mLASSO-Hum: A LASSO-based interpretable human-protein subcellular localization predictor

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

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

Abstract

Knowing the subcellular compartments of human proteins is essential to shed light on the mechanisms of a broad range of human diseases. In computational methods for protein subcellular localization, knowledge-based methods (especially gene ontology (GO) based methods) are known to perform better than sequence-based methods. However, existing GO-based predictors often lack interpretability and suffer from overfitting due to the high dimensionality of feature vectors. To address these problems, this paper proposes an interpretable multi-label predictor, namely mLASSO-Hum, which can yield sparse and interpretable solutions for large-scale prediction of human protein subcellular localization. By using the one-vs-rest LASSO-based classifiers, 87 out of more than 8,000 GO terms are found to play more significant roles in determining the subcellular localization. Based on these 87 essential GO terms, we can decide not only where a protein resides within a cell, but also why it is located there. To further exploit information from the remaining GO terms, a method based on the GO hierarchical information derived from the depth distance of GO terms is proposed. Experimental results show that mLASSO-Hum performs significantly better than state-of-the-art predictors. We also found that in addition to the GO terms from the cellular component category, GO terms from the other two categories also play important roles in the final classification decisions. For readers' convenience, the mLASSO-Hum server is available online at `http://bioinfo.eie.polyu.edu.hk/mLASSOHumServer/`.

Keywords: Protein subcellular localization; Sparse solutions; Interpretable prediction; Depth-dependent information; Multi-label classification.

1. Introduction

Proteins, which abundantly exist in the human body, exert their biological functions in virtually every process within human cells provided that they are located in the correct spatiotemporal cellular contexts. Knowing the subcellular compartments of *homo sapiens* proteins helps biologists elucidate the functions of proteins and identify drug targets [1]. It is also essential to shed light on the mechanisms of a broad range of human diseases due to protein subcellular mislocalization, such as primary human liver tumors [2], Alzheimer's disease [3], breast cancer [4], pre-eclampsia [5], Bartter syndrome [6] and kidney stone [7]. Conventional high quality localization database such as the Human Protein Atlas[1] are obtained via wet-lab experiments such as electron microscopy, cell fractionation and fluorescent microscopy imaging. These methods, however, are time-consuming, costly and laborious, especially with the advent of the avalanche of newly discovered protein sequences after large-scale sequencing projects. Therefore, computational methods are developed for fast and large-scale protein subcellular localization (PSCL).

Computationally, conventional PSCL approaches are classified as sequence-based and knowledge-based. Sequence-based approaches include: (1) amino-acid composition-based methods [8, 9, 10, 11]; (2) homology-based methods [12, 13, 14]; (3) sorting-signals based methods [15, 16, 17]. Knowledge-based methods use information from knowledge databases, such as Gene Ontology (GO)[2] terms [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28], Swiss-Prot keywords [29, 30], or PubMed abstracts [31, 32]. Among them, GO-based methods have demonstrated to be superior to methods based on other features [33, 34, 23, 35]. Actually, GO-based methods have been widely used in many bioinformatics domains, including enzyme class prediction [36, 37], membrane protein type prediction [38] and protein subcellular localization [39, 40, 41, 42, 34].

Conventional predictors specializing for human proteins, such as HSLPred [43] and Hum-PLoc [44], can only deal with single-location proteins. Recent studies have been focusing on predicting both single- and multi-location proteins due to the prevalence of multi-location proteins [45, 46]. These multi-label proteins are found to play important roles in various metabolic activities in more than one cellular compartment. For example, the glucose transporter GLUT4 has been found in both the intracellular vesicles of adipocytes and the plasma membrane [47]; fatty acid β-oxidation has been found

[*] Corresponding author

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

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

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

in the peroxisome and mitochondria, and antioxidant defense is known to reside in the cytosol, mitochondria and peroxisome [48]. Recently, a review [49] has also demonstrated the importance of multi-label protein subcellular localization on medical science and drug development. As pointed in [49], there are two series of subcellular-localization web-servers, namely 'PLoc' series [50, 51, 52, 53, 54, 55] and 'iLoc' series [56, 57, 58, 27, 59, 60, 19], both of which can predict six or seven species, including eukaryote, human, plant, Gram positive bacteria, Gram negative bacteria, virus and animal.

Recently, several state-of-the-art multi-label predictors have been proposed, such as Hum-mPLoc 2.0 [51], iLoc-Hum [27], mGOASVM [61], HybridGO-Loc [62], R3P-Loc [63], mPLR-Loc [64] and other predictors [65, 66, 67]. They all use the GO information as the features and apply different multi-label classifiers to tackle the multi-label classification problem. However, these GO-based predictors lack interpretability and suffer from overfitting due to the high dimensionality of feature vectors. These predictors can only predict where a query protein is located, but they cannot provide biological reasons on why it resides there. This is possibly a common problem for machine-learning based approaches because it is usually difficult to correlate mathematical mechanism of machine-learning approaches with biological phenomena. As far as we know, there is only one subcellular-location predictor called YLoc [68] that is interpretable. However, YLoc requires heterogeneous biological features such as sorting signals, PROSITE[3] patterns and GO terms, which are not always available for every protein. Moreover, except for R3P-Loc, these predictors use feature vectors with dimensions as high as several thousand. In such high-dimensional space, it is likely that many feature components contain irrelevant or redundant information, causing overfitting problems and thus degrading prediction performance.

To tackle the problems mentioned above, this paper proposes an interpretable multi-label predictor, namely **mLASSO-Hum**, which can yield sparse and interpretable solutions for large-scale prediction of both single-label and multi-label human proteins. Given a query protein sequence, a set of GO terms are retrieved by using the procedures described in [63]. The frequencies of GO occurrences are used to formulate frequency vectors with dimensionality of more than 8000. By using the one-vs-rest LASSO-based (least absolute shrinkage and selection operator-based) classifiers, 87 out of these 8,000+ GO terms are selected. Based on these 87 GO terms, the feature vectors are converted to 87-dim vectors by a novel transferring method based on the depth-dependent GO hierarchical information. Subsequently, the dimension-reduced feature vectors are classified by a multi-label LASSO classifier. Experimental results based on a stringent human benchmark dataset demonstrate that mLASSO-Hum outperforms other existing state-of-the-art predictors. More importantly, based on the selected essential GO terms, users of mLASSO-Hum can not only determine where a protein resides, but also explain why it is located there. In other words, the selected essential GO terms are interpretable for the final prediction results. We also found that in addition to the GO terms from the cellular component category, GO terms from the other two categories also play important roles in the final classification decisions.

As demonstrated by a series of recent publications [69, 70, 71, 72, 73] in compliance with Chou's 5-step rule [74], the establishment of a statistical protein predictor involves the following five steps: (a) construction of a valid dataset for training and testing the predictor; (b) formulation of effective mathematical expressions for converting proteins' characteristics to feature vectors that are relevant to the prediction task; (c) development of classification algorithm for discriminating the feature vectors; (d) evaluation of cross-validation tests for measuring the performance of the predictor; and (e) deployment of a user-friendly, publicly accessible web-server for other researchers to use and validate the prediction method. These steps are further elaborated below.

2. Legitimacy of Using GO Information

In terms of using GO information for PSCL, some researchers may have the following concerns. (1) Because the cellular component GO terms have already been annotated with cellular component categories, can the GO-based methods be replaced by a lookup table using the cellular component GO terms as the keys and the component categories as the hashed values? (2) Are cellular component GO terms the only information for PSCL? (3) Are GO-based methods equivalent to transferring annotations from BLAST homologs? The answers for these concerns are all 'no'. The reasons are given as follows.

