https://doi.org/10.1016/j.chemolab.2016.12.014

Gram-LocEN: Interpretable prediction of subcellular multi-localization of Gram-positive and Gram-negative bacterial proteins

Shibiao Wan[a,*], Man-Wai Mak[b,*], Sun-Yuan Kung[a]

[a]*Department of Electrical Engineering, Princeton University, New Jersey, USA*
[b]*Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China.*

Abstract

Bacteria have a highly organized internal architecture at the cellular level. Identifying the subcellular localization of bacterial proteins is vital to infer their functions and design antibacterial drugs. Recent decades have witnessed remarkable progress in bacterial protein subcellular localization by computational approaches. However, existing computational approaches have the following disadvantages: (1) the prediction results are hard to interpret; and (2) they ignore multi-location bacterial proteins. To tackle these problems, this paper proposes an interpretable multi-label predictor, namely Gram-LocEN, for predicting the subcellular localization of both single- and multi-location proteins of Gram-positive or Gram negative bacteria. By using a multi-label elastic-net (EN) classifier, Gram-LocEN is capable of selecting location-specific essential features which play key roles in determining the subcellular localization. With these essential features, not only where a bacterial protein resides can be decided, but also why it locates there can be revealed. Experimental results on two stringent benchmark datasets suggest that Gram-LocEN significantly outperforms existing state-of-the-art multi-label predictors for both Gram-positive and Gram-negative bacteria. For readers' convenience, the Gram-LocEN web-server is available at `http://bioinfo.eie.polyu.edu.hk/Gram-LocEN/`.

Keywords: Interpretable predictor; multi-location proteins; bacterial protein subcellular localization; Gram-positive; Gram-negative.

1. Introduction

As a domain of prokaryotic microorganisms, bacteria were among the first life forms on Earth. Bacteria have a diversity of shapes, including spirals, rods and spheres. With the Gram-staining technique, bacteria are often classified as Gram-positive bacteria and Gram-negative bacteria. The former are stained violet or dark blue, whereas the latter do not retain the stain but instead appear pink after a counterstain is added. Bacteria can form complex relationships with other organisms, including parasitism, mutualism and commensalism. While pathogenic bacteria can cause various kinds of human diseases, such as tuberculosis, foodborne illness, tetanus, leprosy and typhoid fever [1], over thousand types of bacteria in the normal human gut flora contribute to gut immunity, vitamins synthesis and sugars-to-lactic acid conversion [2]. Therefore, studies of bacteria are of paramount significance for antibacterial drug design and basic research.

Despite of the simple cellular structure and the lack of nuclei and complicated organelles such as mitochondrion or chloroplast, bacteria have an intricate subcellular structure in which proteins are dynamically located in particular sites of a cell [3]. Knowing where a protein resides within a bacterial cell is vitally important to understanding its functions. With a huge number of protein sequences discovered in the post-genomic era, it is highly required to develop computational approaches to assist conventional time-consuming and laborious wet-lab experiments.

Recent decades have witnessed a remarkable development in computational prediction of protein subcellular localization (PSCL). Conventional approaches can be divided into four categories: (1) homology-based methods [4, 5]; (2) sorting-signals based methods [6, 7, 8]; (3) amino-acid composition-based methods [9, 10, 11, 12]; and (4) knowledge-based methods. The former three only extract information from protein primary sequences, whereas the last one extracts information from both protein sequences and knowledge databases, including Gene Ontology (GO)[1] based methods [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23], PubMed abstracts based methods [24, 25], or Swiss-Prot keywords based methods [26, 27]. Particularly, GO-based methods were found to be superior [13, 28, 29, 30] among these methods.

However, many of the aforementioned methods fail to predict multi-location proteins which may simultaneously reside in, or move between two or more subcellular compartments. In fact, the multi-location proteins are prevalently found in living organisms [31, 32, 33, 34], playing major roles in various biological activities in multiple cellular compartments. For example, proteins involved in fatty acid β-oxidation are found to

*Corresponding author
Email addresses: `shibiao.wan@princeton.edu` (Shibiao Wan), `enmwmak@polyu.edu.hk` (Man-Wai Mak), `kung@princeton.edu` (Sun-Yuan Kung)

[1] http://www.geneontology.org

locate in the peroxisome and mitochondria; antioxidant defense proteins are known to reside in the peroxisome, cytosol and mitochondria [35]; and the glucose transporter GLUT4 is found in both the plasma membrane and the intracellular vesicles of adipocytes [36, 37].

Recently, several predictors [38, 39, 39, 40, 41, 42, 43] have been proposed to predict gram-positve and gram-negative bacterial proteins. Among them, Gpos-PLoc [38] and Gneg-PLoc [40] can only deal with single-location proteins for Gram-positive and Gram-negative bacteria, respectively. Besides, to the best of our knowledge, only five predictors, namely Gneg-ECC-mPLoc [43], Gpos-mPLoc [39], iLoc-Gpos [44], Gneg-mPLoc [41] and iLoc-Gneg [42], can predict both single- and multi-label bacterial (Gram-positive, Gram-negative or both) proteins. All of these predictors use GO information as the features and adopt various kinds of multi-label classifiers, such as multi-label K-nearest neighbor classifiers, or multi-label ensemble of classifier chains. However, all of these predictors can only predict where the query proteins are located, but cannot give biological insights on why they reside there. In fact, biologists usually want to know not only the predictions results but also the biological reasons that lead to the prediction results. The lack of interpretability may limit the applications of these predictors.

To address the aforementioned problems, this paper proposes an interpretable multi-label predictor, namely Gram-LocEN for large-scale predictions of both single- and multi-location proteins in Gram-positive and Gram-negative bacteria. Specifically, given a query protein, its GO feature information is retrieved from two newly created compact databases by the procedures described in [45]. By using the one-vs-rest multi-label elastic-net (EN) classifier, a small number of GO terms were selected from more than 8000 GO terms, which form a dimension-reduced feature space consisting of essential GO terms responsible for the final predictions. Subsequently, the dimension-reduced feature vectors are classified by a multi-label EN classifier. Compared to existing multi-label bacterial predictors, Gram-LocEN can not only determine where a bacterial protein locates within a cell, but also give insights on why it belongs there. Experimental results based on two stringent bacterial benchmark datasets demonstrate that Gram-LocEN remarkably outperforms existing state-of-the-art predictors.

2. Feature Extraction

2.1. Creating Compact Databases

Conventional GO-based approaches rely on two important databases, namely Swiss-Prot and GOA databases. Typical procedures are as follows: given a query protein, BLAST [46] is used to retrieve its top homologous protein from the Swiss-Prot database, whose accession number (AC) is used as a key to search against the GOA database for the GO information. In this case, the homologous GO information can be transferred to the query protein. However, these methods will become ineffective when there are no GO terms associated with the AC of the top homolog. In such case, some predictors use back-up methods that rely on other features, such as pseudo-amino-acid composition [9] and sorting signals [47]; some predictors [30, 48] use a successive-search strategy to make sure that at least one annotated GO term exists. Nonetheless, these strategies may lead to poor performance and increase computation and storage complexity.

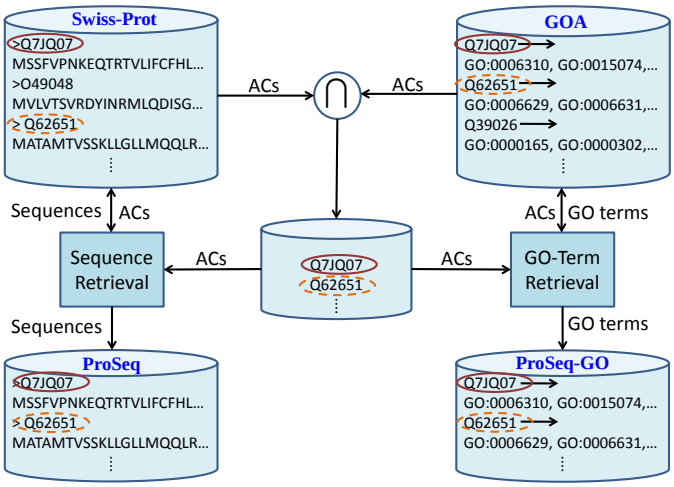


Figure 1: Procedures of creating the compact databases (ProSeq and ProSeq-GO). The circle with the same color represents the same AC. *AC*: accession number; *GO*: gene ontology; *GOA*: gene ontology annotation database.

