prediction based on deep learning methods has also emerged due to their ability to learn high-level features [30, 51–55]. The success achieved by using both composition and homology information [56–61] has led to a plethora of methods using different strategies to improve the prediction quality. However, the best performing SCL systems reported to date are mainly based on multi-label learning and Gene Ontology (GO) annotation which has revolutionized the way to represent biological knowledge so as to be computationally accessible [48, 62, 63]. Basically, GO concept describes the roles of genes across different organisms and allows functional inference for newly discovered genes. The established collection of terms adopted and standardized by the GO Consortium[1] [64] are derived from experimental and electronic annotations. They are organized in three distinct sub-ontologies that represent gene/protein functions aspects: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Molecular function corresponds to activities that can be performed at the molecular level, such as catalytic, binding, or transporter activities. A biological process corresponds to pathways and programs involved in. A cellular component is either the cellular environment or extracellular

1 https://geneontology.org/.

region where the activity is executed. Each category of GO terms is organized as a directed acyclic graph (DAG) with terms as nodes and relationships as edges. There is a structured hierarchy with defined semantic relationships between terms, where each term can have relationships to several parent and child terms. High level terms are more general and low level terms are more specific than their respective parent terms. The most ubiquitous relationships are: "is-a" which describes the fact that child term is an instance of parent and "part-of" which shows that child term is a component of parent. Each GO term has a unique alphanumeric identifier where functional assignment source is indicated in the form of Evidence Code[2] which might be experimental, computational or automatic-assignment evidence. In this paper we investigate the effect of GO annotation and profile alignment on SCL prediction of Gram-negative and Gram-positive bacterial proteins using multi-label learning. These microscopic unicellular prokaryotes distinguished by the lack of cell nucleus play a critical role in health problems. Despite their beneficial effect, they are mostly pathogenic and source of many diseases in humans. There are only a few methods that concentrate on this specie by tackling the SCL prediction problem using both multi-label learning and GO terms [32, 65–71], so, here we tried to estimate how much improvement over the traditional pseudo amino acid composition (PseAAC) [72] might be provided by using position-specific scoring matrix (PSSM) profiles, GO terms and both. Generally, GO terms are retrieved by querying protein accession number against GOA[3] database for annotated proteins or using either InterProScan[4] [73] to scan query proteins for significant matches against the InterPro[5] protein signature databases or BLAST [74] to obtain accession numbers of homologous proteins as the searching keys for uncharacterized proteins. Here, BLAST is used as baseline for similarity search to transfer the GO terms of homologous proteins to target proteins and to infer PSSM profiles. Our proposed model can predict five and six distinct locations on Gram-positive and Gram-negative bacterial proteins, respectively. In order to achieve such a goal, we tried to follow the well-established process of Chou's 5-steps rule [75] for developing fast and reliable computational methods for genomic or proteomic analysis and drug development [76–80]. As it is explicitly described in review papers [81–83] and Wikipedia, our major efforts were thus devoted to: (i) collecting benchmark datasets from experimentally validated protein sequences to train and test the method; (ii) using an effective mapping of protein samples from sequential to vectorized representation; (iii) developing a powerful SCL prediction model for Gram-positive and Gram-negative bacterial proteins, exploiting diversity in both feature and decision spaces; (iv) performing cross-validation tests with appropriate measures to effectively evaluate the prediction model performance; (v) implementing a user-friendly web-server for the proposed prediction model to make it publicly available since such tools allow to avoid going through the complicated scientific and mathematical formulas, and represent the future direction for developing practically more useful predictors. Below, we describe how we performed these steps to obtain our final prediction model.

The rest of the paper is organized as follows. First we present the proposed framework. Next, we briefly describe the benchmark datasets and the evaluation methodologies adopted and then, we summarize the experiments and the results obtained. Finally, in the last section we conclude and present some future research plans.

2 Method

Traditional supervised learning algorithms learn from examples associated with only one single label either for binary or multi-class classification. However, many real-world problems deal with data that does not fail in this category and are referred to as multi-label learning problems, due to the nature of their training examples

2 https://www.geneontology.org/doc/GO.Evidence.html.
3 https://www.ebi.ac.uk/GOA/index.
4 https://www.ebi.ac.uk/interpro/search/sequence/.
5 https://www.ebi.ac.uk/interpro.

which are associated with multiple labels simultaneously. To deal with such situation, two approaches are adopted namely, algorithm adaptation and problem transformation. The first approach modifies or extends the existing algorithms to obtain dedicated versions, taking into account the multi-label nature of the samples. Whereas, the second and the most used approach transforms the original data so as to be able to apply the traditional algorithms. Two strategies are applied for the latest, Binary Relevance (BR) and Label Powerset (LP) transformations. The first model consisting in class binarization principle applies the traditional one-vs-all approach to transform the original multi-label problem to several bi-class sub-problems to apply binary classifiers and combines the predictions. The second model transforms the multi-label problem into a multi-class problem by giving the set of labels associated to each instance, a class identifier to apply any multi-class classifier. BR is relatively a naive approach since each label is learned independently, assuming label independence. However, in multi-label learning from imbalanced data, which is inherent to SCL prediction, it is

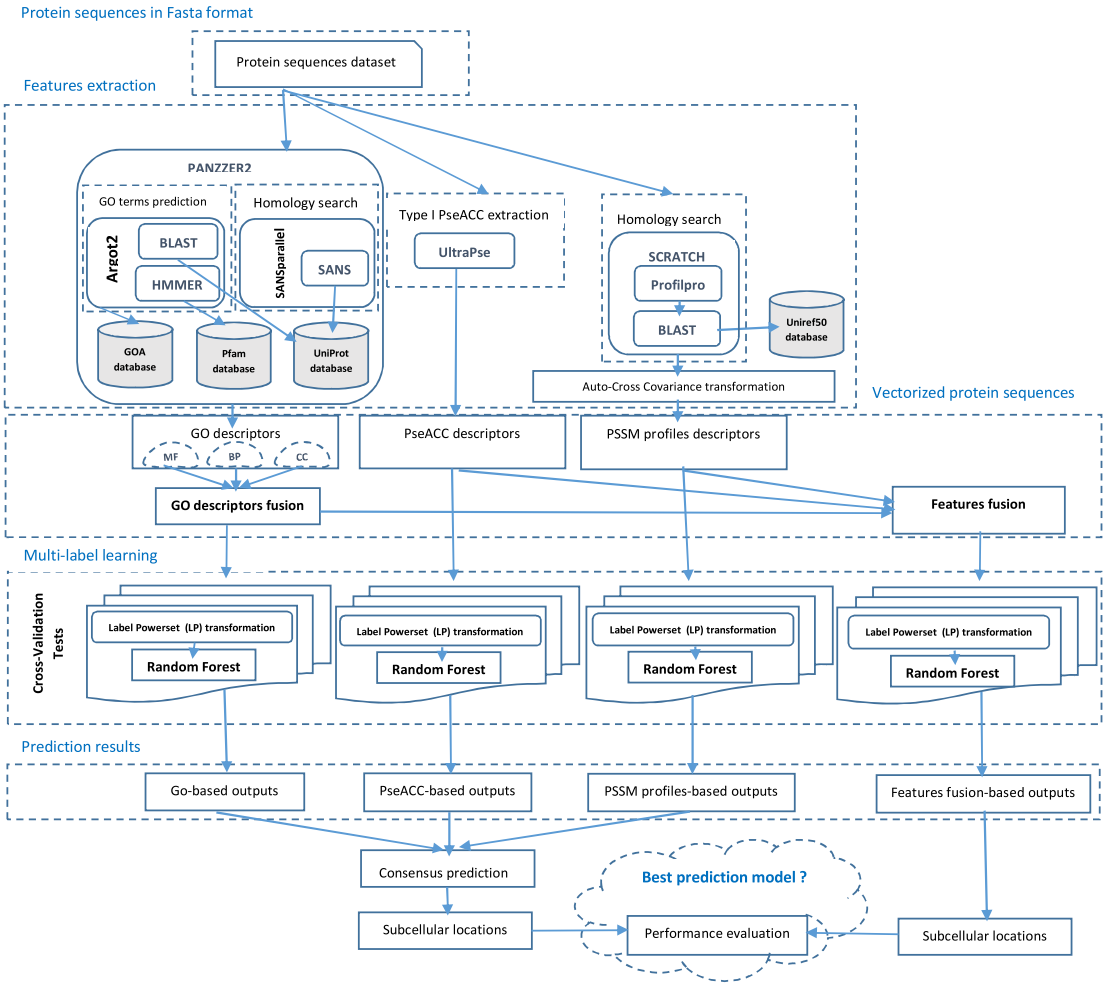

Figure 1: Flowchart for the proposed prediction model for Gram-positive and Gram-negative bacterial proteins subcellular localization. Firstly, protein sequences datasets were collected from the published database. Secondly, they were filtered out and preprocessed using different strategies to obtain a fixed size feature vector representation that can be fed into the learning model. Thirdly, the resulting encoded feature vectors were independently put into the multi-label learning model-based on Label Powerset (LP) transformation to produce independent prediction scores using Random Forest (RF) ensemble method as base classifier. Once optimum performance scores were calculated by using 5-fold cross-validation tests, the final prediction model is built.

important to exploit label relations or dependency to improve the performance. Although, many studies focused on label co-occurrence and correlation to improve the prediction quality, no approach works consistently better on all kinds of multi-label datasets. In this study, we tried to capture the correlation information among labels by using Label Powerset (LP) strategy and the ensemble method Random Forest [84] as baseline classifier. The flowchart in Figure 1 describes the main framework of our proposed method for predicting Gram-positive and Gram-negative bacterial proteins SCL.

2.1 Fusion strategy

In traditional supervised learning, a sample is represented by an instance or feature vector and its associated single class label. Let us denote by $X \subseteq R^d$ a d-dimensional feature space and y a finite set of Q class labels $\{y_1, y_2, \cdots, y_Q\}$. The goal is to learn a function $f : x \to y$ from a set of instances $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where $x_i \in x$ is an instance and $y_i \in y$ is the known label of x_i. In multi-label learning, the number of all label vectors is generally 2^Q. Each label $y_i \in y \subseteq \{1, 0\}^Q$, where $y_i[j] = 1$ if and only if the instance x_i is associated with the jth label, $1 \le j \le Q$. Given the multi-label training set D, the task of multi-label learning consists to learn a predictor h from D, $h : x \to 2^y$ mapping from the instance space to the label space. For each instance (x, y) in test, $\tilde{y} = h(x)$ is the predicted label and $c(y, \tilde{y})$ the cost or penalty of predicting y as $\tilde{y}$, such value might be quantified using the evaluation metrics given in Section 3.2. In Label Powerset the transformation process only modifies the label space, considering the set of distinct unique subsets of labels present in the original training set [85]. In the novel label space the multi-class label $Y = y_{i,\, j,\, \ldots,\, Q}$ means that the respective instance is labeled with the conjunction $y_i \wedge y_j \wedge \ldots \wedge y_Q$. The first prediction model is based on PSSM profiles and the second on GO terms. As the two best individual prediction models will probably provide a reduced set of multi-label outputs, our goal is to improve multiple locations assignation to query proteins. The outputs of the consensus prediction model consist on the union of the individual models predictions. Suppose $\tilde{Y}^{\mathrm{PSSM}} = \tilde{y}^{\mathrm{PSSM}}_{i,\, j,\, \ldots,\, Q}$ be the predicted multi-class label obtained using PSSM profiles and $\tilde{Y}^{\mathrm{GO}} = \tilde{y}^{\mathrm{GO}}_{i,\, j,\, \ldots,\, Q}$ using the predicted GO terms, respectively. The multi-class label obtained by the consensus prediction is as follows:

Table 1: Prokaryotic benchmark datasets statistics. Code column indicates the subcellular location representation in our predictive model. Gram-negative bacteria have five major subcellular localization sites, namely, the cytoplasm, the periplasm, the inner membrane, the outer membrane, and the extracellular space, whereas Gram-positive bacteria do not have an outer cell membrane. However in these benchmark datasets cell wall is absent in Gram-negative dataset and in Gram-positive bacteria, we observe the lack of periplasm proteins.