1. For the first concern, the GO comprises three orthogonal categories whose terms describe the cellular components, biological processes, and molecular functions of gene products. Some researchers argue that the only thing that needs to be done is to create a lookup table using the cellular component GO terms as the keys and the component categories as the hashed values. Such a naive solution, however, is undesirable and will lead to poor performance, as shown and explained in our previous studies [61, 34].
2. The second concern has been explicitly addressed by [75], who demonstrated that GO terms from the molecular function category are also predictive of subcellular localization, particularly for nucleus, extracellular space, membrane, mitochondrion, endoplasmic reticulum and Golgi apparatus. The in-depth analyses of the correlation between the molecular-function GO terms and localization in Lu and Hunter's study provide an explanation of why GO-based methods outperform sequence-based methods. The results in this paper have also refuted this claim.
3. The third concern is explicitly addressed in our previous study [34], which demonstrates that GO-based methods remarkably outperform methods that only use BLAST and homologous transfer. Besides, [76] also suggest that using BLAST alone is not sufficient for reliable prediction.

[3] http://prosite.expasy.org/

A recent review [77] (in Section VI of [77]) and some other studies [78, 42] also provide strong arguments supporting the legitimacy of using GO information for subcellular localization. In particular, as suggested by [42], the good performance of GO-based methods is due to the fact that the feature vectors in the GO space can better reflect their subcellular locations than those in the Euclidean space or any other simple geometric space.

3. Creation of Compact Databases

A challenge for existing GO-based approaches is that GO information is not always available for every protein. While the GOA database[4] allows us to associate the accession number (AC) of a protein with a set of GO terms, for some novel proteins, neither their ACs nor the ACs of their top homologs have any entries in the GOA database; in other words, no GO terms can be retrieved by their ACs or the ACs of their top homologs. In such case, some predictors use back-up methods that rely on other features such as pseudo-amino-acid composition [9] and sorting signals [79]; some predictors [34, 61] use a successive-search strategy to avoid null GO vectors. However, these strategies may lead to poor performance and increase computation and storage complexity.

To address this problem, similar to our earlier work [63], we created two small yet efficient databases: ProSeq and ProSeq-GO. The former is a sequence database and the latter is a GO-term database, which are extracted from Swiss-Prot and GOA databases, respectively. The procedures of creating these databases are shown in Part (A) of Fig. 1. Detailed descriptions of the procedures can be found in [63]. By using ProSeq and ProSeq-GO, we not only guarantee that every query protein can associate with at least one GO term, but also reduce memory consumption.

For BLAST [80], we use the default parameter setting. With the rapid progress of the Swiss-Prot database, in most cases we can always find a homolog for a query protein. In case no close homolog is found for a query protein, we will increase the E-value until we can find homologs. Because the size of ProSeq database is slightly smaller than that of Swiss-Prot [63], the same reason above also applies to ProSeq. Essentially speaking, we use BLAST as a tool to find 'the most similar' protein sequence from the database, where 'the most similar' protein is determined by the score and the E-value [80]. In the extreme but very rare cases, this most similar protein can be an extremely remote homolog of the query protein. Therefore, we can always find a candidate protein sequence by BLAST. In our experiments, by using the default parameter setting, we can find homologs for all of the proteins in the benchmark dataset detailed in Section 7.1.

4. Construction of Conventional GO-Based Vectors

Constructing conventional GO vectors includes two steps: (1) retrieval of GO terms; and (2) construction of GO vectors. The procedures are shown in Part (B) of Fig. 1. For Step 1, given a query protein, its amino acid sequence is presented to BLAST to find its homologs against the ProSeq database, whose ACs are then used as keys to search against the ProSeq-GO database. Compared to our previous works [61, 34, 62], one of the differences is that instead of using Swiss-Prot and GOA databases, mLASSO-Hum uses ProSeq and ProSeq-GO to retrieve GO terms, which can guarantee that GO terms can always be found for a query protein given its amino acid sequence.

For Step 2, given a dataset, the GO terms of all of its proteins are retrieved by the procedures described above. Similar to our earlier works [34, 61], the GO frequency information is used to construct GO feature vectors. Specifically, the GO vector $\mathbf{q}_i$ of the i-th protein $\mathbb{Q}_i$ is defined as:

$$\mathbf{q}_i = [b_{i,1}, \cdots, b_{i,j}, \cdots, b_{i,W}]^\mathsf{T}, b_{i,j} = \begin{cases} f_{i,j} & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \quad (1)$$

where T is a transpose operator, W is the number of distinct GO terms found for the benchmark dataset (see Section 7.1), and $f_{i,j}$ is the number of occurrences of the j-th GO term (term-frequency) in the i-th protein sequence. Detailed information can be found in [61, 34].

5. Multi-label LASSO

LASSO [81], short for Least Absolute Shrinkage and Selection Operator, is an L_1-regularized linear regression model. It has been applied to many bioinformatics domains, such as gene regulation network analysis [82] and microRNA-target regulatory network construction [83]. The L_1 constraint forces the weights of some features to exactly zero [82], and hence LASSO can automatically selects relevant features. Here we apply LASSO to both feature selection and classification, as shown in Part (C) and (D) of Fig. 1.

5.1. Objective Function of LASSO

Suppose for a two-class single-label problem, we are given a set of training data $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathcal{R}^W$ and $y_i \in \{-1, 1\}$. In our case, $\mathbf{x}_i = \mathbf{q}_i$, where $\mathbf{q}_i$ is defined in Eq. 1. Generally speaking, LASSO imposes an L_1-style regularization to ordinary least squares (OLS). More specifically, LASSO minimizes the empirical L_2-norm loss $l(\boldsymbol{\beta})$:

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

subject to $\sum_{j=1}^{W}|\beta_j| \leq t$, where $t > 0$ is a parameter controlling the shrinkage level to be applied to $\boldsymbol{\beta}$, ε_0 is the bias, $x_{i,j}$ is the j-th element of $\mathbf{x}_i$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_j, \ldots, \beta_W]^\mathsf{T}$ is the LASSO estimate vector to be optimized. Eq. 2 is equivalent to minimizing the following equation:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \varepsilon_0 - \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i)^2 + \lambda\sum_{j=1}^{W}|\beta_j|, \quad (3)$$