To address the aforementioned problems, similar to our earlier work [45], we created two small yet efficient databases: ProSeq and ProSeq-GO. The procedures of creating these databases is shown in Fig. 1. As can be seen in Fig. 1, ProSeq is a sequence database extracted from the Swiss-Prot database and ProSeq-GO is a GO-term database extracted from the GOA database.[2] Detailed procedures can be found in [45]. By using ProSeq and ProSeq-GO, we can guarantee that every query protein can be associated with at least one GO term. Also, the memory consumption can be significantly reduced.

2.2. Constructing GO Vectors

Before we extract the features, we need to define the GO space. Let $\mathcal{W}$ denotes a set of distinct GO terms corresponding to a training dataset. $\mathcal{W}$ is constructed in two steps: (1) identifying all of the GO terms in the dataset and (2) removing the repetitive GO terms. Suppose W distinct GO terms are found, i.e., $|\mathcal{W}|= W$; these GO terms form a GO Euclidean space with W dimensions.

Then, features relevant to subcellular localization are extracted in two steps: (1) retrieval of GO terms; and (2) construction of GO vectors. For the former, the amino acid sequence of a query protein is presented to BLAST [46] to find its homologs in the ProSeq database. The homologous ACs are then used as keys to search against the ProSeq-GO database to obtain a set of GO terms associated with the query protein. For the latter,

[2] We have preprocessed the GOA database so that we could obtain a hashtable where the protein ACs are the keys and the corresponding sets of GO terms are the value sets.

the term-frequency method [48, 30] is adopted to construct the GO vectors. Specifically, within the W-dim GO space, the GO vector $\mathbf{q}_i$ of the i-th protein $\mathbb{Q}_i$ is defined as:

$$\mathbf{q}_i = [f_{i,1}, \ldots, f_{i,j}, \ldots, f_{i,W}]^{\mathsf{T}}, \tag{1}$$

where $f_{i,j}$ is the number of occurrences of the j-th GO term (term-frequency) in the i-th protein sequence. Detailed information about GO vectors can be found in [48, 30].

3. Multi-label Elastic Net Based Classifier

We have demonstrated previously [49] that the multi-label LASSO [50] based classifier can be used to interpreting and predicting subcellular localization of human proteins. Due to its sparseness property, LASSO can produce "parsimonious" solutions through which a set of features that are the most relevant to the problem (target variables) can be found. LASSO uses an L_1-regularized linear regression model whose constraint forces the weights of some features to exactly zero [51]. This property enables LASSO to automatically select relevant features. However, LASSO tends to produce very sparse solutions, causing some important information to be excluded from the feature list.

To overcome this disadvantage, the elastic net (EN) was proposed in [52]. EN uses a convex combination of L_1- and L_2-penalties to yield sparse representations similar to LASSO, while encouraging correlated features to be selected or deselected together. Actually, LASSO can be regarded as a special case of EN. EN has been extensively used in various bioinformatics domains, such as ICU mortality risk detection [53], single nucleotide polymorphism (SNP) selection [54] and genetic trait prediction [55].

3.1. Single-Label EN Classifier

Assume that in a two-class single-label problem, we are given a set of training data $\{\mathbf{q}_i, y_i\}_{i=1}^N$, where $\mathbf{q}_i \in \mathcal{R}^W$ and $y_i \in \{-1, 1\}$. $\mathbf{q}_i$ is defined in Eq. 1.

Generally speaking, EN is to impose an $(L_1 + L_2)$-style regularization to ordinary least squares (OLS):

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - f(\mathbf{q}_i))^2 = \sum_{i=1}^{N}\left(y_i - \varepsilon_0 - \sum_{j=1}^{W}\beta_j q_{i,j}\right)^2, \tag{2}$$

subject to

$$\sum_{j=1}^{W}|\beta_j| \leq t_1 \text{ and } \sum_{j=1}^{W}\beta_j^2 \leq t_2,$$

where $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_j, \ldots, \beta_W]^{\mathsf{T}}$ is the EN vector to be optimized, t_1 and t_2 are two positive parameters controlling the shrinkage level applied to $\boldsymbol{\beta}$, ε_0 is a bias,[3] and $q_{i,j}$ is the j-th element of $\mathbf{q}_i$. The constrained minimization in Eq. 2 is equivalent to the following minimization:

$$\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{N}\left(y_i - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{q}_i\right)^2 + \rho\sum_{j=1}^{W}|\beta_j| + \eta\sum_{j=1}^{W}\beta_j^2, \tag{3}$$

where $\rho > 0$ and $\eta > 0$ are the penalty parameters controlling the ridge regression penalty and LASSO penalty, respectively. As can be seen, when $\rho = 0$, Eq. 3 becomes simple ridge regression; when $\eta = 0$, Eq. 3 is exactly the same as LASSO. Besides, by simple transformation, Eq. 3 can be converted to an equivalent LASSO-style problem on augmented data [52]. Because of this property, Eq. 3 can be solved by the same way as LASSO by absorbing the L_2-norm term into the sum of squared errors. Detailed descriptions of the solutions can be found in [52].

3.2. Multi-label EN for Feature Selection

In an M-class multi-label problem, the training data set is written as $\{\mathbf{q}_i, \mathcal{Y}_i\}_{i=1}^N$, where $\mathbf{q}_i \in \mathcal{R}^W$ and $\mathcal{Y}_i \subset \{1, 2, \ldots, M\}$ is a set which may contain one or more labels.

For the multi-label EN, M independent binary one-vs-rest ENs are trained, one for each class. The labels $\{\mathcal{Y}_i\}_{i=1}^N$ are converted to *transformed labels* [48] $y_{i,m} \in \{-1, 1\}$, where $i = 1, \ldots, N$, and $m = 1, \ldots, M$. Then, the EN vector for the m-th class is given by:

$$\hat{\boldsymbol{\beta}}_m = \arg\min_{\boldsymbol{\beta}_m}\left\{\sum_{i=1}^{N}\left(y_{i,m} - \boldsymbol{\beta}_m^{\mathsf{T}}\mathbf{q}_i\right)^2 + \rho_m\sum_{j=1}^{W}|\beta_{j,m}| + \eta_m\sum_{j=1}^{W}\beta_{j,m}^2\right\}, \tag{4}$$

respectively. In Eq. 4, $m = 1, \ldots, M$, $\{y_{i,m}\}_{i=1}^N \in \{-1, 1\}$, ρ_m and η_m are the L_1 and L_2 penalized parameters for the m-th class, respectively. Since L_1 regularization tends to force some weights $\{\beta_{j,m}\}_{j=1}^W$ for the m-th class to exactly zero, EN can be used for feature selection. Compared to LASSO, EN yields less parsimonious solutions.

The GO vectors obtained from Eq. 1 are used for training multi-label one-vs-rest EN classifiers. For an M-class problem (here M is the number of subcellular locations), M independent binary EN classifiers are trained, one for each class. After training, the union of those GO terms with non-zero weights ($\beta_{j,m} \neq 0$) in any one of the M classes constitutes the selected features. EN can remarkably remove those irrelevant features (or GO terms). Suppose S out of the T weights are nonzero, which forms a set $\mathcal{S}$, i.e., $|\mathcal{S}| = S$. They are defined as $\{\beta_{s,m}\}_{s\in\mathcal{S},m=\{1,\ldots,M\}}$ and their corresponding GO terms are called *essential GO terms*. In fact, in our experiments, through the proposed multi-label EN classifiers, 162 (or 245) out of 8110 GO terms were selected for the Gram-positive (or Gram-negative) bacterial dataset. This means that only around 2% (or 3%) of the 8110 GO terms are *essential GO terms* for the Gram-positive (Gram negative) bacterial dataset. In other words, the weights for about 98% (97%) of the 8110 GO terms for the Gram-positive (Gram-negative) bacterial dataset are exactly zero.