No	Subcellular location	Code	Proteins count	
			Gram negative	Gram positive
1	Cytoplasm	C	4,152	349
2	Extracellular	S	272	290
3	Inner membrane	I	1,415	1,779
4	Outer membrane	O	346	–
5	Periplasm	P	422	–
6	Cell wall	W	–	34
7	Vacuole	V	10	4
Multiple localizations			39	8
Total			6,578	2,448

$$\overset{\sim \text{Consensus}}{Y} = \left(\overset{\sim \text{PSSM}}{y} \vee \overset{\sim \text{GO}}{y} \right)_{i,\, j,\, \ldots,\, Q} \tag{1}$$

The consensus prediction for each query protein is obtained by the conjunction (bitwise OR) of its respective predicted single class labels. For example, in the case of Gram-negative bacterial proteins, where the number of predicted subcellular locations is 6 (see Table 1), the prediction model decision is obtained as follows:

$$\begin{aligned}
\overset{\sim \text{PSSM}}{Y} &= [0,\, 0,\, 0,\, 0,\, 1,\, 0]\ \textit{which represents}\ S\ (\textit{Extracellular})\ \ \textit{location} \\
\overset{\sim \text{GO}}{Y} &= [1,\, 0,\, 0,\, 0,\, 0,\, 0]\ \textit{which represents}\ I\ (\textit{Inner Membrane}) \\
\overset{\sim \text{Consensus}}{Y} &= [1,\, 0,\, 0,\, 0,\, 1,\, 0]\ \textit{which is represented by}\ S \Big/ I\ (\textit{Extracellular and Inner Membrane})
\end{aligned}$$

The query protein is thus predicted in both extracellular region and inner membrane.

2.2 Protein sequence representation

There is no doubt that protein sequence representation influences protein subcellular location prediction performance. SCL prediction methods have tried different protein sequence and structure properties to improve the prediction quality but it appears that incorporating GO information is decisive for SCL prediction quality. In the following subsections, we describe protein sequences representation step which has led to several versions of the benchmark datasets.

2.2.1 Pseudo amino acid composition

In computational biology, an important but challenging step is how to represent a biological sequence by a discrete model or a vector that captures its key features without losing sequence-order information, since most of the existing machine-learning algorithms can only handle fixed-length numerical vectors [86, 87]. When dealing with protein sequences, the simplest way to characterize a protein by a fixed-length numerical vector is to extract the information of the protein sequence from the entire amino acid sequence, especially when the protein does not have significant homology to annotated proteins. K.C. Chou proposed PseAAC (pseudo amino acid composition) which has been extensively used in protein-related prediction systems as it has been introduced to enhance the power of the conventional discrete amino acid composition (AAC) which consists of 20 components representing the occurrence frequencies of the 20 naturally occurring amino acids in the sequence. PseAAC incorporates both the sequence order and the length effect [72]. Due to the great success of such representation model for protein sequences in different areas of computational biology, the concept of PseAAC has been extended to represent nucleotide sequences with the concept PseDNC (pseudo-dinucleotide compositions) and the concept of PseKNC (Pseudo K-tuple Nucleotide Composition) [88]. Many Web-servers and stand-alone programs are now available to generate such features and any other desired features for protein/peptide sequences such as PseAAC-Builder[6] [89], propy[7] [90], PseAAC-General[8] [91], Pse-in-One[9] [92] and its very powerful updated version Pse-in-One 2.0[10] [93], and UltraPse[11] software [94] which allows all possible sequence representation modes for user-defined sequence types. Here, UltraPse source code has been downloaded and run locally on our ubuntu platform to extract pseudo amino acid composition (Type I General PseAAC) for each protein sequence in the benchmark datasets.

6 https://pseb.sourceforge.net/.
7 https://code.google.com/p/protpy/downloads/list.
8 https://pseb.sourceforge.net/.
9 https://bioinformatics.hitsz.edu.cn/Pse-in-One/.
10 https://bioinformatics.hitsz.edu.cn/Pse-in-One2.0/.
11 https://github.com/pufengdu/UltraPse.

2.2.2 Generation of PSSM profiles

Protein sequences have been represented by position-specific scoring matrix (PSSM) profiles which reflect the frequencies of each amino acid residue in a specific position of a multiple alignment. To obtain such mapping, we used PROFILpro release 1.1 integrated in SCRATCH[12] server [95] which performs with BLAST version 2.2.26 [74] and a non-redundant UniRef50[13] database (clustered sets of protein sequences that show 50% sequence identity) as the search database. Protein PSSM features were computed by setting the number of iterations to three (–*j*3) and the inclusion e-value to 0.001 (–h0.001). Each feature vector is a $20 \times L$ dimension, where L is the length of the protein sequence. To map the obtained feature vectors of varying lengths to fixed length vectors while preserving the local sequence-order information, auto-cross covariance tranformation (ACC) has been applied [96] which is provided by Pse-in-One 2.0 and protr[14] R package [97]. Each protein sequence is thus represented by a numeric vector of length $lg * 20^2$, where lg is the distance between one amino acid residue and its neighbor along the protein sequence. To describe the ACC transformation, let us denote by $p_{i,j}$ the probability (score) of amino acid i occurring at the position j in the PSSM, if we consider each amino acid as one property and the PSSM as the time sequences of all properties. ACC transformation converts the PSSM of different lengths into a fixed-length vector by measuring the correlation between each pair of properties. It first builds two signal sequences, and then calculates the correlation between them. Let us denote by $\bar{p}_i$ the average score for amino acid i along the whole sequence, expressed by:

$$\bar{p}_i = \frac{1}{L}\sum_{j=1}^{L} p_{i,j} \tag{2}$$

ACC results in two kinds of variables: auto covariance (AC) between the same property, and cross covariance (CC) between two different properties. The AC variable measures the correlation of the same property between two residues separated by a distance of lag along the sequence and can be calculated as follows:

$$AC(i, lag) = \frac{1}{L - lag}\sum_{j=1}^{L-lag}\left(p_{i,j} - \bar{p}_i\right)\left(p_{i,j+lag} - \bar{p}_i\right) \tag{3}$$

where i is one residue of the protein sequence of length L.

In this way, the number of AC variables is $20 \times lg$, where 20 corresponds to the number of columns of the PSSM and lg is the maximum value of lag $(lag = 1, 2, \cdots, lg)$.

The CC variable measures the correlation of two different properties between two residues separated by lag along the sequence as follows:

$$CC(i, j, lag) = \frac{1}{L - lag}\sum_{k=1}^{L-lag}\left(p_{i,k} - \bar{p}_i\right)\left(p_{j,k+lg} - \bar{p}_j\right) \tag{4}$$

where i and j are two different amino acids and $\bar{p}_i$ ($\bar{p}_j$) is the average score for amino acid i (j) along the sequence.

Each protein sequence is thus represented as a vector of ACC-derived variables as combination of AC and CC variables. Here, the parameter value lg is set to one to reduce the computational time for a larger dataset, which is considered as a default parameter of our predictive model. However, it is worth noticing that taking into account the amino acids neighboring effect may be able to improve the prediction quality, so further investigations are required to evaluate the contribution of lg value to the performance of the proposed prediction model.

12 https://scratch.proteomics.ics.uci.edu/.

13 ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50.

14 https://cran.r-project.org/web/packages/protr/index.html.

2.2.3 Gene ontology terms prediction

GO terms prediction is also an active area of research for in silico functional annotation. Uncharacterized proteins do not have accession numbers since they are not repertorized in databases yet and hence direct GO terms retrieval from GOA database is not possible. Such is the case, of newly-discovered, synthetic and hypothetical proteins [98]. Computational methods try to annotate these proteins by transferring GO terms from annotated proteins by sequence or structural similarity [99]. Furthermore, it is assumed that proteins sharing common primary and secondary structures are generally located in the same cellular regions [100–103]. Here, the prediction system takes as input a list of sequences in FASTA format and predicts GO terms using PANZZER2[15] (Protein ANNotation with Z-score) [104, 105] which uses SANSparallel[16] [106] based on SANS (suffix array neighborhood search) algorithm [107] to perform high-performance and fast homology searches in the UniProt[17] database instead of BLAST. The benchmark datasets have been treated in batch mode to obtain both GO annotations and free text protein descriptions (DE) files. Generally, multiple GO terms can be assigned to each query protein which can induce a strong statistical redundancy, while only a relatively small number of unique GO terms is essential for protein annotation [108]. To prevent potential redundancy and select correct annotations PANZZER2 includes implementations of the scoring functions from PANZZER (its basic version), Blast2GO[18] [109, 110] and Argot2[19] (Annotation Retrieval of GO Terms) [63]. The latest exploits a combined approach based on the clustering process of GO terms dependent on their semantic similarities and a weighting scheme which assesses retrieved hits sharing a certain degree of biological features with the sequence to annotate. Where, hits may be obtained by BLAST with UniProt as reference database or HMMER[20] with Pfam[21] using a recent release of UniProtKB-GOA database [111]. Both Blast2GO and Argot are suitable for non-model species, however, our analysis in this study was based on Argot predictions since it has been revisited to increase both accuracy and precision by using an improved weighting scheme, Pfam models and new releases of reference databases. For each query protein sequence a list of scored and ranked GO terms is provided. The GO term set consists of three subsets: molecular function (MF), biological function (BP), and cellular component (CC). Once the relevant GO subspace is obtained, each subset is processed in order to obtain numerical feature vectors of probability estimates using PPV (Positive Predictive Value) which is the normalized prediction score between 0 and 1. We adopted such strategy for the sake of comparison with the common practice that consists to construct feature vectors by using 0/1 value to represent the presence and absence of the predefined GO terms or the frequency of occurrences of GO terms [23].