[4] http://www.ebi.ac.uk/GOA

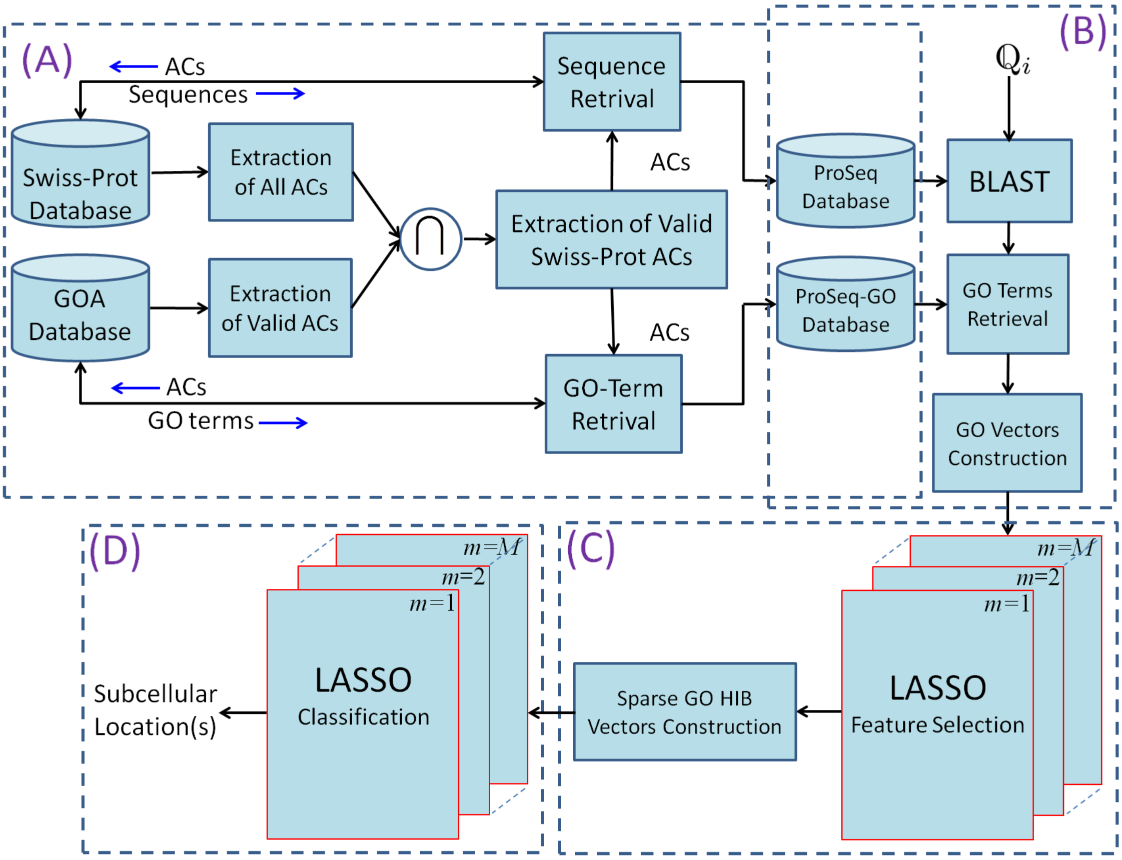


Figure 1: Flowchart of mLASSO-Hum. It consists of four parts: (A) creation of compact databases (ProSeq and ProSeq-GO); (B) Conventional GO-based vectors construction; (C) LASSO-based feature selection and sparse GO hierarchical information based (HIB) vectors construction; and (D) multi-label LASSO classification. $\mathbb{Q}_i$: the i-th query protein.

where $\lambda > 0$ is a penalized parameter to control the degree of regularization. In our experiments, λ was determined by five-fold cross-validation. Eq. 3 is a convex optimization problem, which can be efficiently solved by the least angle regression (LARS) method [84].

5.2. *Multi-label LASSO for Feature Selection*

In an M-class multi-label problem, the training data set is written as $\{\mathbf{x}_i, \mathcal{Y}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathcal{R}^W$ and $\mathcal{Y}_i \subset \{1, 2, \ldots, M\}$ is a set containing one or more labels. M independent binary one-vs-rest LASSOs are trained, one for each class. The labels $\{\mathcal{Y}_i\}_{i=1}^N$ are converted to *transformed labels* [61] $y_{i,m} \in \{-1, 1\}$, where $i = 1, \ldots, N$, and $m = 1, \ldots, M$. Then, the LASSO estimated vector for the m-th class is given by:

$$\hat{\boldsymbol{\beta}}_m = \arg\min_{\boldsymbol{\beta}_m} \left\{ \sum_{i=1}^{N} (y_{i,m} - \varepsilon_{0,m} - \boldsymbol{\beta}_m^{\mathsf{T}} \mathbf{x}_i)^2 + \lambda_m \sum_{j=1}^{W} |\beta_{j,m}| \right\}, \quad (4)$$

where $m = 1, \ldots, M$, $\{y_{i,m}\}_{i=1}^N \in \{-1, 1\}$, $\varepsilon_{0,m}$ and λ_m are the bias and the penalized parameter for the m-th class, respectively. Note that $\hat{\boldsymbol{\beta}}_m$ is solely based on the training data set.

Since L_1 regularization tends to force some of the weights in $\{\beta_{j,m}\}_{j=1}^W$ for the m-th class to exactly zero, LASSO can be used for feature selection. Specifically, the conventional GO vectors obtained from Eq. 1 are used for training multi-label one-vs-rest LASSO classifiers. For an M-class problem (here M is the number of subcellular locations), M independent binary LASSO classifiers are trained, one for each class. After training, the union of those GO terms whose weights are nonzero in any one of the M classes constitutes the selected features. Using LASSO can impressively remove those irrelevant features (or GO terms). Suppose S out of the W weights are nonzero, their corresponding GO terms are called *essential GO terms*. In fact, using the benchmark dataset described in Section 7.1, we found 8110 distinct GO terms from ProSeq-GO. Then, through the proposed multi-label LASSO feature selector, 87 out of 8110 GO terms are selected. Therefore, we have $S = 87$ and $T = 8110$. This means that only around 1% of the GO terms are *essential GO terms* and that the weights for about 99% of the 8110 GO terms are exactly zero.

6. Multi-label LASSO Human Protein Predictor

6.1. *Construction of GO-HIB Vectors*

After feature selection by LASSO, the original W-dim feature vectors become S-dim vectors and the remaining GO terms have been removed. However, because the GO terms in each taxonomy (cellular components, molecular functions or biological processes) are organized within a directed acyclic graph (DAG), it is conducive to make full use of the relations between the essential and the remaining GO terms. Second, because of the structural relationships among the essential GO terms, the essential GO terms are not independent with each other, and hence taking the hierarchical relationships of GO terms into consideration is helpful for the prediction. More importantly, it is likely that some novel proteins associate with non-essential GO terms only; therefore, the relationships between essential and non-essential GO terms should also be considered.

These properties of GO terms inspire us to develop a feature extraction method that makes use of the depth-dependent GO hierarchical information. Specifically, given the t-th query protein $\mathbb{Q}_t$, then its hierarchical information based (HIB) feature vector is constructed as follows:

$$\mathbf{q}_t^{HIB} = [c_{t,1}, \ldots, c_{t,s}, \ldots, c_{t,S}]^\mathsf{T}, \tag{5}$$

where S is the number of essential GO terms obtained in Section 5.2 and

$$c_{t,s} = \sum_{k \in \mathcal{K}_t} f_{t,k}[d_{s,k} = 0] + \max_{k \in \mathcal{K}_t} \left(\frac{f_{t,k}}{2^{d_{s,k}}} \right) [d_{s,k} \neq 0], \tag{6}$$

where $[\cdot]$ is the Iverson bracket, i.e., $[P] = 1$ if P is true and $[P] = 0$ otherwise. In Eq. 6, $\mathcal{K}_t$ is the set of distinct GO terms associated with $\mathbb{Q}_t$ obtained by the procedures detailed in Section 4, $f_{t,k}$ is the number of occurrences of the k-th GO term in the GO-term set $\mathcal{K}_t$, and $d_{s,k}$ is defined as:

$$d_{s,k} = dep(GO_s) - dep(LCA(GO_s, GO_k)), \tag{7}$$

where GO_s and GO_k represent the s-th essential GO term selected in Eq. 4 and the k-th GO term in the GO-term set $\mathcal{K}_t$, respectively; $LCA(GO_s, GO_k)$ is the lowest common ancestor (LCA) of GO_s and GO_k; $dep(GO_s)$ is the depth level of the GO term GO_s and $dep(LCA(GO_s, GO_k))$ is the depth level of the LCA of GO_s and GO_k. The basic properties of GO depth include (1) the depth of the root GO term for each taxonomy is 1, (2) the depth of child terms is larger than their ancestors, and (3) the more specific a GO term, the larger its depth.