[3] For ease of presentation, we omitted the bias in equations in the sequel.

3.3. Multi-label EN for Classification

Besides feature selection, EN can also be used for classification. Specifically, given the t-th query protein $\mathbb{Q}_t$, the feature vector $\mathbf{q}_t \in \mathcal{R}^W$ defined in Eq. 1 is obtained. Then, the elements of $\mathbf{q}_t$ with non-zero weights $\beta_{j,m}$ (Eq. 4) are selected to form a low-dimensional feature vector represented by $\mathbf{q}_t^s \in \mathcal{R}^S$, where $S < T$ is the number of essential GO terms. Similarly, for an M-class problem, M independent binary EN classifiers are trained, one for each class. Then, the score of the m-th EN is:

$$s_m(\mathbb{Q}_t) = \tilde{\boldsymbol{\beta}}_m^{\mathsf{T}} \mathbf{q}_t^s, \tag{5}$$

where $\tilde{\boldsymbol{\beta}}_m$ for EN is given by

$$\tilde{\boldsymbol{\beta}}_m = \arg\min_{\boldsymbol{\alpha}_m} \left\{ \sum_{i=1}^{N} \left(y_{i,m} - \boldsymbol{\alpha}_m^{\mathsf{T}} \mathbf{q}_i^s \right)^2 + \rho_m \sum_{j=1}^{S} |\alpha_{j,m}| + \eta_m \sum_{j=1}^{S} {\alpha_{j,m}}^2 \right\}, \tag{6}$$

where $\boldsymbol{\alpha}_m = [\alpha_{1,m}, \ldots, \alpha_{j,m}, \ldots, \alpha_{S,m}]^{\mathsf{T}}$ is the weight vector to be optimized and $\mathbf{q}_i^s \in \mathcal{R}^S$ is the feature vector for the i-th training protein. Note that $\tilde{\boldsymbol{\beta}}_m$ is obtained based only on the training data.

To predict the subcellular locations of datasets containing both single-label and multi-label proteins, a decision scheme for multi-label EN classifiers should be used. Unlike the single-label problem where each protein has one predicted label only, a multi-label protein should have more than one predicted labels. In this paper, we used the decision scheme described in mGOASVM [48]. In this scheme, the predicted subcellular location(s) of the i-th query protein are given by:

$$\mathcal{M}^*(\mathbb{Q}_t) = \begin{cases} \bigcup_{m=1}^{M} \{m : s_m(\mathbb{Q}_t) > 0\}, \text{ where } \exists\ s_m(\mathbb{Q}_t) > 0\ ; \\ \\ \arg\max_{m=1}^{M} s_m(\mathbb{Q}_t), \text{ otherwise.} \end{cases} \tag{7}$$

For ease of presentation, we refer to the proposed predictor as Gram-LocEN.

4. Experiments

4.1. Datasets

Two stringent bacterial benchmark datasets published recently [39, 41] were used to evaluate the performance of Gram-LocEN. Both datasets were created from Swiss-Prot 55.3. The Gram-positive dataset [39] contains 519 proteins distributed in 4 subcellular locations, whereas the Gram-negative dataset [41] has 1392 proteins distributed in 8 subcellular locations. The sequence identity of both datasets was limited to 25%. Fig. 2 shows the breakdown of the two datasets. The Gram-positive bacterial dataset comprises 519 actual proteins [48] which correspond to 523 locative proteins [48, 56].[4] Among these 519 actual proteins, 515 belong to one location, 4 to two locations and none to more than two locations. In the Gram-negative dataset, there are 1392 actual proteins corresponding to 1456 locative proteins. Among the 1392 actual proteins, 1328 belong to one location, 64 to two locations and none to more than two locations. As can be seen from Fig. 2(a), the majority (97%) of the Gram-positive bacterial proteins are located in the cell membrane, cytoplasm and extracellular, whereas in Fig. 2(b), 78% of the Gram-negative bacterial proteins are located in the cell inner membrane, cytoplasm and periplasm. This means that both datasets are very imbalanced.

[4] Locative proteins are defined as follows. If a protein exists in two different subcellular locations, it will be counted as two locative proteins; if a protein coexists in three locations, then it will be counted as three locative proteins; and so forth.

4.2. Performance Metrics

In multi-label classification scenarios, some sophisticated performance metrics are used for performance measurement to better reflect the multi-label capabilities of classifiers. Two typical measures [57, 48], namely overall locative accuracy (*OLA*) and overall actual accuracy (*OAA*), are often used in multi-label subcellular localization prediction. Specifically, denote $\mathcal{L}(\mathbb{Q}_i)$ and $\mathcal{M}(\mathbb{Q}_i)$ as the true label set and the predicted label set for the i-th protein $\mathbb{Q}_i$ ($i = 1, \ldots, N$), respectively. Then, *OLA* is given by:

$$OLA = \frac{1}{\sum_{i=1}^{N} |\mathcal{L}(\mathbb{Q}_i)|} \sum_{i=1}^{N} |\mathcal{M}(\mathbb{Q}_i) \cap \mathcal{L}(\mathbb{Q}_i)|, \tag{8}$$

and the overall actual accuracy (*OAA*) is:

$$OAA = \frac{1}{N} \sum_{i=1}^{N} \Delta[\mathcal{M}(\mathbb{Q}_i), \mathcal{L}(\mathbb{Q}_i)] \tag{9}$$

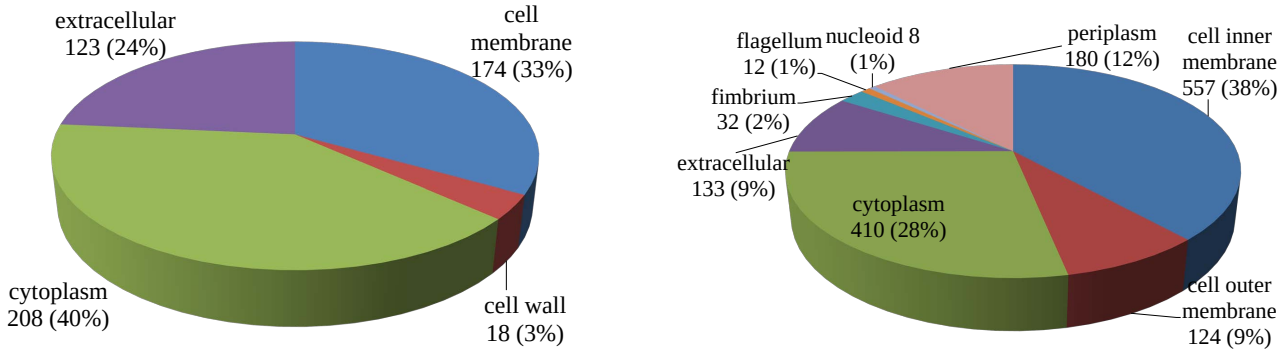
where

$$\Delta[\mathcal{M}(\mathbb{Q}_i), \mathcal{L}(\mathbb{Q}_i)] = \begin{cases} 1 & \text{, if } \mathcal{M}(\mathbb{Q}_i) = \mathcal{L}(\mathbb{Q}_i) \\ 0 & \text{, otherwise.} \end{cases} \tag{10}$$

In addition, some other measures include *Accuracy, Precision, Recall, F1-score (F1)* and *Hamming Loss (HL)*. The definitions of these five measurements for multi-label classification can be found in [58, 45]. Specifically, *Accuracy, Precision, Recall* and *F1* indicate the classification performance. The higher the measures, the better the prediction performance. Among them, *Accuracy* is the most commonly used criteria. *F1-score* is the harmonic mean of *Precision* and *Recall*, which allows us to compare the performance of classification systems by taking the trade-off between *Precision* and *Recall* into account. The *Hamming Loss (HL)* [59, 60] is different from other metrics. The lower the *HL*, the better the prediction performance.

Among all the metrics mentioned above, *OAA* is the most stringent and objective. This is because if some (but not all) of the subcellular locations of a query protein are correctly predicted, the numerators of the other five measures (including *Accuracy, Precision, Recall, F1* and *OLA*) are non-zero, whereas the numerator of *OAA* in Eq. 9 is 0 (thus contributing nothing to the frequency count). More details about the performance metrics can be found in the supplementary materials.