3 Experimental design

In this section, we present the experimental design used to evaluate our SCL prediction model. We first describe the benchmark datasets collected from the literature. The coverage of our proposed model is directly related to the available annotation terms in the datasets used for training, taken as class labels. Each class label is a binary encoded vector of length equal to the number of distinct locations. For protein sequences annotated by two or more locations, multiple locations are encoded by summing up (bitwise OR) each corresponding binary vector. Some evaluation measures typically applied in multi-label learning based prediction are summarized in this section.

15 https://ekhidna2.biocenter.helsinki.fi/sanspanz/.
16 https://ekhidna2.biocenter.helsinki.fi/cgi-bin/sans/sans.cgi.
17 https://www.uniprot.org/.
18 https://www.blast2go.com/
19 https://www.medcomp.medicina.unipd.it/Argot2-5/.
20 https://hmmer.janelia.org.
21 https://pfam.xfam.org/.

3.1 Benchmark datasets

In the protein SCL dataset, a protein might be associated with a set of SCL labels related to the cell type. Prokaryotic cell is typically composed of a cell wall which protects the cell and gives shape, a cell membrane which separates the intracellular environment from the extracellular space which is outside the plasma membrane, and the most abundant cytoplasm where the major cellular processes are performed. Some prokaryotic cells produce gas vacuoles named gas vesicles. One distinguishes Archaea and Bacteria cells, the latest are divided in two broad categories according to their cell wall, Gram-positive, and Gram-negative. Gram-negative bacteria have five major SCL sites, which are the cytoplasm, the periplasm, the inner membrane and the outer membrane. The inner membrane separates the cytoplasm from the periplasm. The outer membrane protects the cell against some antibiotics, and the extracellular space. The outer membrane is absent in Gram-positive bacteria which allows antibiotics reception. The volume of periplasm is much smaller than in Gram-negative bacteria and it is characterized by a thicker cell wall. Our SCL prediction model has been benchmarked on two independent datasets of experimentally determined annotations on SCL. They were collected from curated set of bacterial protein sequences Gram-positive and Gram-negative, taken from the Swiss-Prot database release of May 17th 2011 [112], available here. These datasets contain protein sequences having less than 98% sequence identity to reduce sequence redundancy. Here, protein datasets have been filtered out so as to only consider protein sequences with the 20 standard amino acids and excluded sequences containing *X* symbol because of their ambiguity. The number of proteins in each main localization obtained after the filtering process are summarized in Table 1 for Gram negative and Gram positive bacteria datasets. The class referred to as Vacuole contains the gas Vesicle proteins incorporated in both datasets, such class or proteins appears only in certain prokaryotic organisms. Gram-positive and Gram-negative bacteria are chiefly differentiated by their cell wall structure, Table 1 lists the number of proteins in different localization sites in the datasets and Figure 2 and Figure 3 report the datasets statistics. The two benchmark datasets are imbalanced since the distributions of the proteins in different locations is uneven. The majority of Gram-negative bacterial proteins are located in the cytoplasm, the inner membrane and the periplasm, whereas the Gram-positive bacterial proteins are located in the cell inner membrane, the cytoplasm and the extracellular space. As we can see, the number of protein samples in Vacuole (V) which represent the gas Vesicle location is 10 in Gram-negative dataset and only 4 in Gram-positive dataset, which represents the minority class. It is worth

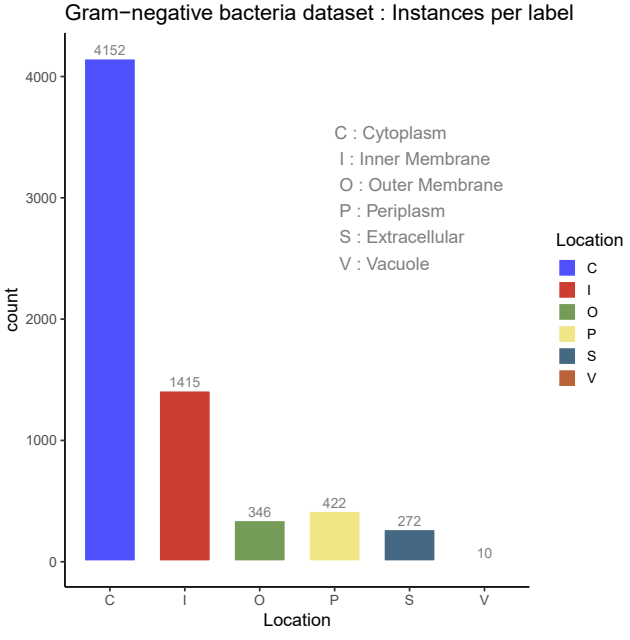

Figure 2: Gram-negative bacteria dataset.

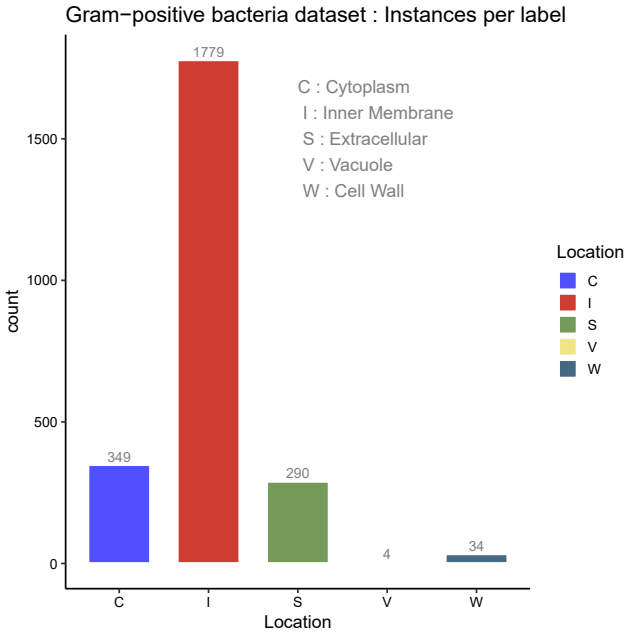

Figure 3: Gram-positive bacteria dataset.

noticing that such class is rare in bacterial proteins, it is present in aquatic and marine bacteria, while it represents an important class in archaea and planktonic species. The same trend is observed for the cell wall (W) compartment. Cytoplasm is the most abundant class in Gram-negative dataset whereas it is significantly dominated by Inner membrane (I) class in Gram-positive dataset. In these benchmark datasets cell wall is absent in Gram-negative dataset and in Gram-positive bacteria, we observe the lack of periplasm proteins which indicates that our predictive model will not predict the cell wall location for Gram-negative bacteria and periplasm location for Gram-positive bacteria.

Multiple sites proteins are shown in Figure 4 for Gram-negative bacteria and Figure 5 for Gram-positive bacteria. As it can be observed, multiple locations proteins are limited to a pair sites which is generally the case of the majority of multiple sites proteins [2].

3.2 Performance measures

The performance evaluation of multi-label learning based prediction models needs specific metrics, since each instance could be associated with two or more labels simultaneously. Various meaningful metrics have been used in the literature such as example-based metrics and label-based metrics [113, 114]. The first category evaluates the generalization performance on each test instance and returns the average value for the entire test set, whereas the second category proceeds first on each class label separately, and then the average value is calculated across all class labels. The latest measures are derived from the four common values used in binary classification, namely TP (true positive), FP (false positive), TN (true negative), and FN (false negative). The term macro is used for a measure such as recall, precision, F1 score derived by assuming equal importance for each label while micro corresponds to that derived by assuming equal importance for each example [115]. In our study, we adopted both example-based and label-based metrics. They are implemented in both mldr[22] [116] and utiml[23] [117] R packages. Given a test instance x_i, $i = 1, \ldots, N$, y the set of all labels, $Y_i \subseteq y$ the set of true labels and $\hat{Y}_i$ the set of predicted labels for x_i, the metrics are thus described in the following subsections.

22 https://cran.r-project.org/web/packages/mldr/index.html.
23 https://cran.r-project.org/web/packages/utiml/index.html.

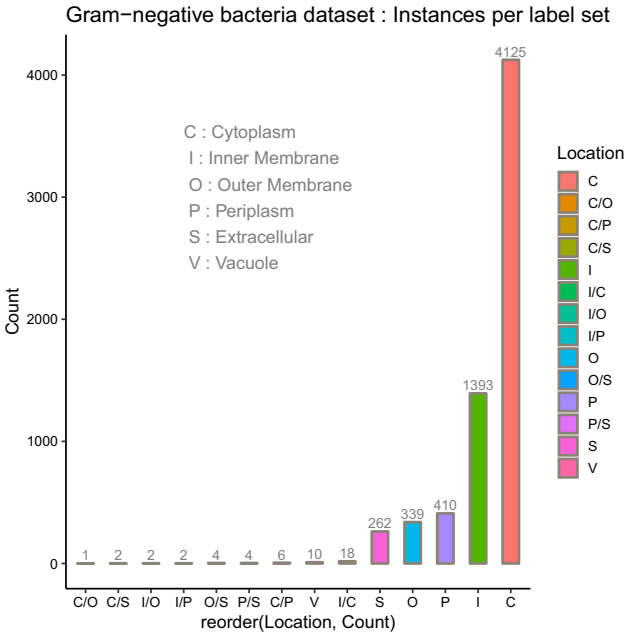

Figure 4: Observed localization sites of proteins in Gram-negative bacteria dataset.

3.2.1 Example-based measures

Accuracy score computes the percentage of correctly predicted class labels among all predicted and true class labels for each instance, it is defined as follows:

$$Accuracy\,score = \frac{1}{N}\sum_{i=1}^{N}\frac{\|Y_i \cap \tilde{Y}_i\|_1}{\|Y_i \cup \tilde{Y}\|_1} \tag{5}$$

It is important to note that using Accuracy metric alone may mislead the analysis when dealing with imbalanced data and high scores do not necessarily indicate good performance.

Precision is the proportion of TP examples from all the examples predicted as positive.

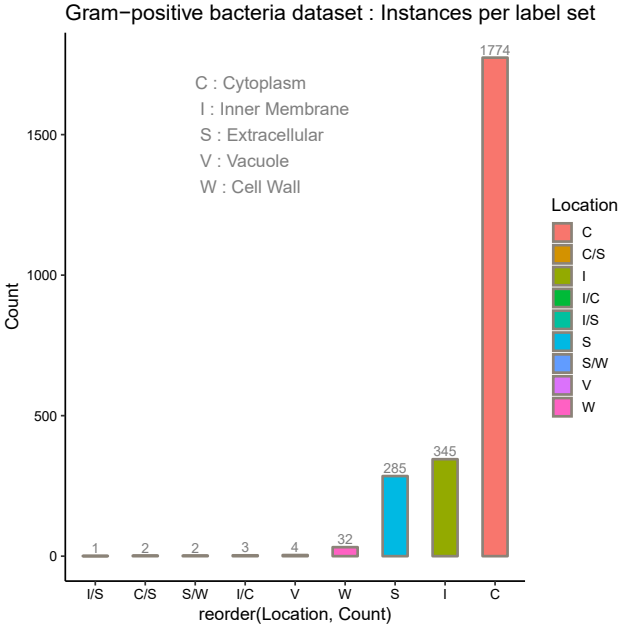

Figure 5: Observed localization sites of proteins in Gram-positive bacteria dataset.