Fig. 2 illustrates the working principle of Eq. 6 and Eq. 7. For ease of reference, we label the GO terms from GO:0044699 to GO:0006082 as A to F, as shown in Fig. 2. Here, GO term C (GO:0044281, small molecule metabolic process), which belongs to biological processes, is an essential GO term and its depth is 4 (suppose the depth of the root biological process is 1). Assume that there is only one GO term, GO term F (GO:0006082), associated with the t-th protein and this GO term appears 3 times. Because GO term F is a child term of GO term C, the LCA of these two terms is GO term C itself. Then, according to Eq. 7, $d_{s,k} = 0$; and thus $c_{t,s} = 3 + 0 = 3$ in Eq. 6. On the other hand, if we only found one GO term, GO term D (GO:0044763), for the t-th query protein with occurrence of frequency 3. From Fig. 2, we can see that the LCA of GO term D and GO term C is GO term A (GO:0044699), whose depth is 2. Then, according to Eq. 7, $d_{s,k} = 4 - 2 = 2$; and thus $c_{t,s} = 0 + 3/2^2 = 0.75$. In this case, we can see that the $c_{t,s}$ is smaller for GO term D than for GO term F, despite with the same frequency occurrences. This is also consistent with our observation that GO term F has a closer relationship with GO term C than with GO term D.

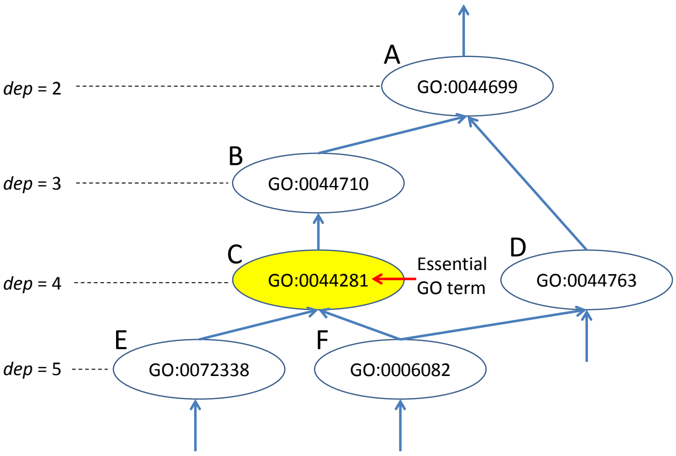


Figure 2: An example demonstrating the working principle of Eq. 6 and Eq. 7. *dep*: the depth of the corresponding GO term. Here, GO:0044281 (in yellow) is the essential GO term. The blue arrows represent the 'is-a' relationship. For ease of reference, we label the GO terms from GO:0044699 to GO:0006082 as A to F.

The rationale for Eq. 6 and Eq. 7 is that if the k-th GO term of $\mathbb{Q}_t$ is a child term of the s-th essential GO term, the former should be regarded as equivalent to the latter, because the former contributes as significantly as the latter to the predictions. In this case, $LCA(GO_s, GO_k)$ in Eq. 7 is GO_s and thus $d_{s,k} = 0$ in Eq. 6. Therefore, only the first term of Eq. 6 has contribution to $c_{t,s}$. If the k-th GO term of $\mathbb{Q}_t$ is not a child term of the s-th essential GO term, $LCA(GO_s, GO_k)$ in Eq. 7 is not GO_s and thus $d_{s,k} \neq 0$ in Eq. 6. Thus, only the second term of Eq. 6 has contribution. The contribution of the k-th GO term to the prediction will diminish exponentially fast when the depth distance between the two GO terms increases. Since our previous studies [61, 34] have demonstrated the superiority of term-frequency, we incorporate this information ($f_{t,k}$) in Eq. 6.

Because those essential GO terms and their child terms (in this case $d_{s,k} = 0$) play far more important roles than the remaining GO terms, to highlight the significance of the former and suppress that of the latter, we fully count the frequency of the former whereas only selects the maximum weighted frequency of the latter. By accumulating the contributions of every GO term of the query protein to each essential GO term, the depth-dependent GO hierarchical information is incorporated in our new feature vectors represented in Eq. 5.

6.2. Multi-label LASSO for Classification

Besides feature selection, LASSO can also be used for classification. Compared to using LASSO for feature selection, one of the differences is that we use a new method introduced in Section 6.1 to generate the feature vectors for training multi-label one-vs-rest LASSO classifiers. Specifically, given the t-th query protein $\mathbb{Q}_t$, the feature vector $\mathbf{x}_t^{HIB} \in \mathcal{R}^S$ is defined in Eq. 5, where $S < T$ is the number of essential GO terms. Similarly, for an M-class problem (here M is the number of subcellular locations), M independent binary LASSO classifiers are trained, one for each class. Then, the score of the m-th LASSO is:

$$s_m(\mathbb{Q}_t) = \tilde{\beta}_m^{\mathsf{T}} \mathbf{x}_t^{HIB}, \tag{8}$$

where $\tilde{\beta}_m$ is given by Eq. 4 with $\mathbf{x}_i$ replaced by $\mathbf{x}_i^{HIB}$ and with W replaced by S.

To predict the subcellular locations of datasets containing both single-label and multi-label proteins, a decision scheme for multi-label LASSO classifiers should be used. In this paper, we used the decision scheme described in mGOASVM [61]. In this scheme, the predicted subcellular location(s) of the i-th query protein are given by:

$$\mathcal{M}^*(\mathbb{Q}_t) = \begin{cases} \bigcup_{m=1}^{M} \{m : s_m(\mathbb{Q}_t) > 0\}, \text{ where } \exists\ s_m(\mathbb{Q}_t) > 0\ ; \\ \\ \arg\max_{m=1}^{M}\ s_m(\mathbb{Q}_t), \text{ otherwise.} \end{cases} \tag{9}$$

For ease of presentation, we refer to the proposed predictor as mLASSO-Hum. The flowchart of mLASSO-Hum is shown in Fig. 1.

7. Experiments

7.1. Datasets

In this paper, a recent human benchmark dataset [51] was used to evaluate the performance of mLASSO-Hum. The human dataset was created from Swiss-Prot 55.3, which is a publicly accessible protein database[5]. This benchmark dataset is downloadable from the hyperlink in the mLASSO-Hum server. The human dataset contains 3106 human proteins distributed in 14 locations. Of the 3106 proteins, 2580 belong to one subcellular location, 480 belong to two locations, 43 belong to three locations, 3 belong to four locations and none to five or more locations. This means that the number of locative proteins [27, 61] is ($2580\times1+480\times2+43\times3+3\times4+\sum_{m=5}^{14} 0\times m = 3681$). These locative proteins are distributed as follows: 77 in centrosome, 817 in cytoplasm, 79 in cytoskeleton, 229 in endoplasmic reticulum, 24 in endosome, 385 in extracellular, 161 in Golgi apparatus, 77 in lysosome, 24 in microsome, 364 in mitochondrion, 1021 in nucleus, 47 in peroxisome, 354 in plasma membrane and 22 in synapse. The sequence identity of the dataset was cut off at 25%.

The breakdown of the human dataset is listed in Fig. 3. As can be seen, the majority (79.9%) of the human proteins are located in cytoplasm, nucleus, extracellular, mitochondrion and plasma membrane while proteins located in the rest 9 subcellular locations totally account only around 20%. This means that the dataset is multi-class distributed and imbalanced.