Statistically speaking, leave-one-out cross validation (LOOCV) is considered to be the most rigorous and bias-free procedure [61] for evaluating classifiers' performance. Thus, LOOCV was used to examine the performance of Gram-LocEN.



(a) The Gram-Positive Bacterial Dataset

(b) The Gram-Negative Bacterial Dataset

Figure 2: Breakdown of datasets. In (a), the Gram-positive dataset contains 519 actual proteins corresponding to 523 locative proteins, whereas the Gram-negative dataset in (b) contains 1392 actual proteins corresponding to 1456 locative proteins.

5. Results and Discussions

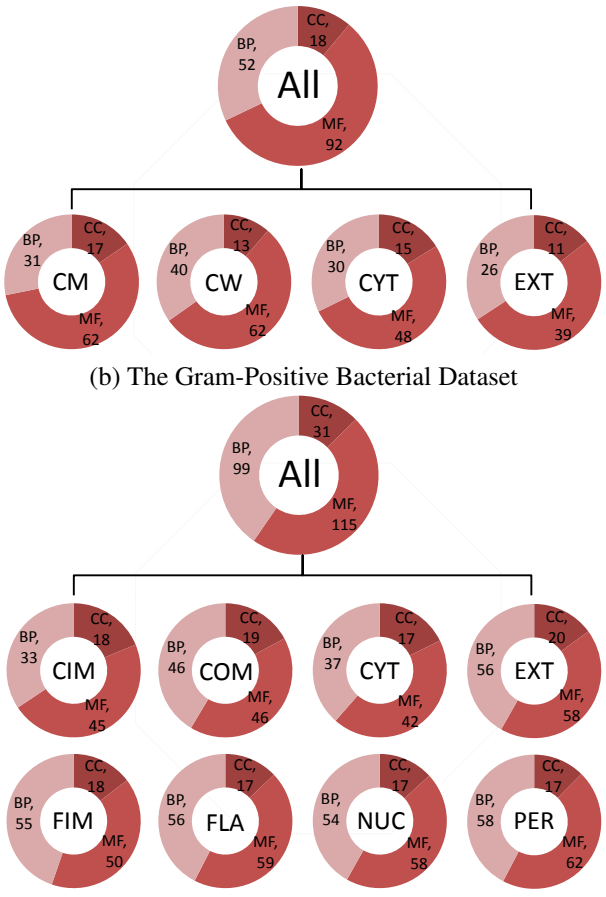
5.1. Statistical Analysis of the Essential GO Terms

Fig. 3 shows the location-specific categorical breakdown of the essential GO terms found by Gram-LocEN for (a) the Gram-positive bacterial dataset and (b) the Gram-negative bacterial dataset. Fig. 3(a) shows that for each subcellular location, the essential GO terms come from not only the cellular-component category, but also from the other two categories. Besides, we can observe that the number of essential GO terms from the molecular-function category is significantly larger than that from the other two categories. For example, for *cell membrane*, 110 essential GO terms contribute to the final decisions, of which 62 belong to the molecular-function (MF) category; the remaining 17 and 31 belong to cellular-component (CC) and biological-process (BP) categories, respectively; in other words, more than half of the essential GO terms are from the MF category. However, as shown in Fig. 3(b), the percentage of MF GO terms found by Gram-LocEN is much smaller for the Gram-negative bacterial dataset. For example, for *fimbrium*, only around 40% (50 out of 123) belongs to MF, whereas 45% (55 out of 123) essential GO terms are from BP.

When we group the essential GO terms according to their GO categories (BP, CC, and MF) without taking the location-specific information into account, we obtain the pie charts (labeled with "All") at the top of Fig. 3(a) and Fig. 3(b). Obviously, the number of essential GO terms across all locations (upper pie chart) is much smaller than the sum of the location-specific GO terms (lower pie charts), suggesting that some of essential GO terms appear in multiple classes and may contribute to the prediction of more than one location.

5.2. Significance of Location-Specific GO Terms

To quantitatively demonstrate how and to what extent essential GO terms contribute to the prediction of subcellular locations, we analyzed the location-specific weights $\{\tilde{\boldsymbol{\beta}}_m\}_{m=\{1,\ldots,M\}}$ defined in Eq. 6 for the essential GO terms.[5] Fig. 4 shows the



(b) The Gram-Positive Bacterial Dataset

(d) The Gram-Negative Bacterial Dataset

Figure 3: Location-specific categorical breakdown of essential GO terms for (a) the Gram-positive bacterial dataset and (b) the Gram-negative bacterial dataset. In (a), the 4 subcellular locations for the Gram-positive bacterial dataset include: cell membrane (CM), cell wall (CW), cytoplasm (CYT) and extracellular (EXT); In (b), the 8 subcellular locations for the Gram-negative bacterial dataset include: cell inner membrane (CIM), cell outer membrane (COM), cytoplasm (CYT), extracellular (EXT), fimbrium (FIM), flagellum (FLA), nucleoid (NUC) and periplasm (PER). *CC*: cellular component; *MF*: molecular function; and *BP*: biological function.

[5] Specific weights $\{\beta_{s,m}\}_{s\in\mathcal{S},m=\{1,\ldots,M\}}$ of each subcellular location for both Gram-positive and Gram-negative bacterial datasets by Gram-LocEN can be found in the supplementary materials.