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\left|Y_i \cap \tilde{Y}_i\right|\right|_1}{\left|\left|\tilde{Y}_i\right|\right|_1} \tag{6}$$

Recall is the proportion of TP examples predicted as positive.

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\left|Y_i \cap \tilde{Y}_i\right|\right|_1}{\|Y_i\|_1} \tag{7}$$

F1 score is the harmonic mean between Precision and Recall, expressed as follows:

$$F1\,score = \frac{1}{N}\sum_{i=1}^{N}\frac{\|2Y_i \cap \tilde{Y}_i\|_1}{\|Y_i\|_1 + \|\tilde{Y}_i\|_1} \tag{8}$$

Subset-accuracy

$$Subset\_accuracy = \frac{1}{N}\sum_{i=1}^{N} I\left(Y_i = \tilde{Y}_i\right) \tag{9}$$

where I function is defined as I(true) = 1 and I(false) = 0. This metric takes into account only exact matches and by ignoring partially correct matches, it will not be able to recognize nearly exact prediction from totally incorrect prediction. While, it would be interesting to know if an example is correctly assigned to at least one of the labels it belongs to, especially when dealing with imbalanced data.

Hamming-Loss gives the fraction of labels that are incorrectly predicted. It is the widespread evaluation metric in multi-label learning systems, expressed as:

$$Hamming - loss = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}\sum_{j=1}^{Q}\|Y_i[j] \neq \tilde{Y}_i[j]\|_1 \tag{10}$$

Rank-loss evaluates the average proportion of label pairs that are incorrectly ordered for an example. It is defined as follows :

$$Rank - loss = \frac{1}{N}\sum_{i=1}^{N}\sum_{Y_i[j]>Y_i[l]}\left(\left[\left[\tilde{Y}_i[j] < \tilde{Y}_i[l]\right]\right] + \frac{1}{2}\left[\left[Y_i[j] = \tilde{Y}_i[l]\right]\right]\right) \tag{11}$$

The higher the value of Accuracy and F1 score, the better the performance of the learning algorithm and the smaller the value of Hamming loss and Rank loss, the better the performance. The Area Under the ROC Curve (AUC) is also used since it is a good indicator of performance, especially when dealing with multi-label learning problems.

Table 2: Extracted GO terms statistics for the three components of GO namespace, namely, Molecular Function (MF), Biological Process (BP), and Celullar Component (CC).

Dataset	GO terms count		
	MF	BP	CC
Gram negative	1,140	1,316	223
Gram positive	570	667	116

3.2.2 Label-based measures

As it has been mentioned above, these metrics are obtained for all labels by either maco-averaging or micro-averaging.

Macro_precision is defined by the fraction of the number of TPs by the number of both TPs and false positives for the label y_j considered as a binary class.

$$Macro\_precision = \frac{1}{Q}\sum_{j=1}^{Q}\frac{TP_j}{TP_j + FP_j} \tag{12}$$

Recall is defined by the fraction of the number of TPs by the number of both TPs and false negatives for the label y_j.

$$Macro\_recall = \frac{1}{Q}\sum_{j=1}^{Q}\frac{TP_j}{TP_j + FN_j} \tag{13}$$

Macro_F1 score is the harmonic mean between Macro_precision and Macro_recall, expressed as follows:

$$Macro\_F1score = \frac{2 * Macro\_precision * Macro\_recall}{Macro\_precision + Macro\_recall} \tag{14}$$

Micro_precision

$$Micro\_precision = \frac{\sum_{j=1}^{Q} TP_j}{\sum_{j=1}^{Q}(TP_j + FP_j)} \tag{15}$$

Micro_recall

$$Micro\_recall = \frac{\sum_{j=1}^{Q} TP_j}{\sum_{j=1}^{Q}(TP_j + FN_j)} \tag{16}$$

Micro_F1 score is the harmonic mean between Micro_precision and Micro_recall, expressed as follows:

$$Micro\_F1score = \frac{2 * Micro\_precision * Micro\_recall}{Micro\_precision + Micro\_recall} \tag{17}$$

3.3 Results and discussion

This section presents the results of the study and a synthesis of the experiments carried out, starting by the homology-based GO extraction. The studied SCL prediction models have been assessed using cross-validation tests to infer the best ensemble model with a special interest to multiple sites proteins as it is illustrated in Figure 1.

3.3.1 Extracted GO terms

First we extracted from the learning datasets three sets of distinct GO terms which are the top ranked GO terms provided by PANNZER2, corresponding to the three sub-ontology molecular function (MF), biological process (BP), and cellular component (CC) by removing the repetitive GO terms. Then the feature vector has been constructed for each protein given in FASTA format from the union of these essential GO terms for two reasons: (i) to increase the possibility of getting at least one GO term for protein encoding to reduce the scenario where annotation is totally absent, (ii) to enhance the prediction quality since it has been found that not only CC GO terms are indicative of cellular component but also both MF and BP GO terms contribute to the final predictions. Many studies characterized a protein by a feature vector of 0/1 values indicating whether the protein is annotated with a predefined GO term or not, or the frequency of such GO term in [118]. Here, we investigated

Table 3: Performance evaluation results of cross-validation tests on Gram-negative bacteria dataset for different predictions: pseudo-amino acid composition (PseAAC), position-specific scoring matrix (PSSM) profiles, gene ontology (GO) terms 0/1-based representation, GO terms PPV-based representation, features fusion, a consensus prediction using both GO terms and PSSM profiles outputs, and a consensus of PseAAC, GO terms and PSSM profiles outputs. Italic values correspond to the best predictive model.

Sequence features	Example-based metrics							Label-based metrics					
	Accuracy	Precision	Recall	F1 score	Subset_ accuracy	Hamming-loss	Rank-loss	Macro_ precision	Macro_ recall	Macro_F1 score	Micro_ precision	Micro_recall	Micro_F1 score
PseACC	0.882	0.884	0.882	0.883	0.880	0.039	0.117	0.873	0.656	0.732	0.884	0.879	0.881
PSSM profiles	0.940	0.941	0.940	0.940	0.937	0.020	0.060	0.919	0.815	0.860	0.941	0.936	0.939
GO terms 0/1	0.963	0.965	0.963	0.963	0.960	0.012	0.037	0.952	0.865	0.903	0.965	0.960	0.962
GO terms ppv	0.965	0.967	0.965	0.965	0.962	0.011	0.035	0.958	0.869	0.908	0.967	0.962	0.964
PseAAC+ PSSM+GO	0.951	0.953	0.951	0.952	0.949	0.016	0.048	0.938	0.835	0.880	0.953	0.948	0.951
PSSM+GO	0.951	0.954	0.951	0.952	0.949	0.016	0.048	0.939	0.850	0.890	0.954	0.948	0.951
$Consensus_{PseAAC\ +PSSM+GO}$	0.920	0.920	0.986	0.941	0.858	0.030	0.043	0.852	0.930	0.884	0.855	0.984	0.915
$Consensus_{PSSM+GO}$	*0.953*	*0.954*	*0.982*	*0.963*	*0.922*	*0.016*	*0.029*	*0.907*	*0.922*	*0.911*	*0.923*	*0.980*	*0.951*

Table 4: Performance evaluation results of cross-validation tests on Gram-positive bacteria dataset for different predictions: pseudo-amino acid composition (PseAAC), position-specific scoring matrix (PSSM) profiles, gene ontology (GO) terms 0/1-based representation, GO terms PPV-based representation, features fusion, a consensus prediction using both GO terms and PSSM profiles outputs, and a consensus of PseAAC, GO terms and PSSM profiles outputs. Italic values correspond to the best predictive model.

Sequence features	Example-based metrics							Label-based metrics					
	Accuracy	Precision	Recall	F1 score	Subset_ accuracy	Hamming-loss	Rank-loss	Macro_ precision	Macro_recall	Macro_F1 score	Micro_ precision	Micro_recall	Micro_F1 score
PseACC	0.896	0.897	0.895	0.896	0.894	0.041	0.104	0.708	0.510	0.559	0.897	0.894	0.895
PSSM profiles	0.937	0.938	0.937	0.937	0.935	0.025	0.062	0.909	0.763	0.812	0.938	0.935	0.937
GO terms 0/1	0.959	0.959	0.958	0.959	0.957	0.016	0.041	0.952	0.785	0.791	0.959	0.957	0.958
GO terms ppv	0.962	0.962	0.961	0.962	0.960	0.015	0.038	0.927	0.803	0.819	0.962	0.960	0.961
PseAAC+PSSM+GO	0.947	0.949	0.947	0.948	0.946	0.020	0.052	0.931	0.772	0.824	0.949	0.946	0.947
PSSM+GO	0.948	0.950	0.948	0.949	0.946	0.020	0.051	0.920	0.822	0.851	0.950	0.947	0.948
$\text{Consensus}_{\text{PseAAC+PSSM+GO}}$	0.924	0.924	0.978	0.941	0.871	0.033	0.050	0.855	0.858	0.839	0.871	0.977	0.921
$\text{Consensus}_{\text{PSSM+GO}}$	*0.950*	*0.950*	*0.974*	*0.958*	*0.924*	*0.021*	*0.038*	*0.886*	*0.854*	*0.851*	*0.924*	*0.973*	*0.948*

Table 5: Performance evaluation by 5-fold cross-validation tests on Gram-negative bacteria proteins using a consensus of PSSM and GO terms-based predictions. The multi-label confusion matrix reflects well the predictions performance for each location separetely.

Subcellular locations		Metrics										
		TP	FP	FN	TN	Correct	Wrong	% TP	% FP	% FN	% TN	% Correct
4,152	Cytoplasm (C)	4,132	157	20	2,269	6,401	177	0.63	0.02	0	0.34	0.97
1,415	Inner membrane (I)	1,388	238	27	4,925	6,313	265	0.21	0.04	0	0.75	0.96
346	Outer membrane (O)	315	29	31	6,203	6,518	60	0.05	0	0	0.94	0.99
422	Periplasm (P)	393	72	29	6,084	6,477	101	0.06	0.01	0	0.92	0.98
272	Extracellular (S)	250	39	22	6,267	6,517	61	0.04	0.01	0	0.95	0.99
10	Vacuole (V)	8	0	2	6,568	6,576	2	0	0	0	1	1

both 0/1 representation and the GO term representation by the PPV (positive predictive value) score assigned by Argot2 predictor. The numbers of GO terms in the CC, MF, and BP sub-ontologies for each benchmark dataset are reported in Table 2.