7.2. Performance Metrics

Compared to traditional single-label classification, multi-label classification requires more sophisticated performance metrics to better reflect the multi-label capabilities of classifiers. These measures include *Accuracy, Precision, Recall, F1-score (F1)* and *Hamming Loss (HL)*. 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.[6]

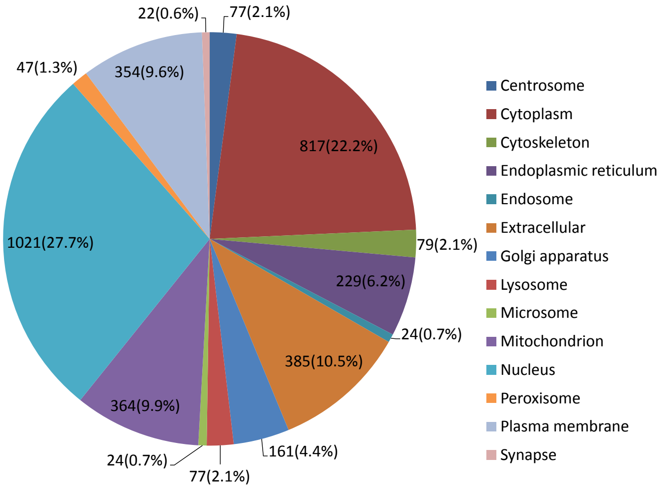


Figure 3: Breakdown of the human dataset. The number of proteins shown in each subcellular location represents the number of 'locative proteins' [27, 61]. Here, 3106 actual proteins have 3681 locative proteins. The plant proteins are distributed in 14 subcellular locations, including centrosome, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracellular, Golgi apparatus, lysosome, microsome, mitochondrion, nucleus, peroxisome, plasma membrane and synapse.

Then the five measurements are defined as follows:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|}{|\mathcal{M}(\mathbb{Q}_i)\cup\mathcal{L}(\mathbb{Q}_i)|}\right) \tag{10}$$

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|}{|\mathcal{M}(\mathbb{Q}_i)|}\right) \tag{11}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|}{|\mathcal{L}(\mathbb{Q}_i)|}\right) \tag{12}$$

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{2|\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|}{|\mathcal{M}(\mathbb{Q}_i)|+|\mathcal{L}(\mathbb{Q}_i)|}\right) \tag{13}$$

$$HL = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|\mathcal{M}(\mathbb{Q}_i)\cup\mathcal{L}(\mathbb{Q}_i)|-|\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|}{M}\right) \tag{14}$$

where $|\cdot|$ means counting the number of elements in the set therein and $\cap$ represents the intersection of sets. An intuitive description of multi-label metrics can also be found in Eq. 16 of [77].

As can be seen from Eq. 14, when all of the proteins are correctly predicted, i.e., $|\mathcal{M}(\mathbb{Q}_i)\cup\mathcal{L}(\mathbb{Q}_i)| = |\mathcal{M}(\mathbb{Q}_i)\cap\mathcal{L}(\mathbb{Q}_i)|$ ($i = 1, \ldots, N$), then $HL = 0$; whereas, other metrics will be equal to 1. On the other hand, when the predictions of all proteins are completely wrong, i.e., $|\mathcal{M}(\mathbb{Q}_i)\cup\mathcal{L}(\mathbb{Q}_i)| = M$ and $|\mathcal{M}(\mathbb{Q}_i)\cap$

[5] http://www.uniprot.org/

[6] In our case, $N = 3106$ for the human dataset.

$\mathcal{L}(\mathbb{Q}_i)|= 0$, then $HL = 1$; whereas, other metrics will be equal to 0.

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)* [85, 86] is different from other metrics. The lower the *HL*, the better the prediction performance.

Two additional measurements [27, 61] are often used in multi-label subcellular localization prediction. They are overall locative accuracy (*OLA*) and overall actual accuracy (*OAA*). Then, *OLA* is given by:

$$OLA = \frac{1}{\sum_{i=1}^{N}|\mathcal{L}(\mathbb{Q}_i)|}\sum_{i=1}^{N}|\mathcal{M}(\mathbb{Q}_i) \cap \mathcal{L}(\mathbb{Q}_i)|, \tag{15}$$

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

$$OAA = \frac{1}{N}\sum_{i=1}^{N}\Delta[\mathcal{M}(\mathbb{Q}_i), \mathcal{L}(\mathbb{Q}_i)] \tag{16}$$

where

$$\Delta[\mathcal{M}(\mathbb{Q}_i), \mathcal{L}(\mathbb{Q}_i)] = \begin{cases} 1 & \text{, if } \mathcal{M}(\mathbb{Q}_i) = \mathcal{L}(\mathbb{Q}_i) \\ 0 & \text{, otherwise.} \end{cases} \tag{17}$$

According to Eq. 15, a locative protein is considered to be correctly predicted if any of the predicted labels matches any labels in the true label set. On the other hand, Eq. 16 suggests that an actual protein is considered to be correctly predicted only if *all* of the predicted labels match those in the true label set exactly. For example, for a protein coexist in, say, three subcellular locations, if only two of the three are correctly predicted, or the predicted result contains a location not belonging to the three, the prediction is considered to be incorrect. In other words, when and only when all the subcellular locations of a query protein are exactly predicted without any overprediction or underprediction, can the prediction be considered as correct. Therefore, *OAA* is a more stringent measure as compared to *OLA*. *OAA* is also more objective than *OLA*. This is because locative accuracy is liable to give biased performance measure when the predictor tends to over-predict, i.e., giving large $|\mathcal{M}(\mathbb{Q}_i)|$ for many $\mathbb{Q}_i$. In the extreme case, if every protein is predicted to have all of the M subcellular locations, according to Eq. 15, the *OLA* is 100%. But obviously, the predictions are wrong and meaningless. On the contrary, *OAA* is 0% in this extreme case, which definitely reflects the real 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 predict, the numerators of the other five measures (including *Accuracy, Precision, Recall, F1* and *OLA*) are non-zero, whereas the numerator of *OAA* in Eq. 16 is 0 (thus contribute nothing to the frequency count).

In statistical prediction, leave-one-out cross validation (LOOCV) is considered to be the most rigorous and bias-free method [87]. Hence, LOOCV was used to examine the performance of mLASSO-Hum.

8. Results and Discussions

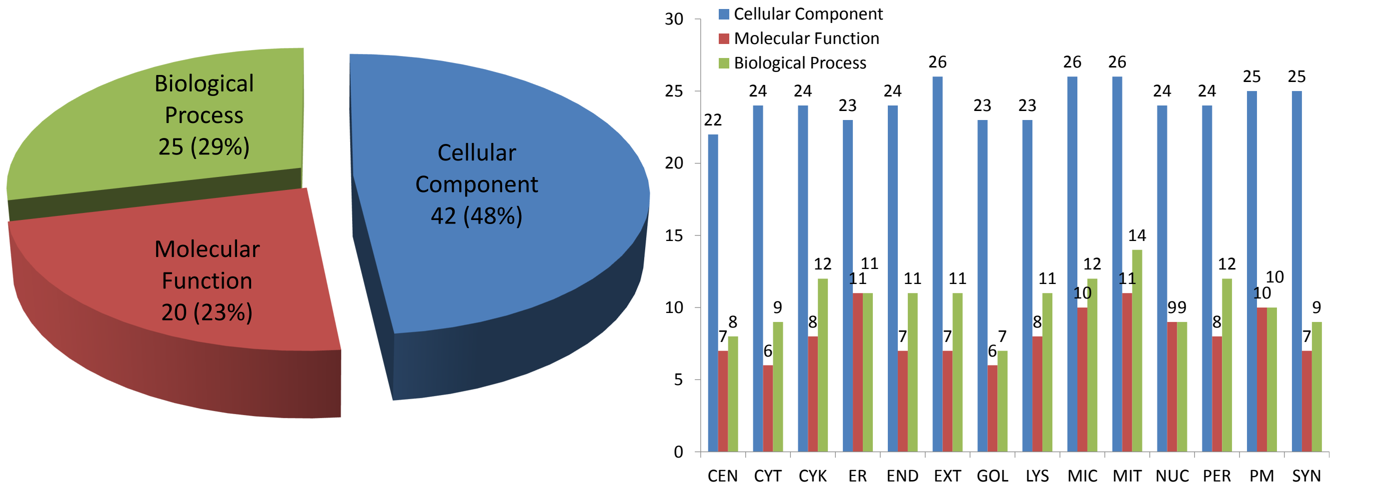
8.1. Statistical Analysis of the Essential GO Terms

Fig. 4(a) shows the statistics of the 87 essential GO terms selected by mLASSO-Hum. As can be seen, among the 87 essential GO terms, about half (42) of them belong to the cellular component category, 25 belong to biological process and the remaining 20 belong to molecular function. This suggests that not only GO terms from cellular components contributes to the prediction of mLASSO-Hum, those from the other categories also play important roles in determining the subcellular localization of proteins.