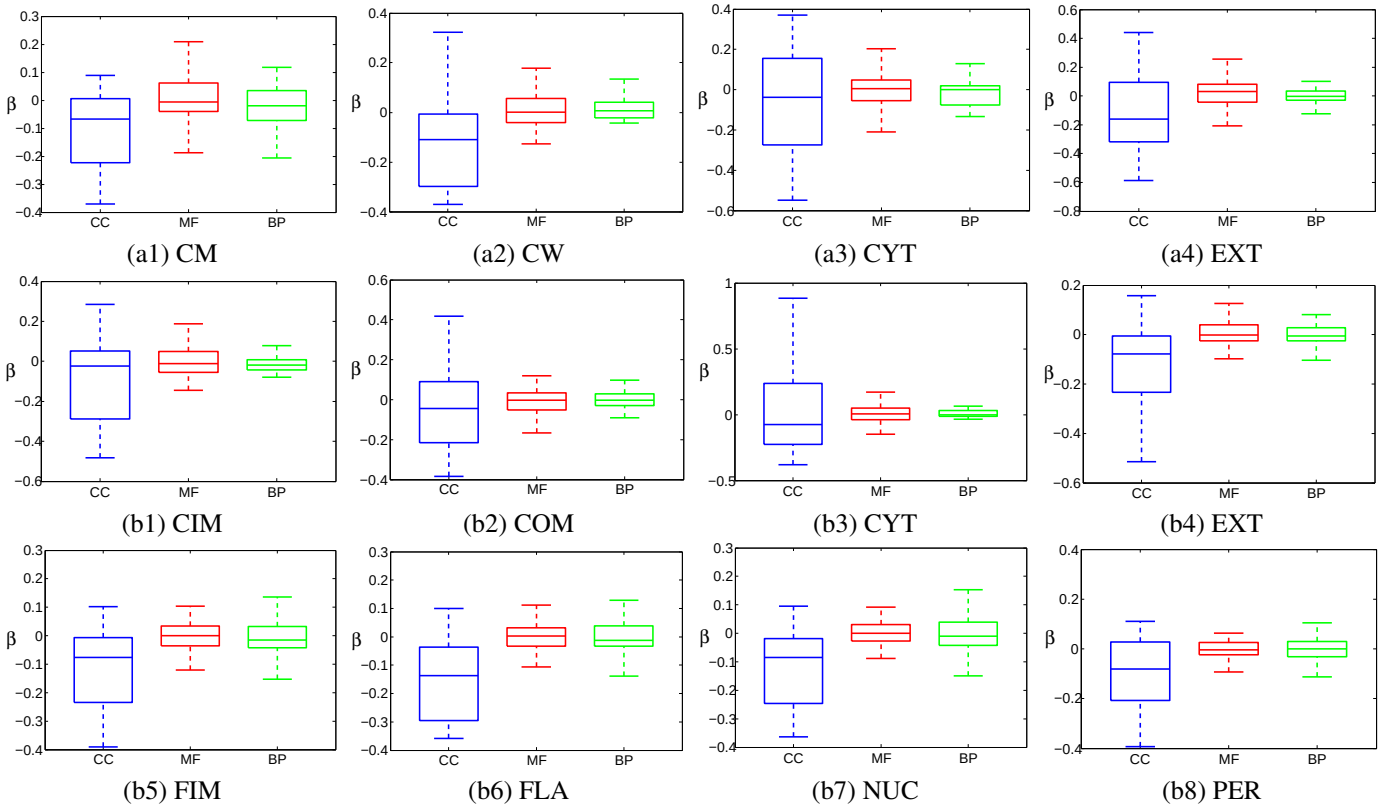


Figure 4: Categorical significance of essential GO terms for different subcellular locations for the Gram-positive bacterial and Gram-negative bacterial datasets. (a1)–(a4) are for the 4 subcellular locations of the Gram-positive bacterial dataset, whereas (b1)–(b8) are for the 8 subcellular locations of the Gram-negative bacterial dataset. The 4 subcellular locations for the Gram-positive bacterial dataset include: cell membrane (CM), cell wall (CW), cytoplasm (CYT) and extracellular (EXT); the 8 subcellular locations for the Gram-negative bacterial dataset include: cell inner membrane (CIM), cell outer membrane (COM), cytoplasm (CYT), extracellular (EXT), fimbrium (FIM), flagellum (FLA), nucleoid (NUC) and periplasm (PER).

boxplots of categorical significance of location-specific essential GO terms for both the Gram-positive and Gram-negative bacterial datasets. Specifically, Fig. 4(a1)–(a4) are for the weights of essential GO terms of the 4 subcellular locations of the Gram-positive bacterial dataset, whereas Fig. 4(b1)–(b8) are for those of the 8 subcellular locations of the Gram-negative bacterial dataset. For simplicity, $\beta_{s,m}$ is abbreviated as β in the figures. Note that GO terms with positive weights and negative weights are both essential GO terms which contribute to the final prediction decisions. GO terms with positive weights for a particular subcellular location indicate that the query protein resides in this subcellular location; the larger the positive weight of a GO term is, the higher confidence the indication has. On the contrary, GO terms with negative weights for a particular subcellular location indicate that the query protein does **not** belong to this subcellular location; the smaller the negative weight of a GO term is, the higher confidence the indication has.

As shown in Fig. 4(a1), for the *cell membrane*, we can see that the median and maximum weights of CC GO terms are smaller than those of GO terms from the other two categories. Besides, the essential GO term with the largest weight (> 0.2) is from the MF category. These results suggest that the essential GO terms from the CC category may play a less significant role in indicating the presence of a query protein in *cell membrane*. However, the cases are different for the other three subcellular locations as shown in Fig. 4(a2)–(a4), where the CC GO terms have a wider range of weights and essential GO terms with the minimum and the maximum weights are from the CC category. On the other hand, the weights of the MF and BP GO terms are within a much smaller range and the absolute values of their minimum/maximum weights are much smaller than those from the CC category. This indicates that the CC GO terms contribute more to the final prediction of these three subcellular locations than those GO terms from the MF and BP categories.

For the Gram-negative bacterial dataset, as shown in Fig. 4(b1)–(b8), the CC GO terms have a wider range than the GO terms of the other two categories for all of the 8 subcellular locations. However, the CC GO terms for the first three subcellular locations (*cell inner membrane, cell outer membrane* and *cytoplasm*) have larger positive weights than those for the other five subcellular locations. This suggests that the CC GO terms are indicative of the presence of the query proteins in the first three subcellular locations with higher confidence whereas for the other five subcellular locations, most of the CC GO terms are indicative of being not present in the corresponding subcellular location(s).

5.3. Circular Networks for Essential GO Terms and Subcellular Locations

To gain a comprehensive impression of the relationships between the essential GO terms and the subcellular locations (GO-SCL relationships) for Gram-LocEN, we have drawn two circular networks linking the essential GO terms and subcellular locations, namely Fig. 5(a) and Fig. 5(b), to show the GO-SCL relationships for the Gram-positive and the Gram-negative bacterial datasets, respectively. In both figures, small green dots represent the GO terms and the large dots in different colors represent different subcellular locations. A line connecting an essential GO term and a subcellular location means that the GO term contributes to the prediction of the subcellular location. On the other hand, if there is no connection between an essential GO term and a subcellular location, then this GO term does not provide any information about the presence or absence of a protein in this particular subcellular location.

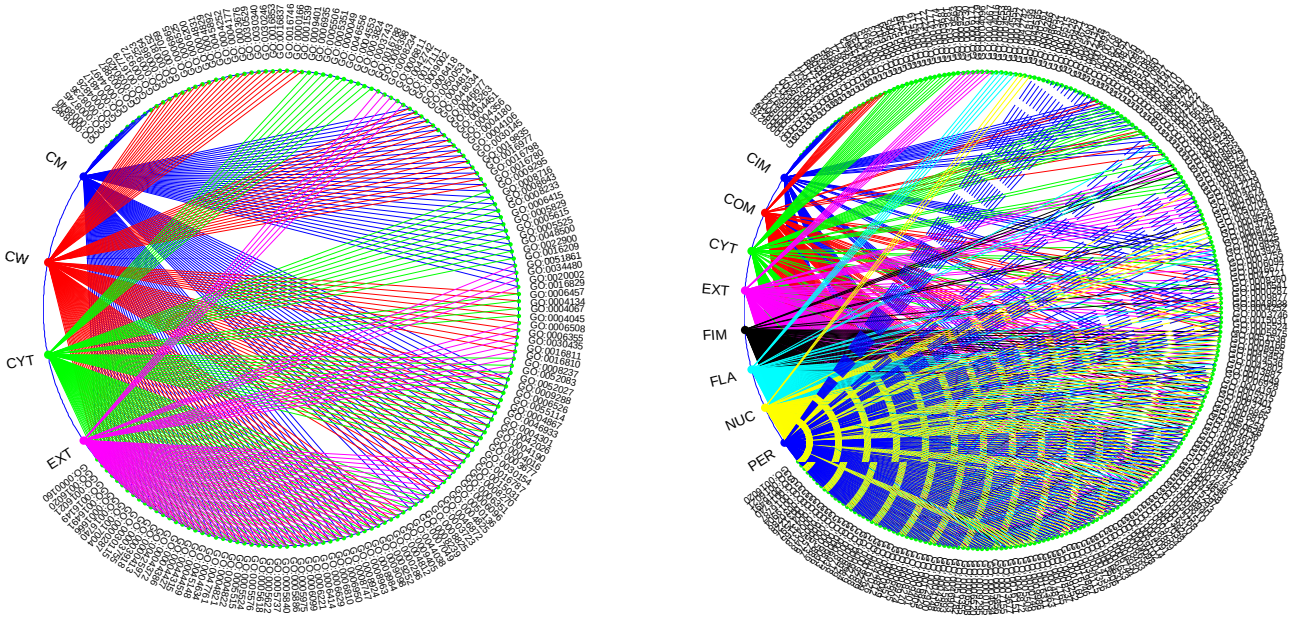
Starting from the top-left green dot to the bottom-left green dot in clockwise direction in Fig. 5, the degree of overlapping among the lines gradually increases, meaning that the number of subcellular locations to which a GO term contributes also gradually increases. For example, in Fig. 5(a), the first 10 GO terms (GO:0008982, GO:0008940, GO:0008745, GO:0008236, GO:0004826, GO:004817, GO:0004497, GO:0003887, GO:0030420 and GO:0016779) are indicative of *cell membrane* only, i.e., suggesting whether a Gram-positive bacterial protein belongs to *cell membrane* or not. Similarly, in Fig. 5(b), the first 13 GO terms can only indicates whether a Gram-negative bacterial protein is located in *cell inner membrane* or not. On the other hand, for the Gram-positive bacterial proteins, GO:0009002 is indicative for both *cell membrane* and *cell wall*; for the Gram-negative bacterial proteins, GO:0006099 contributes to the prediction of both *cell inner membrane* and *cytoplasm*. More aggressively, the last several GO terms, such as GO:0000160, GO:0016020 and GO:0016021, contribute to the prediction of all of the 4 subcellular locations for the Gram-positive bacterial proteins, whereas for the Gram-negative bacterial proteins, the last several GO terms such as GO:0016491, GO:0016021 and GO:0016020 is indicative for all of the 8 subcellular locations. These essential GO terms are indicators of whether a protein resides in one or more subcellular location(s) or not.