In these experiments, a total of 2679 GO terms were selected for the Gram-negative bacterial dataset and 1353 GO terms for Gram-positive bacterial dataset. The number of BP GO terms is significantly larger than that from the other two sub-ontologies; however, our aim in this study is not to assess which of these specific GO terms are influential in the prediction but to combine them to ensure that each query protein has at least one GO term. To predict the subcellular locations both 0/1- and PPV values-based representations of GO terms are assessed to build the final best predictive model.

3.3.2 Cross-validation tests

We adopted the cross-validation method to evaluate the generalization ability of our proposed prediction model against the other studied models since it is still a good validation method for large datasets. To do so, we performed 5-fold cross-validation by randomly dividing protein sequences of each benchmark dataset into five mutually exclusive parts of approximately equal sizes so as the model learns from four parts, and tests are made on the remaining part. Then, the process is repeated and evaluated for all five possible combinations. Firstly, we have evaluated all the models built using different features individually to investigate the impact of the features on the prediction quality, namely, PseAAC, PSSM profiles, and GO terms descriptors. Then, we evaluated the performance using these features fusion for SCL prediction. We therefore set out to test how well Random Forest (RF) and Support Vector Machine (SVM) [119, 120] would perform as baseline classifiers for the multi-label approach used, as SVM is often claimed to be the best at dealing with complex classification problems. The results reported in Supplementary material S1 show that SVM outperformed RF only in the case

Table 6: Performance evaluation by 5-fold cross-validation tests on Gram-positive bacteria proteins using a consensus of both GO terms and PSSM profiles predictions. The multi-label confusion matrix reflects well the predictions performance for each location separetely.

Subcellular locations		Metrics										
		TP	FP	FN	TN	Correct	Wrong	% TP	% FP	% FN	% TN	% Correct
349	Cytoplasm (C)	323	30	26	2,069	2,392	56	0.13	0.01	0.01	0.85	0.98
1,779	Inner membrane (I)	1,768	61	11	608	2,376	72	0.72	0.02	0	0.25	0.97
290	Extracellular (S)	282	101	8	2,057	2,339	109	0.12	0.04	0	0.84	0.96
34	Cell wall (W)	13	3	21	2,411	2,424	24	0.01	0	0.01	0.98	0.99
4	Vacuole (V)	4	0	0	2444	2448	0	0	0	0	1	1

when it learned from PseACC protein features, while RF was the best model when using PSSM profiles and GO terms as feature vectors. Once, the baseline classifier was selected, we exploited each SCL prediction model to infer a consensus prediction so as to increase the possibility of obtaining additional multi-label outputs, since individual predictions provide a poor set of multi-label outputs. The results reported in Table 3 and Table 4 show the performance of each prediction model using 5-fold cross-validation tests on Gram-negative and Gram-positive bacteria respectively. A total of eight prediction models has been obtained using PseAAC features, PSSM profiles, GO terms with both 0/1-, and PPV-based representations, fusion of all the extracted features that we represented by *PseAAC + PSSM + GO*, fusion of only PSSM and GO features represented by *PSSM+GO*, and the fusion of the individual predictions by our proposed consensus approach, where Consensus $_{PSSM+GO}$ stands for the consensus decision obtained from the individual predictions of PSSM profiles- and GO terms-based models and Consensus$_{PseAAC+PSSM+GO}$, from all the three individual prediction models.

From both Table 3 and Table 4, we observe that PSSM profiles-based prediction is significantly better than PseAAC-based prediction, whereas GO terms-based model clearly outperformed the two former individual prediction models. As it will be also observed, the two GO terms-based prediction models performed comparably well. However, the PPV-based representation of GO terms gave better performances than 0/1-based representation which suggests that substituting GO terms by 0/1 values do not truly reflect GO information. Moreover, features fusion by incorporating PseAAC, PSSM profiles and GO terms gave similar performances as by integrating only PSSM profiles and GO terms, which means that no significant gain in precision has been obtained in the presence of PseAAC features. It is clear that the most prominent improvement is achieved by the consensus model that combines the predictions of both PSSM profiles and GO terms-based individual models. The ROC curve provides a more realistic view of the prediction model performance, showing the specificity and sensitivity. The larger the area under the ROC curve, the better is the prediction quality. Here, the ROC curves in Figure 6 and Figure 7 depict well the performance of the individual prediction models built using different features and different ensemble models on Gram-negative bacteria and Gram-positive bacteria datasets. The difference between the studied prediction models is highly visible. It is clear that PseAAC-based model mostly performed rather weak, while the ROC curve of the Consensus prediction model that combines the decisions of both individual models based on PSSM profiles and GO terms is well above. Features fusion curves coincide, reflecting the poor influence of PseAAC features on the prediction quality, however such information would be useful when dealing with proteins that do not share any homology with annotated proteins.

For both datasets, the performances achieved by all eight prediction models are roughly the same. Capturing the most relevant biological features for protein characterization is crucial for prediction effectiveness. Here, it is clear that PseAAC features alone give inconsistent prediction, whereas, PSSM profiles and

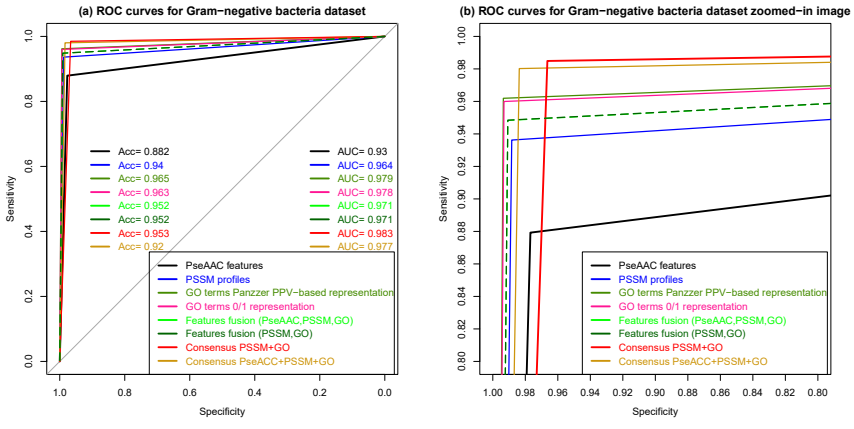

Figure 6: ROC curves of cross-validation tests on Gram-negative bacteria dataset for different predictions pseudo-amino acid composition (PseAAC), PSSM profiles, GO terms 0/1-based representation, GO terms PPV-based representation, features fusion, a consensus prediction using both GO terms and PSSM profiles outputs and a consensus of PseAAC, GO terms, PSSM profiles outputs. In (b) the zoomed-in image of the different curves.

Table 7: The proposed SCL prediction model predictions for multiple sites proteins of Gram-positive bacterial dataset versus CELLO2GO, BUSCA, and UniLoc predictions.

Protein name True location (s)	Predicted essential GO terms			Top ranked CC description	Predicted location (s)						
	MF	BP	CC		PseAAC	PSSM profiles	GO terms	Consensus	CELLO2GO	BUSCA	UniLoc
COOS2_CARHZ C/I	GO:0043885 GO:0018492 GO:0016151 GO:0051539 GO:0051538 GO:0008198 GO:0042803 GO:0016746	GO:0006091 GO:0055114 GO:0006084 GO:0015977	GO:0005886 GO:0005737 GO:0032991	plasma membrane cytoplasm protein-containing complex	I	I	C/I	C/I	C	M	C/M
COMGG_BACSU I/S		GO:0030420	GO:0016021 GO:0005576 GO:0005886	integral component of membrane extracellular region plasma membrane	I	C	C	C	S/M	M	M/S
COOS1_CARHZ C/I	GO:0043885 GO:0018492 GO:0016151 GO:0051539 GO:0051538 GO:0008198 GO:0042803 GO:0050418 GO:0004601	GO:0006091 GO:0055114 GO:0042542 GO:0098869 GO:0006807	GO:0005886 GO:0005737 GO:0032991	plasma membrane cytoplasm protein-containing complex	I	I	I	I	C	M	C/M
ECPA_STAEP S/W	GO:0008234	GO:0006508 GO:0009405	GO:0005576 GO:0005618	extracellular region cell wall	S	S	S	S	S	C	S/W
LYS_CLOAB C/S	GO:0003796 GO:0008233 GO:0008745	GO:0016998 GO:0009253 GO:0042742 GO:0019835 GO:0006508	GO:0005576 GO:0005737 0016021	extracellular region cytoplasm integral component of membrane	S	S	S	S	S	C	S
HYES_CORS2 C/I	GO:0016787 GO:0016746 GO:0140096	GO:0019439 GO:0006508	GO:0005886 GO:0005737	plasma membrane cytoplasm	I	I	C	C/I	M/C	C	M
CPLR_DESHA S/W	GO:0050781 GO:0031419 GO:0016787 GO:0000166 GO:0061783	GO:0046193 GO:0009166 GO:0055114 GO:0016998 GO:0009253	GO:0009275 GO:0005576 GO:0005886 GO:0016021	cell wall extracellular region plasma membrane integral component of membrane	I	S	C	C/S	S	S	S/M/W

Table 7: (continued)

Protein name True location (s)	Predicted essential GO terms			Top ranked CC description	Predicted location (s)						
	MF	BP	CC		PseAAC	PSSM profiles	GO terms	Consensus	CELLO2GO	BUSCA	UniLoc
	GO:0046872	GO:0042742									
	GO:0003723	GO:0019835									
	GO:0016301	GO:0016311									
		GO:0015031									
		GO:0016310									
PLC_LISMO	GO:0008081	GO:0006629	GO:0005576	extracellular region	I	S	S	S	S	M	S
C/S	GO:0004436	GO:0009405	GO:0005737	cytoplasm							
		GO:1901575	GO:0016021	integral component of membrane							
		GO:0006508									

Table 8: The proposed SCL prediction model predictions for some multiple sites proteins of Gram-negative bacterial dataset versus CELLO2GO, BUSCA, and UniLoc predictions.