Fig. 4(b) shows the the categorical breakdown of essential GO terms in each subcellular location. As can be seen, for a particular subcellular location, among the 87 unique essential GO terms, only around 30~50 essential GO terms simultaneously determine where a protein resides. For example, for centrosome, 37 (= 22 + 7 + 8) essential GO terms are contributive to the final decisions, while the remaining 50 GO terms have no roles in determining whether a protein belongs to centrosome or not. Besides, a large portion (around half) of the essential GO terms belong to cellular components, e.g., 22 out of 37 in centrosome, 24 out of 39 in cytoplasm, etc. This is also consistent with the percentage in the statistics of overall essential GO terms shown in Fig. 4(a). The results indicate that cellular component GO terms contribute more to the final prediction than those GO terms from the other two categories. Another observation is that essential GO terms for one subcellular location overlap with those for another subcellular location. This is because the total sum of the essential GO terms of all the subcellular locations is much larger than 87, the total number of unique essential GO terms.

8.2. Comprehensive Networks between Essential GO Terms and SCLs

To understand the relationship between the essential GO terms and the subcellular locations, we have drawn a network connecting the 14 subcellular locations and the 87 essential GO terms in Fig. 5. Small green dots represent the GO terms and the large dots in different colors represent the 14 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. For example, the first 7 GO terms (GO:0007275, GO:0006915, GO:0006355, GO:0005643, GO:0005524, GO:0048471 and GO:0004674) are only contributive to *cytoplasm*, indicating whether a protein belongs to *cytoplasm* or not; GO:0005509 can only indicates whether a protein is located in *endoplasmic reticulum* or not. On the other hand, GO:0005815 is indicative for both *centrosome* and *cytoskeleton*; GO:0005635 contributes to the prediction of both *cytoplasm* and *nucleus*. More aggressively, the



(a) Statistics of essential GO terms

(b) Breakdown of the essential GO terms

Figure 4: Information of the essential GO terms for the human dataset, including (a) the statistics of essential GO terms and (b) the categorical breakdown of the essential GO terms in each subcellular location. *CEN*: centrosome; *CYT*: cytoplasm; *CYK*: cytoskeleton; *ER*: endoplasmic reticulum; *END*: endosome; *EXT*: extracellular; *GOL*: Golgi apparatus; *LYS*: lysosome; *MIC*: microsome; *MIT*: mitochondrion; *NUC*: nucleus; *PER*: peroxisome; *PM*: plasma membrane; *SYN*: synapse.

last several GO terms, such as GO:0016787, GO:0046872 and GO:0005515, may contribute to the prediction of all of the 14 subcellular locations. On the contrary, if there is no line connecting an essential GO term with a particular subcellular location, then this GO term cannot indicate any information about the presence or absence of a protein in this particular subcellular location.

In summary, these essential GO terms are indicators of whether a protein resides in one or more subcellular location(s) or not.

8.3. Location-Specific Significance of Essential GO Terms

To quantitatively demonstrate how and to what extent essential GO terms contribute to the prediction of subcellular locations, Fig. 6 shows the distributions of the non-zero weights in $\{\beta_{s,1}\}_{s\in\mathcal{S}}$ defined in Eq. 4 for the essential GO terms in *centrosome*, where $\mathcal{S}$ is a set of indexes corresponding to non-zero weights. For simplicity, $\beta_{s,1}$ is abbreviated as β in the figure. Specifically, in Fig. 6, there are 37 GO terms whose weights (β) are nonzero, among which, there are 5 positive weights and 32 negative weights. This means that a majority of the weights are negative.

According to Eq. 8 and Eq. 9, the presence of a positive weight provides a piece of evidence that the query protein locates in the corresponding subcellular location; on the contrary, the presence of a negative weight adds a piece of evidence that the query protein does not belongs to the particular subcellular location. This suggests that the presence of one or more of GO:0008543, GO:0005815, GO:0005813, GO:0005515 and GO:0016605 indicates that the query protein belongs to *centrosome*; on the other hand, the presence of any of the other 32 GO terms shown in Fig. 6 suggests that the query protein does **not** belong to *centrosome*. Moreover, the larger the β, the higher the confidence the query protein belongs to or does not belong to a particular subcellular location. For example, both GO:0005815 and GO:0016605 are indicative of *centrosome*; however, the former provide higher confidence than the latter for the indication. On the contrary, the presence of GO:0016020 provides stronger evidence than that of GO:0005829 for concluding that the query protein does not belong to *centrosome*. Similar conclusions can be drawn for the remaining 13 subcellular locations in Fig. S1(b)∼Fig. S1(n).

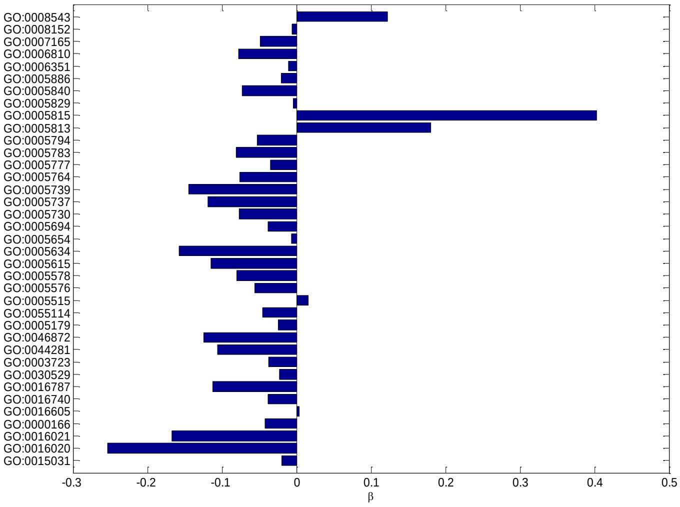


Figure 6: The distributions of non-zero weights $\{\beta_{s,1}\}_{s\in\mathcal{S}}$ defined in Section 6.2 for the essential GO terms in *centrosome*, where $\mathcal{S}$ is a set of indexes corresponding to non-zero weights. Figures for the rest 13 subcellular locations can be found in supplementary materials Fig. S1(b)∼Fig. S1(n) in supplementary materials. For simplicity, $\beta_{s,1}$ is abbreviated as β in the figures.