For readers' convenience, all the essential GO terms found by Gram-LocEN for both the Gram-positive and Gram-negative bacterial proteins are listed in supplementary materials.

5.4. Comparing with State-of-the-Art Predictors

Table 1 and Table 2 compare the performance of Gram-LocEN against several state-of-the-art multi-label predictors on the Gram-positive bacterial benchmark dataset and the Gram-negative bacterial benchmark dataset, respectively, based on leave-one-out cross-validation (LOOCV). Among them, Gpos-PLoc [38] and Gneg-PLoc [40] can only predict single-location proteins, whereas the other predictors can deal with both single- and multi-location proteins. All of the predictors use some forms of GO vectors as features. From the classification perspective, Gpos-PLoc and Gneg-PLoc use an optimized evidence-theoretic based K-nearest neighbors (OET-KNN) classifier; Gpos-mPLoc [39] and Gneg-mPLoc [41] use a multi-label version of OET-KNN classifier; iLoc-Gpos [44] and iLoc-Gneg [42] use an improved multi-label KNN (ML-KNN) classifier; Gpos-ECC-mPLoc [43] and Gneg-ECC-mPLoc [43] use an ensemble of SVM classifier chains; and the proposed Gram-LocEN uses multi-label EN classifier.

As shown in Table 1, in terms of all performance metrics, Gram-LocEN performs significantly better than the other four



(a) The Gram-Positive Bacterial Dataset

(b) The Gram-Negative Bacterial Dataset

Figure 5: A network showing the relationship between the essential GO terms and each subcellular location for (a) the Gram-positive bacterial dataset and (b) the Gram-negative bacterial dataset. Small green dots represent the GO terms and the large dots in different colors represent the 4 subcellular locations. A line connecting an essential GO term and a subcellular location denotes that the GO term contributes to the prediction of the subcellular location. On the contrary, if there is no line connecting an essential GO term with a particular subcellular location, then this GO term does not provide any information about the presence or absence of a protein in this particular subcellular location. The 4 subcellular locations for the Gram-positive bacterial dataset include: cell membrane (CM), cell wall (CW), cytoplasm (CYT) and extracellular (EXT); the 8 subcellular locations for the Gram-negative bacterial dataset include: cell inner membrane (CIM), cell outer membrane (COM), cytoplasm (CYT), extracellular (EXT), fimbrium (FIM), flagellum (FLA), nucleoid (NUC) and periplasm (PER).

Table 1: Comparing Gram-LocEN with state-of-the-art multi-label predictors using the Gram-positive bacterial dataset based on leave-one-out cross-validation (LOOCV) tests. "–" means the corresponding references do not provide the related metrics.

Label	Subcellular Location	LOOCV Test Locative Accuracy (LA)				
		Gpos-PLoc [38]	Gpos-mPLoc [39]	iLoc-Gpos [44]	Gpos-ECC-mPLoc [43]	Gram-LocEN
1	Cell membrane	–	–	167/174= 0.960	168/174 = 0.965	170/174= 0.977
2	Cell wall	–	–	12/18= 0.667	12/18 = 0.667	17/18= 0.944
3	Cytoplasm	–	–	198/208= 0.952	200/208= 0.962	202/208= 0.971
4	Extracellular	–	–	110/123= 0.894	114/123 = 0.927	117/123= 0.951
Overall Locative Accuracy (*OLA*)		379/523 = 0.725	430/523 = 0.822	487/523 = 0.931	494/523 = 0.944	506/523 = **0.968**
Overall Actual Accuracy (*OAA*)		–	–	482/519 = 0.929	488/519 = 0.940	500/519 = **0.963**
Accuracy		–	–	–	–	**0.968**
Precision		–	–	–	–	**0.971**
Recall		–	–	–	–	**0.970**
F1		–	–	–	–	**0.970**
HL		–	–	–	–	**0.016**

predictors for the Gram-positive bacterial dataset. Specifically, the *OLA* of Gram-LocEN is more than 24% (absolute), 14% (absolute), 3% (absolute) and 2% (absolute) better than that of Gpos-PLoc, Gpos-mPLoc, iLoc-Gpos and Gpos-ECC-mPLoc, respectively; the *OAA* of Gram-LocEN is 3% (absolute) than that of both iLoc-Gpos and Gpos-ECC-mPLoc; the individual locative accuracies of Gram-LocEN are remarkably higher than those of iLoc-Gpos and Gpos-ECC-mPLoc in terms of all of the four subcellular locations. The results suggest that Gram-LocEN performs better than the state-of-the-art classifiers.

Similar conclusions can be drawn from Table 2, in which Gram-LocEN also outperforms the other existing state-of-the-art predictors for the Gram-negative bacterial dataset.

6. Predicting and Interpreting Novel Proteins

To further exemplify how Gram-LocEN predicts and interprets the subcellular localization of proteins, we collected several novel proteins as test proteins (including both Gram-positive and Gram-negative bacterial proteins), which were manually reviewed and were added to Swiss-Prot after September, 2016. The novelty of these proteins can impartially demonstrate the prediction powers and the interpretability of our proposed predictors. Table 3 shows the prediction results of the 6 novel proteins (3 from the Gram-positive bacteria and 3 from the Gram-negative bacteria) by Gram-LocEN. As can be seen, although these proteins are totally unseen to our training dataset (created before 2009), all of them are correctly predicted. The essential GO terms that contribute to the prediction decisions are also presented in Table 3. A comparison between the essential GO terms in Fig. 5 and the last column in Table 3 reveals that not all of the essential GO terms contribute to the final predictions. For example, for the protein D3JTC1, only 3 out of 6 GO terms are useful for determining the subcellular localization. Interestingly, even if two proteins are predicted to the same subcellular location, the essential GO terms for the two proteins are not necessarily the same. For example, for C5C7X8 and P0DOB6, although both of them are correctly predicted to locate in *cytoplasm*, their essential GO terms are different.

Fig. 6 demonstrates how researchers can use Gram-LocEN to interpret the prediction results of query proteins. Fig. 6(a) shows an screenshot of location-specific scores produced by Eq. 7 in descending order using the query protein D3JTC1 (Table 3) as input. Also, the columns "Weight" and "Term-Freq" represent non-zero elements of $\tilde{\boldsymbol{\beta}}_m$ in Eq. 6 and $\mathbf{q}_t^s$ in Eq. 5, and the column "Feature Score" represents the product of Weight and Term-Freq. The higher the feature score, the more contribution is the corresponding GO term to the prediction result. Since only one score is positive, the subcellular location is determined by the positive score(s), which corresponds to *extracellular*. The scores and weights for the essential GO terms in *extracellular* are also shown in Fig. 6(a).[6] As can be seen, only 3 (See Table 3) essential GO terms contribute to the scores corresponding to *extracellular*. Fig. 6(b) shows the case for a Gram-negative bacterial protein (O07838). Similarly, there is only one positive score, suggesting that the query protein is located in a single location, namely *cell inner membrane*.

7. Conclusions

This paper proposes a multi-label elastic-net based classifier, namely Gram-LocEN, for predicting and interpreting subcellular localization of both single- and multi-location proteins of Gram-positive and Gram-negative bacteria. By using one-vs-rest EN classifiers, 162 and 245 out of more than 8,000 GO terms are selected for Gram-positive and Gram-negative bacteria, respectively. Based on these selected essential GO terms, the prediction results can be easily interpreted. The contributions of this paper can be summarized as follows: (1) Gram-LocEN performs impressively better than existing state-of-the-art predcitors for both Gram-positive and Gram-negative bacterial proteins; (2) with the essential GO terms, the predictions made by Gram-LocEN are interpretable; (3) experimental results are consistent with biological annotations, i.e., the key GO terms play greater roles in determining subcellular localization of proteins; and (4) Besides cellular component GO terms, GO terms from the categories of biological processes and molecular functions also contribute to the prediction.