Protein name	Predicted essential GO terms			Top ranked CC description	Predicted location (s)						
True location (s)	MF	BP	CC		PseAAC	PSSM profiles	GO terms	Consensus	CELLO2GO	BUSCA	UniLoc
CH60_NEIGO	GO:0051082	GO:0042026	GO:0005737	cytoplasm	C	C	C	C	C	C	C
C/O	GO:0005524	GO:0006458	GO:0101031	chaperone complex							
			GO:0009279	cell outer membrane							
ENO_ECOLI	GO:0004634	GO:0006096	GO:0000015	phosphopyruvate hydratase complex	C	C	C	C	C	C	C/S
C/S	GO:0000287	GO:0044210	GO:0009986	cell surface							
	GO:0003883	GO:0006541	GO:0005576	extracellular region							
	GO:0042802		GO:0005856	cytoskeleton							
	GO:0005524		GO:0016020	membraneane							
SPIC_SALTY		GO:0009405	GO:0005576	extracellular region	C	I	S	I/S	S	C	C/S
C/S		GO:0015031	GO:0009279	cell outer membrane							
		GO:0035592	GO:0005737	cytoplasm							
		GO:0002790									
		GO:0032940									
MXIG_SHIFL		GO:0009405	GO:0016021	extracellular region	C	I	I	I	I/C	C	M
I/O			GO:0009279	cell wall							
			GO:0005886								
PGPB_ECOLI	GO:0008962	GO:0016311	GO:0016021	integral component of membrane	I	I	I	I	I	M	M
I/O	GO:0000810	GO:0009395	GO:0009279	cell outer membrane							
	GO:0008195	GO:0006655	GO:0005886	plasma membrane							
	GO:0050380	GO:0009252									
LEPA_ECOLI	GO:0043022	GO:0045727	GO:0005886	integral component of membrane	C	C	C	C	I/C	C	C/M
I/P	GO:0003746	GO:0006414	GO:0005829	cell outer membrane							
	GO:0003924	GO:0009651		plasma membrane							
	GO:0005525	GO:0009268									
	GO:0043024	GO:0009409									
	GO:0043023										
	GO:0042802										
	GO:0016779										
LEPA_SALTY	GO:0043022	GO:0045727	GO:0005886	plasma membrane	I	S	C	C/S	I/C	C	C/M
I/P	GO:0003746	0006414		cytosol							
	GO:0003924										
	GO:0005525										
TIBA_ECOLI			GO:0019867	outer membrane	S	S	S	S	S/O	C	M
O/S			GO:0009986	cell surface							

Table 8: (continued)

Protein name	Predicted essential GO terms			Top ranked CC description	Predicted location (s)						
True location (s)	MF	BP	CC		PseAAC	PSSM profiles	GO terms	Consensus	CELLO2GO	BUSCA	UniLoc
	GO:0005509	GO:0009405	GO:0044462	external encapsulating structure part							
	GO:0004674	GO:0007155	GO:0030313	cell envelope							
	GO:0005524	GO:0006468	GO:0042597	periplasmic space							
			GO:0005576	extracellular region							
			GO:0016021								
FRPC_NEIMC	GO:0005509	GO:0009405	GO:0005576	extracellular region	O	S	S	S	S/C	C	S/C
O/S	GO:0090729	GO:0016311	GO:0009279	cell outer membrane							
	GO:0004035	GO:0005975	GO:0016021	integral component of membrane							
	GO:0008780	GO:0007156									
	GO:0004553										
YOPM_YERPE	GO:0004842	GO:0016567	GO:0045335	phagocytic vesicle	C	C	C	C	C	S/I/C	S/C
O/S	GO:0016874	GO:0007154	GO:0016021	integral component of membrane							
	GO:0043531	GO:0023052									
	GO:0004672	GO:0051716									
	GO:0016787	GO:0050794									
	GO:0005524	GO:0043326									
FRPA_NEIMC	GO:0005509	GO:0009405	GO:0005576	extracellular region	O	S	S	S	S	C	M/S
O/S	GO:0004035	GO:0016311	GO:0016020	membrane							
	GO:0004190	GO:0006629									
		GO:0006508									
TCPS_VIBCH		GO:0009297	GO:0042597	periplasmic space	C	I	S	I/S	P/C	S	P/S
P/S		GO:0009405	GO:0005576	extracellular region							
PPBL_PSEAE	GO:0042301	GO:0035435	GO:0043190	ATP-binding cassette (ABC) trans-	S	S	P	S/P	S	S	S/P/M
S/P	GO:0004035	GO:0016311	GO:0042597	porter complex							
		GO:0015628	GO:0005576	periplasmic space							
		GO:0043952		extracellular region							
HLYE_ECOLI	GO:0090729	GO:0044179	GO:0020002	host cell plasma membrane	C	C	S	S/C	S	C	P/M/S
S/P	GO:0042802	GO:0009405	GO:0042597	periplasmic space							
		GO:0044532	GO:0005576	extracellular region							
			GO:0016021	integral component of membrane							
PPBH_PSEAE	GO:0090729	GO:0016311	GO:0042597	periplasmic space	C	P	P	P	P	S	P/S
S/P	GO:0042802	GO:0015628	GO:0005576	extracellular region							
		GO:0043952									

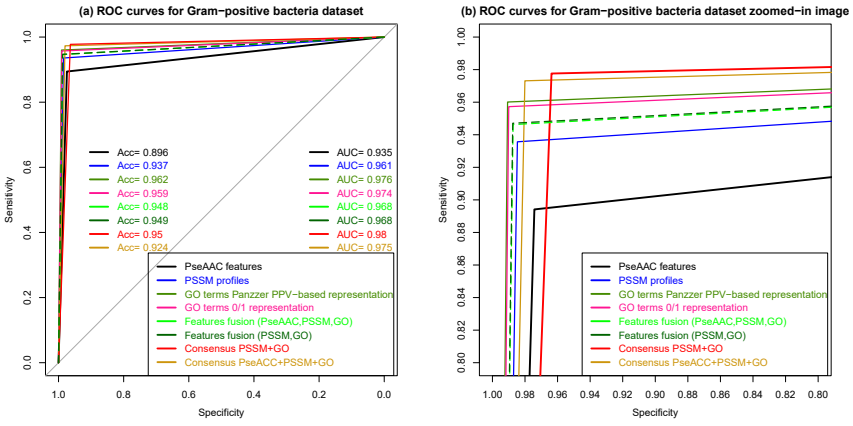

Figure 7: ROC curves of cross-validation tests on Gram-positive bacteria dataset for different predictions pseudo-amino acid composition (PseAAC), PSSM profiles, GO terms 0/1-based representation, GO terms PPV-based representation, features fusion, a consensus prediction using both GO terms and PSSM profiles outputs and a consensus of PseAAC, GO terms,PSSM profiles outputs. In (b) the zoomed-in image of the different curves.

GO terms give significantly better predictive performances, especially the latest which outperformed the two others. Such result demonstrates that evolutionary information has a significant effect on the prediction performance and that GO information is a very good indicator for protein subcellular location. The results also show that in features fusion, when incorporating PseAAC features, no significant gain in performance is obtained since the ROC curves coincide. In addition, the prediction accuracy of the consensus model that considers PseAAC prediction model decisions declines for both Gram-positive and Gram- negative SCL benchmarks. Finally, the highest results are achieved by the consensus of the individual predictions based on PSSM profiles and GO terms. To provide more information about the statistical significance of our proposed predictive model achieved results, we examined the confusion matrix which gives statistics by subcellular location as it is shown in Table 5 for Gram-negative bacteria and Table 6 for Gram-positive bacteria, respectively. The results show that the overall performance of the combined predictions of each individual model is significantly improved, accordingly, the consensus decision of individual models decisions based on different features works better than features fusion in discriminating between the different compartments. The results support our assumption that combining the decisions of diverse individual prediction models can significantly enhance the performance of SCL prediction. Moreover, the incorporated features are biologically more meaningful and need further attention to be well analyzed and exploited in proteins related problems.

3.4 Multiple location prediction

There are only eight multiple sites proteins in Gram-positive bacterial dataset with the statistics I/S (1), C/S (2), S/W (2), and I/C (3) as it is shown in Figure 5. Gram-negative bacterial dataset contains 39 multiple sites proteins among them the minority mutli-label classes are C/O (1), C/S (2), I/O (2), I/P (2), O/S (4), and P/S (4); whereas I/C (18) and C/P (6) are more populated as is highlighted in Figure 4. We have thus, compared the predictions provided by our consensus model against three state-of-art SCL prediction methods, namely CELLO2GO[24] [49], BUSCA[25] [32] and UniLoc[26] [121]. CELL2GO is based on GO terms and uses BLAST to search for homologous sequences. The recent method BUSCA combines three SCL prediction methods (BaCelLo [41], MemLoci [122] and SChloro [123]) and methods for identifying signal and transit peptides, glycosphosphatidylinositol (GPI)-anchors and transmembrane domains. Whereas UniLoc SCL prediction is based on the implicit similarity between proteins, it identifies template proteins based on the number of shared related

24 https://cello.life.nctu.edu.tw/cello2go/.
25 https://busca.biocomp.unibo.it/.
26 https://bioapp.iis.sinica.edu.tw/UniLoc/.

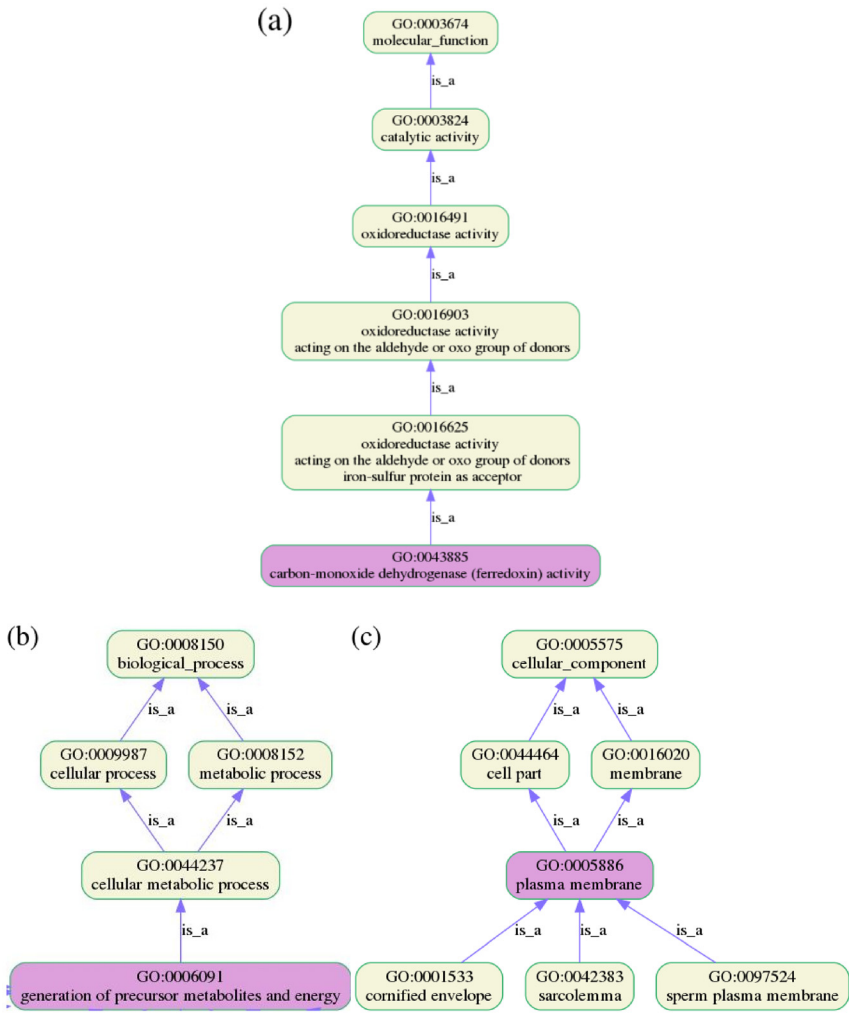

Figure 8: Effect of the presence of the MF GO term GO:0043885, the BP GO term GO:0006091, and the CC GO term GO:0005886 on the proposed model prediction of COOS2_CARHZ (C/I) and COOS1_CARHZ(C/I) proteins where it really succeeded, while they are predicted as cytoplasm (C) or membrane (M) by the others predictors.

words using PSI-BLAST search against NCBI nr[27] database. The results reported in Table 7 and Table 8 show how the predicted locations have been obtained by the consensus model for the minority multi-label classes in Gram-positive and Gram-negative bacterial proteins datasets respectively. One should be careful in drawing conclusions when comparing SCL methods, since they differ in many aspects such as the learning strategy, features that they learn from since they are extracted from different sources of information, and their coverage of different localizations which depends on the available classes in the learning datasets [124]. In addition each SCL prediction method has its strengths and disadvantages and an objective comparison is practically difficult. Here we focused on the minority multi-label classes in both Gram-positive and Gram-negative bacterial benchmarks to observe how they are discerned by the different predictors. It appears that the number of essential GO terms from the MF and BP categories is significantly larger than that from the CC category for both Gram-negative and Gram-positive bacterial datasets. From Table 7, we observe that our proposed model predicts membrane proteins (assigned to M class by the others) as either cytoplasmic or inner membrane or both which is somehow a good result as shown by Figure 8 obtained using GOATOOLS[28] software [125].