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

Table 1 compares the performance of mLASSO-Hum against several state-of-the-art multi-label predictors on the human

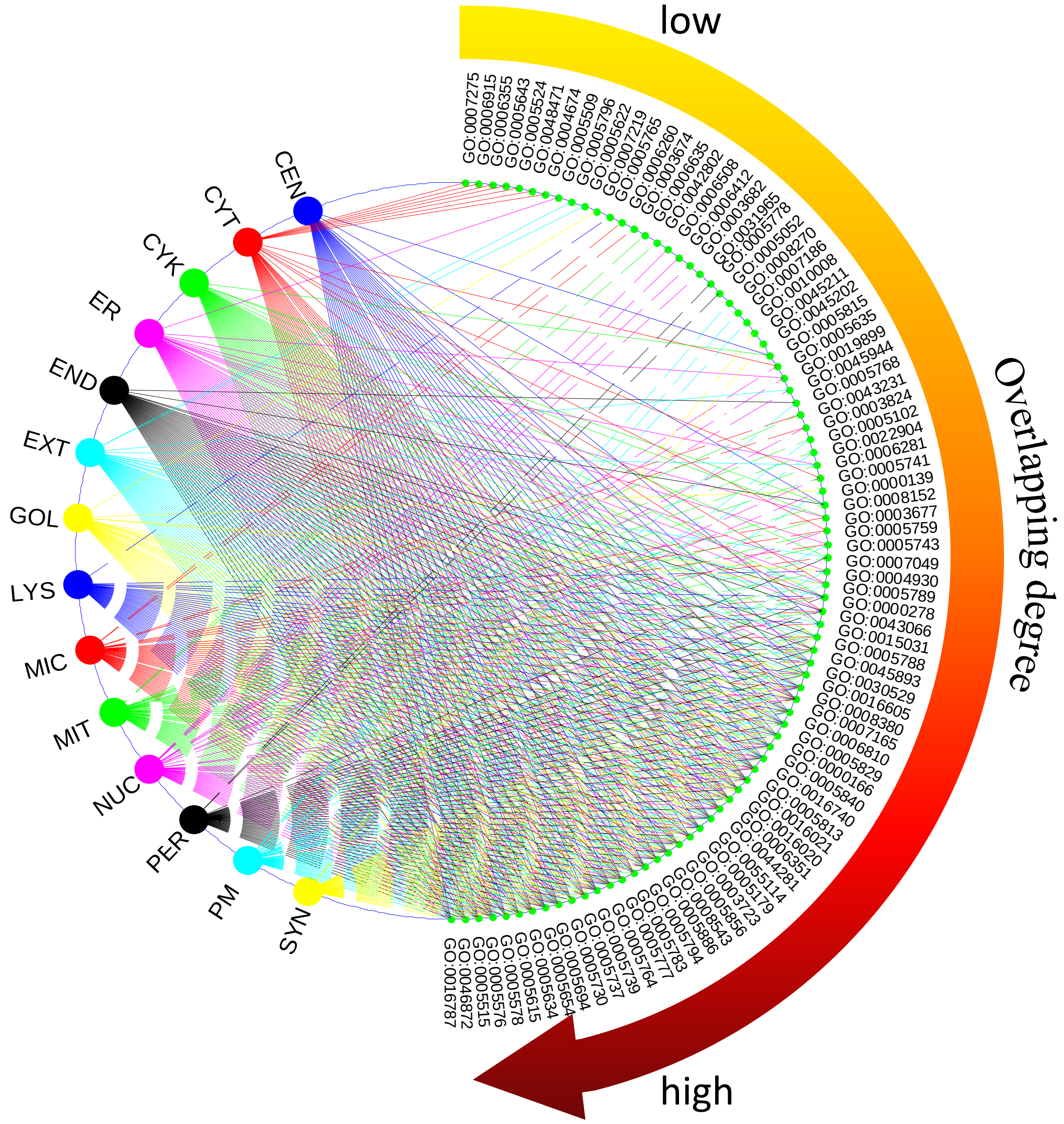


Figure 5: A network showing the relationship between the essential GO terms and each subcellular location. Small green dots on the right represent the GO terms and the large dots in different colors on the left represent the 14 subcellular locations. A line connecting an essential GO term and a subcellular 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 cannot indicate any information about the presence or absence of a protein in this particular subcellular location. Starting from the top green dot to the bottom green dot in clockwise direction, the degree of overlapping among the lines gradually changes from low to high, suggesting that the number of subcellular locations to which an essential GO term contributes changes from small to large. See the caption of Fig. 4 for the acronyms of subcellular locations.

Table 1: Comparing mLASSO-Hum with state-of-the-art multi-label predictors based on leave-one-out cross-validation on the human dataset. "–" means the corresponding references do not provide the related metrics. Note that *OAA* is the most stringent and objective among all the metrics.

Label	Subcellular Location	LOOCV Locative Accuracy (LA)		
		iLoc-Hum [58]	mGOASVM [61]	mLASSO-Hum
1	Centrosome	56/77 = 0.727	64/77 = 0.831	56/77 = 0.727
2	Cytoplasm	561/817 = 0.687	683/817 = 0.836	703/817 = 0.861
3	Cytoskeleton	27/79 = 0.342	44/79 = 0.557	31/79 = 0.392
4	Endoplasmic reticulum	166/229 = 0.725	193/229 = 0.843	188/229 = 0.821
5	Endosome	1/24 = 0.042	9/24 = 0.375	3/24 = 0.125
6	Extracellular	325/385 = 0.844	344/385 = 0.894	327/385 = 0.849
7	Golgi apparatus	99/161 = 0.615	131/161 = 0.814	133/161 = 0.826
8	Lysosome	56/77 = 0.727	71/77 = 0.922	73/77 = 0.948
9	Microsome	7/24 = 0.292	18/24 = 0.750	1/24 = 0.042
10	Mitochondrion	284/364 = 0.780	339/364 = 0.931	343/364 = 0.942
11	Nucleus	918/1021 = 0.899	931/1021 = 0.912	929/1021 = 0.910
12	Peroxisome	20/47 = 0.426	43/47 = 0.915	41/47 = 0.872
13	Plasma membrane	277/354 = 0.783	288/354 = 0.814	282/354 = 0.797
14	Synapse	12/22 = 0.546	12/22 = 0.546	4/22 = 0.182
Overall Actual Accuracy (*OAA*)		2118/3106 = 0.682	2251/3106 = 0.725	2324/3106 = **0.748**
Overall Locative Accuracy (*OLA*)		2809/3681 = 0.763	3170/3681 = **0.861**	3114/3681 = 0.846
Accuracy		–	0.821	**0.833**
Precision		–	0.851	**0.874**
Recall		–	**0.888**	0.879
F1		–	0.853	**0.862**
HL		–	0.029	**0.027**

benchmark dataset. To the best of our knowledge, iLoc-Hum [27] is the best state-of-the-art predictor specializing for predicting multi-label human protein subcellular localization. Because mGOASVM [61] is not trained for predicting human proteins, we retrained it on the human dataset so that the results can be compared with those obtained from mLASSO-Hum. All of the predictors use the information of GO terms as features. From the classification perspective, iLoc-Hum use a multi-label KNN classifier; mGOASVM [61] uses a multi-label SVM classifier; and the proposed mLASSO-Hum uses a multi-label LASSO classifier.

As shown in Table 1, mLASSO-Hum performs significantly better than iLoc-Hum. The *OLA* and *OAA* of mLASSO-Hum are 8% (absolute) and 6% higher than those of iLoc-Hum, respectively. When comparing with mGOASVM, the *OAA* of mLASSO-Hum is more than 2% (absolute) higher than that of mGOASVM, although it is a bit lower than that of mGOASVM in terms of *OLA* and *Recall*. In terms of *Accuracy, Precision, F1* and *HL*, mLASSO-Hum performs better than mGOASVM. The results suggest that the proposed mLASSO-Hum performs better than the state-of-the-art predictors. The individual locative accuracies of mLASSO-Hum are remarkably higher than that of iLoc-Hum, and are comparable to mGOASVM.

9. Discussion

We observe that among the essential GO terms, some GO terms have much larger absolute weights (i.e. $|\beta_{s,m}|$) than the rest, suggesting that they play more significant roles in making the predictions. Specifically, if the weight of an essential GO term for a particular subcellular location is larger than a certain positive threshold, it has high confidence to indicate that the query protein resides in this subcellular location; on the contrary, if the weight is smaller than a certain negative threshold, it has high confidence to indicate that the query protein does not belong to the corresponding subcellular location. We refer the former GO terms and the latter GO terms to as *significantly essential positive GO terms (SEPos GO terms)* and *significantly essential negative GO terms (SENeg GO terms)*, respectively.