Acknowledgment

This work was in part supported by the RGC of Hong Kong SAR Grant Nos. PolyU152068/15E and PolyU152518/16E.

References

[1] R. A. Harvey, C. C. N., B. D. Fisher, Quick review of clinically important microorganisms, in: R. A. Harvey (Ed.), Lippincott's Illustrated Reviews: Microbiology (Third Edition), Lippincott Williams & Wilkins, Hagerstown, MD, 2013, Ch. 32, pp. 332–353.

[2] A. M. O'Hara, F. Shanahan, The gut flora as a forgotten organ, EMBO reports 7 (7) (2006) 688–693.

[3] D. Z. Rudner, R. Losick, Protein subcellular localization in bacteria, Cold Spring Harbor Perspectives in Biology 2 (4) (2010) a000307.

[4] R. Mott, J. Schultz, P. Bork, C. Ponting, Predicting protein cellular localization using a domain projection method, Genome research 12 (8) (2002) 1168–1174.

[5] M. W. Mak, J. Guo, S. Y. Kung, PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM, IEEE/ACM Trans. on Computational Biology and Bioinformatics 5 (3) (2008) 416 – 422.

[6] O. Emanuelsson, H. Nielsen, S. Brunak, G. von Heijne, Predicting subcellular localization of proteins based on their N-terminal amino acid sequence, J. Mol. Biol. 300 (4) (2000) 1005–1016.

[7] H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, Int. J. Neural Sys. 8 (1997) 581–599.

[6] The scores and weights for the essential GO terms for all of the 4 subcellular locations can be seen by inputing the query protein sequence to our Gram-LocEN web-server by selecting the Gram-positive bacteria as the organism.

Table 2: Comparing Gram-LocEN with state-of-the-art multi-label predictors using the Gram-negative bacterial dataset based on leave-one-out cross-validation (LOOCV) tests. "–" means the corresponding references do not provide the related metrics.

Label	Subcellular Location	LOOCV Test Locative Accuracy (LA)				
		Gneg-PLoc [40]	Gneg-mPLoc [41]	iLoc-Gneg [42]	Gneg-ECC-mPLoc [43]	Gram-LocEN
1	Cell inner membrane	454/557= 0.815	525/557= 0.943	539/557 = 0.968	532/557 = 0.955	541/557 = 0.971
2	Cell outer membrane	68/124= 0.548	105/124= 0.847	103/124 = 0.831	526/557 = 0.944	111/124 = 0.895
3	Cytoplasm	362/410= 0.883	357/410= 0.871	367/410 = 0.895	378/410 = 0.922	379/410 = 0.924
4	Extracellular	59/133= 0.444	79/133= 0.594	115/133 = 0.865	124/133 = 0.932	129/133 = 0.970
5	Fimbrium	11/32= 0.344	28/32= 0.875	30/32 = 0.938	30/32 = 0.938	32/32 = 1.000
6	Flagellum	0/12= 0.000	0/12= 0.000	12/12 = 1.000	12/12 = 1.000	12/12 = 1.000
7	Nucleoid	0/8= 0.000	0/8= 0.000	4/8 = 0.500	7/8 = 0.875	7/8 = 0.875
8	Periplasm	87/180= 0.483	154/180= 0.856	161/180 = 0.894	170/180 = 0.944	161/180 = 0.894
Overall Locative Accuracy (*OLA*)		1041/1456= 0.715	1248/1456= 0.857	1331/1456 = 0.914	1370/1456 = 0.941	1387/1456 = **0.953**
Overall Actual Accuracy (*OAA*)		–	–	1252/1392 = 0.899	1286/1392 = 0.924	1315/1392 = **0.945**
Accuracy		–	–	–	–	**0.931**
Precision		–	–	–	–	**0.952**
Recall		–	–	–	–	**0.944**
F1		–	–	–	–	**0.942**
HL		–	–	–	–	**0.027**

Subcellular Location	Scores	Detailed Info			
Extracellular	1.2157	**Essential GO Terms**	Feature Scores	Weight	Term-Freq
		GO:0005576	1.3224	0.4408	3
		GO:0046872	-0.048	-0.048	1
		GO:0016829	-0.0587	-0.0587	1
Cell-Membrane	-1.1436	**Essential GO Terms**	Feature Scores	Weight	Term-Freq
		GO:0016829	-0.0067	-0.0067	1
		GO:0046872	-0.0272	-0.0272	1
		GO:0005576	-1.1097	-0.3699	3
⋮	⋮	⋮			

(a) Predicting Gram-Positive Bacterial Protein D3JTC1

Subcellular Location	Scores	Detailed Info			
Cell-Inner-Membrane	0.9444	**Essential GO Terms**	Feature Scores	Weight	Term-Freq
		GO:0005886	0.852	0.284	3
		GO:0016021	0.1212	0.0606	2
		GO:0016020	0.0505	0.0505	1
		GO:0006810	-0.0793	-0.0793	1
Flagellum	-0.8992	**Essential GO Terms**	Feature Scores	Weight	Term-Freq
		GO:0006810	-0.0332	-0.0332	1
		GO:0016021	-0.0802	-0.0401	2
		GO:0016020	-0.0955	-0.0955	1
		GO:0005886	-0.6903	-0.2301	3
⋮	⋮	⋮			

(b) Predicting Gram-Negative Bacterial Protein O07838

Figure 6: Screenshots showing how Gram-LocEN predicts and interprets subcellular localization of (a) a Gram-positive bacterial protein (D3JTC1) and (b) a Gram-negative bacterial protein (O07838). *Score*: the score determined in Eq. 7; *Feature Score*: the score that each essential GO term contributes to the final prediction; *Term-freq*: the frequency of occurrence of an essential GO term.

Table 3: Prediction results of 7 novel proteins by Gram-LocEN. *AC*: UniProtKB accession number; *Ground-truth location(s)*: the experimentally-validated actual subcellular location(s); *GO Total Number*: the total number of GO terms retrieved for a given query protein.

Organism	AC	Date of Creation	Ground-truth	Prediction Results	GO Total Number	Essential GO Terms
Gram-Positive	D3JTC1	07-Sep-2016	Extracellular	Extracellular	6	GO:0005576, GO:0046872, GO:0016829
	P0DOB6	05-Oct-2016	Cytoplasm	Cytoplasm	12	GO:0005524, GO:0006096, GO:0000166, GO:0016740, GO:0046872, GO:0005737
	C5C7X8	07-Sep-2016	Cytoplasm	Cytoplasm	13	GO:0005524, GO:0003918, GO:0000287, GO:0003677, GO:0016853, GO:0046677, GO:0000166, GO:0046872, GO:0005737
Gram-Negative	O07838	16-Mar-2016	Cell inner membrane	Cell inner membrane	5	GO:0005886, GO:0016021, GO:0016020, GO:0006810
	Q88RS0	05-Oct-2016	Cytoplasm	Cytoplasm	12	GO:0005737, GO:0016311, GO:0008270, GO:0000287, GO:0009103, GO:0005975, GO:0016787, GO:0046872
	Q128M1	07-Sep-2016	Periplasm	Periplasm	2	GO:0006810, GO:0030288

[8] K. Nakai, M. Kanehisa, Expert system for predicting protein localization sites in gram-negative bacteria, Proteins: Structure, Function, and Genetics 11 (2) (1991) 95–110.

[9] K. C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins: Structure, Function, and Genetics 43 (2001) 246–255.

[10] H. Nakashima, K. Nishikawa, Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies, J. Mol. Biol. 238 (1994) 54–61.

[11] G. P. Zhou, K. Doctor, Subcellular location prediction of apoptosis proteins, Proteins: Structure, Function, and Genetics 50 (2003) 44–48.