We also observe that the other proteins are reasonably predicted and that in the particular case of a total absence of GO terms, which means that even when the prediction is left to chance (arbitrary prediction), the

27 ftp://ftp.ncbi.nlm.nih.gov/blast/db/.
28 https://github.com/tanghaibao/goatools.

proposed prediction model remains robust. From Table 7 and Table 8, we can show that our prediction model recognizes plasma membrane as either inner membrane or cytoplasm and has a strong ability to recognize both cytoplasm and extracellular space classes. Periplasmic space is also well recognized whereas the main difficulty remains in cell wall class prediction. The results show that some GO terms have a critical influence on the decision even in the presence of the others GO terms. For example, the pathogenesis BP GO term GO:0009405 is a good indicator of inner membrane as well as the CC GO term GO:0016021. The BP GO term GO:0016311 reflects the periplasm class whereas the presence of the MF GO term GO:0005509 induces extracellular space. However, it is noticed that even in the absence of MF GO terms, the prediction remains correct, which suggests that the three types of GO terms are complementary for the SCL prediction and when all the three GO categories are present, the prediction is more effective. The reported results show that the decisions of the compared models are roughly complementary and seemingly, taking into account all their decisions might enhance the prediction quality. Further results concerning the more populated multi-label classes I/C (18) and C/P (6) are reported in Supplementary material S2. This experimental evaluation has shown the effectiveness of the proposed multi-label prediction model. However, more investigations are necessary to learn more about GO terms influence on the prediction model decision. We believe that the proposed model effectiveness would be more significant when the benchmark datasets contain more training samples and cover more subcellular location sites. Finally, in order to satisfy the Chou's fifth rule in our future work, we shall provide a web-server for the method presented in this paper. The related datasets can be download from: https://github.com/hb-sources/Protein-SCL-Prediction and the source code for implementing in this study is available from the author upon request.

4 Conclusion

Protein SCL prediction is a challenging problem by its nature since it is an imbalanced multi-label classification problem. Imbalanced because most of the training datasets have an uneven distribution of the proteins in different organelles and muti-label, due to the inherent ability of proteins to simultaneously reside at, or move between two or more different subcellular location sites. In this study, we have proposed an ensemble multi-label SCL prediction system that exploits the potential discriminative power of evolutionary information in the form of PSSM profiles and GO terms to tackle Gram-positive and Gram-negative SCL prediction problem. We have shown that combining individual prediction models decisions is better than features fusion and the assumption that better performance could be expected by combining uncorrelated output predictions since each individual model performs differently proved to be more realistic. Moreover, the results show the superiority of evolutionary information-based prediction, especially when GO annotation is considered, which highlights the usefulness of sequence and structure homology for inferring protein localization and improving the prediction correctness. In the proposed prediction model we have exploited the correlations embedded in label space by using label powerset (LP) transformation strategy with in mind a flat organization structure of the SCLs. However, since proteins trafficking in the cell is highly correlated to the subcellular location sites relationships and organization, in our following research attempts will be made to implement a multiple prediction system considering interdependences between subcellular locations. Investigating the effect of an hierarchical structure organization of the location sites might be an interesting avenue in order to obtain a more effective prediction. Further work will include efforts on collecting more annotated proteins to learn from larger datasets, to extend the coverage scope to further subcellular locations at least for Gram-negative bacteria such as fimbrium, flagellum and nucleoid, and to extend the approach to other organisms. However, further investigations are required to shed light on the underlying mechanisms that govern the positioning of proteins in specific cell sites and on how they are implicated in human diseases.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

1. Wang X, Li S. Protein mislocalization: mechanisms, functions and clinical applications in cancer. Acta Biochim Biophys Sin 2014;1846:13–25.
2. Horton P, Mukai Y, Nakai K. Protein subcellular localization prediction. In: Wong L for Infocomm Research, editors. Review Volume practical-bioinformatician. Singapore: World Scientific Publishing Co. Pte. Ltd; 2004, vol 2, ch 9:193–216 pp.
3. Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Protein Struct Funct Genet 1991;11:95–110.
4. Nakai K, Kanehisa M. A knowledge base for predicting protein localisation sites in eukaryotic cells. Genomics 1992;14: 897–911.
5. Horton P, Nakai K. A probabilistic classification system for predicting the cellular localization sites of proteins. In: Proceedings of intelligent systems in molecular biology. St. Louis, USA; 1996:109–15 pp.
6. Horton P, Nakai K. Better prediction of protein cellular localization sites with the K-nearest neighbors. In: Proceedings of intelligent systems in molecular biology. St. Louis, USA: AAAI Press; 1997:368–83 pp.
7. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999;24:34–6.
8. Lorena A, Carvalho A. Protein cellular localization prediction with support vector machines and decision trees. Comput Biol Med 2007;37:115–25.
9. Scott M, Calafell SJ, Thomas DY, Hallett MT. Refining protein subcellular localization. PLoS Comput Biol 2005;1:e66.
10. King BR, Guda C. ngLOC: an n-gram-based bayesian method for estimating the subcellular proteomes of eukaryotes. Genome Biol 2007;8:R68.
11. Briesemeister S, Rahnenführer J, Kohlbacher O. YLoc-an interpretable web server for predicting subcellular localization. Nucleic Acids Res 2010;38:W497–502.
12. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998; 26:2230–6.
13. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 2000;300:1005–16.
14. Anastasiadis A, Magoulas G. Analysing the localisation sites of proteins through neural networks ensembles. Neural Comput Appl 2006;15:277–88.
15. Shen H, Yang J, Chou K. Methodology development for predicting subcellular localization and other attributes of proteins. Expet Rev Proteonomics 2007;4:453–63.
16. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–8.
17. Chou K. Prediction of protein cellular attributes using pseudo-amino acid composition. Protein Struct Funct Genet 2001;4: 246–55.
18. Cai Y, Liu X, Xu X, Chou K. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence – order effect. J Cell Biochem 2002;84:343–8.
19. Chou K, Cai Y. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 2002;277:45765–9.
20. Bhasin M, Raghava G. SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 2004;32:W414–19.
21. Wang J, Sung W, Krishnan A, Lin K. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. BMC Bioinf 2005;6. https://doi.org/10.1186/1471-2105-6-174.
22. Pierleoni A, Martelli P, Fariselli P, Casadio R. Bacello: a balanced subcellular localization predictor. Bioinformatics 2006;22: 3963–9.
23. Wan S, Mak M, Kung S. mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinf 2012;13:290.
24. Cui Q, Jiang T, Liu B, Ma S. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. BMC Bioinf 2004;5:66.
25. Gardy J, Spencer C, Wang K, Ester M, Tusnady G, Simon I, et al. PSORT-B: improving protein subcellular localization prediction for gram-negative bacteria. Nucleic Acids Res 2003;31:3613–7.

26. Chou K, Shen H. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochemical and biophysical research communications. Biochem Biophys Res Commun 2006;347:150–7.
27. Yu C, Chen Y, Lu C, Hwang J. Prediction of protein subcellular localization. Proteins Struct Funct Bioinf 2006;64:643–51.
28. Guo J, Lin Y, Liu X. GNBSL: a new integrative system to predict the subcellular location for gram-negative bacteria proteins. Proteomics 2006;6:5099–105.
29. Magnusa M, Pawlowskia M, Bujnicki J. MetaLocGramN: a meta-predictor of protein subcellular localization for gram-negative bacteria. Biochim Biophys Acta 2012;1824:1425–33.
30. Wan S, Mak M, Kung S. Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins. BMC Bioinf 2016;17. https://doi.org/10.1186/s12859-016-0940-x.
31. Sperschneider J, Catanzariti A, DeBoer K, Petre B, Gardiner D, Singh K, et al. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Tech Rep, Sci Rep 2017;7. https://doi.org/10.1038/srep44598.
32. Savojardo C, Martelli P, Fariselli P, Profiti G, Casadio R. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res 2018;46:W459–66.
33. Claros M. MitoProt, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci CABIOS 1995;11:441–7.
34. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 1999;8:978–84.
35. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 2002;18:298–305.
36. Krings A, Brameier M, MacCallum R. NucPred-predicting nuclear localization of proteins. Bioinformatics 2007;23:1159–60.
37. Bhasin N, Raghava G. ESLpred: SVM based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 2004;32:W414–9.
38. Sarda D, Chua G, Li K, Krishnan A. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinf 2005;6:152.
39. Tantoso E, Li K. AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids 2008;35:345–53.
40. Briesemeister S, Blum T, Brady S, Lam Y, Kohlbacher O, Shatkay H. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. J Proteome Res 2009;8:5363–6. PMID: 19764776.2.
41. Pierleoni A, Martelli P, Fariselli P, Casadio R. BaCelLo: a balanced subcellular localization predictor. Bioinformatics 2006;22: e408–16.
42. Marcotte E, Xenarios I, vander Bliek A, Eisenberg D. Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci 2000;97:12115–20.
43. Adelfio A, Volpato V, Pollastri G. SCLpredT: Ab initio and homology-based prediction of subcellular localization by N-to-1 neural networks. SpringerPlus 2013;2:502.
44. Lee K, Chuang H, Beyer A, Sung M, Huh W, Lee B, et al. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res 2008;36:e136.
45. Park S, Yang J, Jang S, Kim S. Construction of functional interaction networks through consensus localization predictions of the human proteome. J Proteome Res 2009;8:3367–76.
46. Mondal A, Lin J, Hu J. Network based subcellular localization prediction for multi-label proteins. IEEE Int Conf Bioinf Biomed Workshop (BIBMW) 2011:473–80. https://doi.org/10.1109/bibmw.2011.6112416.
47. Wan S, Mak M, Kung S. Semantic similarity over gene ontology for multi-label protein subcellular localization. Engineering 2013;5:68–72.
48. Wan S, Mak M, Kung S. HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. PLoS One 2014;9:e89545.
49. Yu C, Cheng C, Su W, Chang K, Huang S, Hwang J, et al. CELLO2GO:a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. PLoS One 2014;9:e99368.
50. Shatkay H, Höglund A, Brady S, Blum T, Dönnes P, Kohlbacher O. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. Bioinformatics 2007;23:1410–7.
51. Nielsen H, Almagro A, José J, Sonderby C, Sonderby S, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33:3387–95.
52. Shao W, Ding Y, Shen H, Zhang D. Deep model-based feature extraction for predicting protein subcellular localizations from bio-images. Front Comput Sci 2017;11:243.
53. Zhang N, Rao R, Salvato F, Havelund J, Moller I, Thelen J, et al. MU-LOC: a machine-learning method for predicting mitochondrially localized proteins in plants. Front Plant Sci 2018;9:634.
54. Pang L, Wang J, Zhao L, Wang C, Zhan H. A novel protein subcellular localization method with CNN-XGBoost model for alzheimer's disease. Front Genet 2019;9:751.
55. Yao Y, Li M, Xu H, Yan S, He P.-A, Dai Q, et al. Protein subcellular localization prediction based on PSI-BLAST profile and principal component analysis. Curr Proteomics 2019;16. https://doi.org/10.2174/1570164616666190126155744.