[12] G.-L. Fan, Q.-Z. Li, Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition, Journal of Theoretical Biology 304 (2012) 88–95.

[13] K. C. Chou, H. B. Shen, Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers, J. of Proteome Research 5 (2006) 1888–1897.

[14] S. Wan, M. W. Mak, S. Y. Kung, Sparse regressions for predicting and interpreting subcellular localization of multi-label proteins, BMC Bioinformatics 17 (97). doi:10.1186/s12859-016-0940-x.

[15] K. C. Chou, Y. D. Cai, Prediction of protein subcellular locations by GO-FunD-PseAA predictor, Biochem. Biophys. Res. Commun. 320 (2004) 1236–1239.

[16] S. Wan, M. W. Mak, S. Y. Kung, mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction, Analytical Biochemistry 473 (2015) 14–27.

[17] K. C. Chou, Z. C. Wu, X. Xiao, iLoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites, Molecular BioSystems 8 (2012) 629–641.

[18] S. Wan, M. W. Mak, B. Zhang, Y. Wang, S. Y. Kung, Ensemble random projection for multi-label classification with application to protein subcellular localization, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14),, IEEE, 2014, pp. 5999–6003.

[19] S. Wan, M. W. Mak, Machine learning for protein subcellular localization prediction, De Gruyter, 2015.

[20] W. Z. Lin, J. A. Fang, X. Xiao, K. C. Chou, iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, Molecular BioSystems 9 (4) (2013) 634–644.

[21] S. Wan, M. W. Mak, S. Y. Kung, Semantic similarity over gene ontology for multi-label protein subcellular localization, Engineering 5 (2013) 68–72.

[22] S. Mei, Multi-label multi-kernel transfer learning for human protein subcellular localization, PLoS ONE 7 (6) (2012) e37716.

[23] S. Wan, M. W. Mak, Predicting subcellular localization of multi-location proteins by improving support vector machines with an adaptive-decision scheme, International Journal of Machine Learning and Cybernetics to appear (2015) 1–13. doi:10.1007/s13042-015-0460-4.

[24] A. Fyshe, Y. Liu, D. Szafron, R. Greiner, P. Lu, Improving subcellular localization prediction using text classification and the gene ontology, Bioinformatics 24 (2008) 2512–2517.

[25] S. Brady, H. Shatkay, EpiLoc: a (working) text-based system for predicting protein subcellular location, in: Pac. Symp. Biocomput., 2008, pp. 604–615.

[26] R. Nair, B. Rost, Sequence conserved for subcellular localization, Protein Science 11 (2002) 2836–2847.

[27] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, R. Eisner, Predicting subcellular localization of proteins using machine-learned classifiers, Bioinformatics 20 (4) (2004) 547–556.

[28] W. L. Huang, C. W. Tung, S. W. Ho, S. F. Hwang, S. Y. Ho, ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization, BMC Bioinformatics 9 (2008) 80.

[29] S. M. Chi, D. Nam, WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms, Bioinformatics 28 (7) (2012) 1028–1030.
URL `http://bioinformatics.oxfordjournals.org/content/28/7/1028.short`

[30] S. Wan, M. W. Mak, S. Y. Kung, GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition, Journal of Theoretical Biology 323 (2013) 40–48.

[31] R. F. Murphy, communicating subcellular distributions, Cytometry 77 (7) (2010) 686–92.

[32] A. H. Millar, C. Carrie, B. Pogson, J. Whelan, Exploring the function-location nexus: using multiple lines of evidence in defining the subcellular location of plant proteins, Plant Cell 21 (6) (2009) 1625–1631.

[33] S. Zhang, X. F. Xia, J. C. Shen, Y. Zhou, Z. Sun, DBMLoc: A database of proteins with multiple subcellular localizations, BMC Bioinformatics 9 (2008) 127.

[34] L. J. Foster, C. L. D. Hoog, Y. Zhang, Y. Zhang, X. Xie, V. K. Mootha, M. Mann, A mammalian organelle map by protein correlation profiling, Cell 125 (2006) 187–199.

[35] J. C. Mueller, C. Andreoli, H. Prokisch, T. Meitinger, Mechanisms for multiple intracellular localization of human mitochondrial proteins, Mitochondrion 3 (2004) 315–325.

[36] S. Rea, D. James, Moving GLUT4: the biogenesis and trafficking of GLUT4 storage vesicles, Diabetes 46 (1997) 1667–1677.

[37] R. Russell, R. Bergeron, G. Shulman, H. Young, Translocation of myocardial GLUT-4 and increased glucose uptake through activation of AMPK by AICAR, American Journal of Physiology 277 (1997) H643–649.

[38] H. B. Shen, K. Chou, Gpos-PLoc: An ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins, Protein Engineering, Design and Selection 20 (2007) 39–46.

[39] H. B. Shen, K. C. Chou, Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bac-

terial proteins, Protein and Peptide Letters 16 (12) (2009) 1478–1484.
[40] K. C. Chou, H. B. Shen, Large-scale predictions of gram-negative bacterial protein subcellular locations, Journal of Proteome Research 5 (2006) 3420–3428.
[41] H. B. Shen, K. C. Chou, Gneg-mPLoc: A top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins, J. Theor. Biol 264 (2010) 326–333.
[42] X. Xiao, Z. C. Wu, K. C. Chou, A multi-label learning classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites, PLoS ONE 6 (6) (2011) e20592.
[43] X. Wang, J. Zhang, G. Z. Li, Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble, BMC bioinformatics 16 (12) (2015) 1.
[44] Z. C. Wu, X. Xiao, K. C. Chou, iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins, Protein and Peptide Letters 19 (2012) 4–14.
[45] S. Wan, M. W. Mak, S. Y. Kung, R3P-Loc: A compact multi-label predictor using ridge regression and random projection for protein subcellular localization, Journal of Theoretical Biology 360 (2014) 34–45.
[46] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[47] K. Nakai, Protein sorting signals and prediction of subcellular localization, Advances in Protein Chemistry 54 (1) (2000) 277–344.
[48] S. Wan, M. W. Mak, S. Y. Kung, mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines, BMC Bioinformatics 13 (2012) 290.
[49] S. Wan, M. W. Mak, S. Y. Kung, mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization predictor, Journal of Theoretical Biology 382 (2015) 223–234.
[50] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.
[51] B. Zhang, H. Li, R. B. Riggins, M. Zhan, J. Xuan, Z. Zhang, E. P. Hoffman, R. Clarke, Y. Wang, Differential dependency network analysis to identify condition-specific topological changes in biological networks, Bioinformatics 25 (4) (2009) 526–532.
[52] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.
[53] B. J. Marafino, W. J. Boscardin, R. A. Dudley, Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes, Journal of Biomedical Informatics 54 (2015) 114–120.
[54] K. L. Ayers, H. J. Cordell, SNP Selection in genome-wide and candidate gene studies via penalized logistic regression, Genetic Epidemiology 34 (8) (2010) 879–891.
[55] D. He, Z. Wang, L. Parida, Data-driven encoding for quantitative genetic trait prediction, BMC Bioinformatics 16 (Suppl 1) (2015) S10.
[56] Z. C. Wu, X. Xiao, K. C. Chou, iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites, Molecular BioSystems 7 (2011) 3287–3297.
[57] X. Xiao, Z. C. Wu, K. C. Chou, iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites, Journal of Theoretical Biology 284 (2011) 42–51.
[58] S. Wan, M. W. Mak, S. Y. Kung, HybridGO-Loc: Mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins, PLoS ONE 9 (3) (2014) e89545.
[59] K. Dembczynski, W. Waegeman, W. Cheng, E. Hullermeier, On label dependence and loss minimization in multi-label classification, Machine Learning 88 (1-2) (2012) 5–45.
[60] W. Gao, Z. H. Zhou, On the consistency of multi-label learning, in: Proceedings of the 24th Annual Conference on Learning Theory, 2011, pp. 341–358.
[61] T. Hastie, R. Tibshirani, J. Friedman, The element of statistical learning, Springer-Verlag, 2001.