56. Dehzangi A, Sohrabi S, Heffernan R, Sharma A, Lyons J, Paliwal K, et al. Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features. BMC Bioinf 2015;16. https://doi.org/10.1186/1471-2105-16-s4-s1.
57. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary based descriptors into Chouś general PseAAC. J Theor Biol 2015;364:284–94.
58. Yu B, Li S, Qiu W, Wang M, Du J, Zhang Y, et al. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. BMC Genom 2018;19:478.
59. Cheng X, Lin W, Xiao X, Chou K. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. Bioinformatics 2018;35:398–406.
60. Uddin M, Sharma A, Farid D, Rahman M, Dehzangi A, Shatabda S. EvoStruct-Sub: an accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. J Theor Biol 2018;443:138–46.
61. Xiao X, Cheng X, Chen G, Mao Q, Chou K. pLoc_bal-mGpos: predict subcellular localization of gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics 2019;111:886–92.
62. Wan S, Mak MW, Kung SY. mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction. Anal Biochem 2015;473:14–27.
63. Lavezzo E, Falda M, Fontana P, Bianco L. Enhancing protein function prediction with taxonomic constraints-the Argot 2.5 web server. Methods 2016;93:15–23.
64. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
65. Shen H, Chou K. Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. Protein Pept Lett 2009;16:1478–84.
66. Shen H, Chou. Gneg-mPLoc: a top-down approach to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J Theor Biol 2010;264:326–33.
67. Xiao X, Wu Z, Chou K. A multi-label learning classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0020592.
68. Wu Z, Xiao X, Chou K. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram- positive bacterial proteins. Protein Pept Lett 2012;19:4–14.
69. Wang X, Zhang J, Li G. Multi-location gram-positive and gram- negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinf 2015;16. https://doi.org/10.1186/1471-2105-16-s12-s1.
70. Wan S, Mak M, Kung S. Gram-LocEN: interpretable prediction of subcellular multi-localization of gram- positive and gram-negative bacterial proteins. Chemometr Intell Lab Syst 2017;162:1–9.
71. Cheng X, Xiao X, Chou K. pLoc-mGneg: predict subcellular localization of gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics 2018;110:231–9.
72. Chou K. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 2009;6:262–74.
73. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–20.
74. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
75. Chou K. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21: 10–9.
76. Chou K. Artificial intelligence (AI) tools constructed via the 5-steps rule for predicting post-translational modifications. Trends Artif Intell 2019;3:60–74.
77. Du X, Diao Y, Liu H, Li S. MsDBP: exploring DNA-binding proteins by integrating multi-scale sequence information via Chou's 5-steps rule. J Proteome Res 2019;18:3119–32.
78. Ju Z, Wang S. Prediction of lysine formylation sites using the composition of k-spaced amino acid pairs via Chou's 5-steps rule and general pseudo components. Genomics 2019;112:859–66.
79. Butt A, Khan Y. Prediction of S-sulfenylation sites using statistical moments based features via Chou's 5-step rule. Int J Pept Res Therapeut 2019. https://doi.org/10.1007/s10989-019-09931-2.
80. Kabir M, Ahmad S, Iqbal M, Hayat M. iNR-2L: a two-level sequence-based predictor developed via Chou's 5-steps rule and general PseAAC for identifying nuclear receptors and their families. Genomics 2020;112:276–85.
81. Chou K. Some remarks on protein attribute prediction, pseudo amino acid composition (50th anniversary year review). Journal of Theor Biol 2011;273:236–47.
82. Chou K. Recent progresses in predicting protein subcellular localization with artificial intelligence (AI) tools developed via the 5-steps rule. Jpn J Gastroenterol Hepatol 2019;2:1–21.
83. Chou K. Advance in predicting subcellular localization of multi-label proteins and its implication for developing multi- target drugs. Curr Med Chem 2019;26:4918–43.
84. Breiman L. Random forests. Mach Learn 2001;45:5–32.
85. Cherman E, Monard M, Metz J. Multi-label problem transformation methods: a case study. CLEI Electron J 2011;14:4.

86. Chou K. Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 2004;11:2105–34.
87. Chou K. Some illuminating remarks on molecular genetics and genomics as well as drug development. Mol Genet Genom 2020;295:261–74.
88. Chen W, Lei T, Jin D, Lin H, Chou K. PseKNC: a flexible web-server for generating pseudo k-tuple nucleotide composition. Anal Biochem 2014;456:53–60.
89. Du P, Wang X, Xu C, Gao Y. PseAAC-builder: a cross-platform stand-alone program for generating various special Chouś pseudo amino acid compositions. Anal Biochem 2012;425:117–9.
90. Cao D, Xu Q, Liang Y. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 2013;29:960–2.
91. Du P, Gu S, Jiao Y. PseAAC-general: fast building various modes of general form of Chou's pseudo amino acid composition for large-scale protein datasets. Int J Mol Sci 2014;15:3495–506.
92. Liu B, Liu F, Wang X, Chen J, Fang L, Chou K. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 2015;43:W65–71.
93. Liu B, Wu H, Chou K. Pse-in-one 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci 2017;9:67–91.
94. Du P, Zhao W, Miao Y, Wei L, Wang L. UltraPse: a universal and extensible software platform for representing biological sequences. Int J Mol Sci 2017;18. https://doi.org/10.3390/ijms18112400.
95. Cheng J, Randall A, Sweredoski M, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33:W72–6.
96. Jonsson SWJ, Sjörström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 1993; 277:239–53. https://doi.org/10.1016/0003-2670(93)80437-p.
97. Xiao N, Cao D, Zhu M, Xu Q. protr/protrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015;31:1857–9.
98. Chou K, Shen H. Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Sci 2010;2:1090–103.
99. Sokolov A, Ben-Hur A. Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinf Comput Biol 2010;8:357–76.
100. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci 2002;11:2836–47.
101. Yu C, Chen Y, Hwang J. Prediction of protein subcellular localization. Proteins Struct Funct Bioinf 2006;64:643–51.
102. Juncker A, Jensen L, Pierleoni A, Bernsel A, Tress M, Bork P, et al. Sequence-based feature prediction and annotation of proteins. Genome Biol 2009;10. https://doi.org/10.1186/gb-2009-10-2-206.
103. Zhang D, Huang H, Bai X, Xang X, Zhang Y. A high-precision hybrid algorithm for predicting eukaryotic protein subcellular localization. bioRxiv 2019. https://doi.org/10.1101/620179.
104. Koskinen P, Törönen P, Nokso-Koivisto J, Holm L. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics 2015;31:1544–52.
105. Toronen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation webserver. Nucleic Acids Res 2018;46:W84–8.
106. Somervuo P, Holm L. SANSparallel: interactive homology search against uniprot. Nucleic Acids Res 2015;43:W24–9.
107. Koskinen J, Holm L. SANS: high-throughput retrieval of protein sequences allowing 50% mismatches. Bioinformatics 2012;18: i438–43.
108. Jantzen S, Sutherland B, Minkley D, Koop B. GO trimming: Systematically reducing redundancy in large gene ontology datasets. BMC Res Notes 2011;4:267.
109. Conesa A, Götz S, García-Gómez J, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674–6.
110. Conesa A, Götz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. Int J Plant Genom;2008:12, 6198322008.
111. Barrell D, Dimmer E, Huntley R, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009 – an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009;37:D396–403.
112. King B, Vural S, Pandey S, Barteau A, Guda C. ngLoc: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. BMC Res Notes 2012;5:351.
113. Tsoumakas G, Katakis I. Multi-label classification: an overview. Int J Data Warehous Min 2007;3:1–13.
114. Madjarov G, Kocev D, Gjorgjevikj D, Deroski S. An extensive experimental comparison of methods for multi-label learning. Pattern Recogn 2012;45:3084–104.
115. Zhang M, Zhou Z. A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 2014;26:1819–37.
116. Charte, F, Charte, D. Working with multilabel datasets in R: the MLDR package. R J 2015;7:149–62.
117. Rivolli A, de Carvalho A. The utiml package: multi-label classification in R. R J 2018. https://doi.org/10.32614/rj-2018-041.
118. Chou K, Cai Y. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 2003;311:743–7.
119. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
120. Vapnik V, Statistical learning theory. New York: John Wiley & Sons, Inc.; 1998.

121. Lin H, Chen C, Sung T, Hsu W. UniLoc: a universal protein localization site predictor for eukaryotes and prokaryotes. bioRxiv, 2018. https://doi.org/10.1101/252916.
122. Pierleoni A, Martelli P, Casadio R. MemLoci: predicting subcellular localization of membrane proteins in eukaryotes. Bioinformatics 2011;27:1224–30.
123. Savojardo C, Martelli P, Fariselli P, Casadio R. SChloro: directing viridiplantae proteins to six chloroplastic sub-compartments. Bioinformatics 2016;33:347–53.
124. Assfalg J, Gong J, Kriegel H, Pryakhin A, Wei T, Zimek A. Supervised ensembles of prediction methods for subcellular localization. J Bioinf Comput Biol 2009;7:269–85.
125. Klopfenstein D, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: a python library for gene ontology analyses. Sci Rep 2018;8:10872.

Supplementary Material: The online version of this article offers supplementary material https://doi.org/10.1515/jib-2019-0091.