Lists of the SEPos GO terms and SENeg GO terms for all of the 14 human subcellular locations can be found in Table 2. For ease of comparison, we have also listed the *key GO terms*, which are defined as those GO terms whose names are exactly the same as the names of subcellular locations according to the GO annotations. Note that there are no key GO terms for *extracellular* and *microsome*. As can be seen, there are around 1~3 SEPos GO terms for each subcellular location, whose presences indicate high confidence of residing in the corresponding subcellular location. More importantly, Table 2 shows that the SEPos GO term(s) selected by mLASSO-Hum for a particular subcellular location incorporate the corresponding key GO term

Table 2: Significantly essential positive and negative GO terms for the 14 human subcellular locations. *Key GO terms* are those GO terms whose names are exactly the same as the names of subcellular locations according to the GO annotations; *SEPos GO terms*: significantly essential positive GO terms, whose weights are larger than 0.1; *SENeg GO terms*: significantly essential negative GO terms, whose weights are smaller than −0.1. "–" means that there is no GO term whose name is exactly the same as the name of the corresponding subcellular location in the GO annotations.

Label	Subcellular Location	Key GO Terms	SEPos GO Terms	SENeg GO Terms
1	Centrosome	GO:0005813	GO:0005813, GO:0005815	GO:0016020, GO:0016021, GO:0016787, GO:0044281, GO:0046872, GO:0005615 GO:0005634 , GO:0005737 , GO:0005739
2	Cytoplasm	GO:0005737	GO:0005737	GO:0016020,GO:0016021,GO:0005615, GO:0005634 , GO:0005739,GO:0005764, GO:0005783,GO:0005794,GO:0005813, GO:0005856
3	Cytoskeleton	GO:0005856	GO:0005856	GO:0016020,GO:0016021,GO:0016787, GO:0044281,GO:0046872,GO:0005615, GO:0005634,GO:0005737,GO:0005739
4	Endoplasmic reticulum	GO:0005783	GO:0005783, GO:0005789	GO:0016020,GO:0046872,GO:0005578 GO:0005615,GO:0005634,GO:0005730, GO:0005737,GO:0005739,GO:0005764, GO:0005856
5	Endosome	GO:0005768	GO:0005768	GO:0016020,GO:0016021,GO:0016787, GO:0044281,GO:0046872,GO:0005578, GO:0005615,GO:0005634,GO:0005737, GO:0005739,GO:0005856
6	Extracellular	—	GO:0005578, GO:0005615	GO:0015031,GO:0016020,GO:0016021, GO:0044281,GO:0005634,GO:0005737, GO:0005739,GO:0005856,GO:0008543
7	Golgi apparatus	GO:0005794	GO:0000139, GO:0005794	GO:0016020,GO:0016021,GO:0044281, GO:0046872,GO:0005615,GO:0005634, GO:0005737,GO:0005739,GO:0005764, GO:0005783,GO:0005856
8	Lysosome	GO:0005764	GO:0005764, GO:0005765	GO:0016020,GO:0016021,GO:0046872, GO:0005615,GO:0005634,GO:0005737, GO:0005739,GO:0005783,GO:0005856
9	Microsome	—	GO:0043231	GO:0016020,GO:0016021,GO:0016787, GO:0046872,GO:0005578,GO:0005615, GO:0005634,GO:0005737,GO:0005739, GO:0005856
10	Mitochondrion	GO:0005739	GO:0005739	GO:0016020,GO:0016021,GO:0016787, GO:0005615,GO:0005634,GO:0005737, GO:0005783,GO:000585
11	Nucleus	GO:0005634	GO:0031965, GO:0005634, GO:0005730	GO:0016020,GO:0016021,GO:0005615, GO:0005737,GO:0005739,GO:0005783, GO:0005840
12	Peroxisome	GO:0005777	GO:0005777	GO:0016020,GO:0016021,GO:0016787, GO:0046872,GO:0005615,GO:0005634, GO:0005737,GO:0005739,GO:0005856
13	Plasma membrane	GO:0005886	GO:0005886	GO:0015031,GO:0016740,GO:0046872, GO:0005578,GO:0005615,GO:0005634, GO:0005737,GO:0005739,GO:0005764, GO:0005783,GO:0005789,GO:0005856
14	Synapse	GO:0045202	GO:0045202, GO:0045211	GO:0016020,GO:0016021,GO:0016787, GO:0046872,GO:0005615,GO:0005634, GO:0005737,GO:0005739,GO:0005856

(if any). For example, the SEPos GO terms for *centrosome* include its key GO term "GO:0005813"; the SEPos GO term for *cytoplasm* is the same as its key GO term "GO:0005737". This suggests that our experimental results are consistent with GO annotations, i.e., the key GO terms play more significant roles in determining the final predictions of mLASSO-Hum. There is one exception—GO:0008543, which has been removed from the list of SEPos GO terms although its weights are larger than 0.1. This is because it appears to act as a SEPos GO term in 13 subcellular locations, which in fact has lost the indicative functions of SEPos GO term for localization. We have found that only less than 1% (24 out of 3106) proteins have this GO term, which, despite its larger weights, will have limited contributions to the final decisions.

Table 2 also lists the SENeg GO terms for the 14 subcellular locations. The presence of these SENeg GO terms can be used to indicate that the query protein is **not** located in the corresponding subcellular location. Comparing the weights of these SENeg GO terms and the SEPos GO terms reveals the degree of exclusiveness between different subcellular locations, namely the possibility that a protein cannot co-localize in two or more particular subcellular locations. For example, in the Row 10 and 13 of Table 2, we note that "GO:0005739" is a SEPos GO term for *mitochondrion*, whereas it is a SENeg GO term for *plasma membrane*; "GO:0005886" is a SEPos GO term for *plasma membrane*, whereas its direct ancestor "GO:0016020" is a SENeg term for *mitochondrion*.[7] This means that a protein is highly likely to **not** co-localize in both *mitochondrion* and *plasma membrane*. This is actually the case in the human benchmark dataset, where only one out of the 3106 proteins co-localizes in *mitochondrion* and *plasma membrane*.

Moreover, from the list of SENeg GO terms, we can find that SENeg GO terms for a particular subcellular location are likely to be overlapped with those SEPos GO terms for other subcellular locations. This means that some GO terms may be indicative for a query protein to reside in a subcellular location, whereas simultaneously they are indicative of not residing in other subcellular location. In other words, these essential GO terms are not favorable to making multi-label decisions. This is understandable because the percentage (17%, 526 out of 3106) of multi-label proteins is remarkably lower than that (83%) of single-label proteins. Nevertheless, this does not mean that mLASSO-Hum cannot make predictions on multi-label proteins, because the final prediction is based on the overall significance of all of the essential GO terms. For example, in Fig. S1(b) of supplementary materials, the weight (0.33) for "GO:0005737" is even larger than the absolute weight (|−0.19|) for "GO:0005634" (the key GO term of *nucleus*); and in Fig. S1(k), the weight (0.28) for "GO:0005634" is also larger than the absolute weight (|−0.16|) for "GO:0005737". This means that the presences of both "GO:0005634" and "GO:0005737" for a query protein may still indicate that it locates in both *cytoplasm* and *nucleus*.

10. Conclusions

This paper proposes an interpretable multi-label predictor, namely mLASSO-Hum, which is based on a depth-dependent GO hierarchical information-based method and a multi-label LASSO classifier to predict subcellular localization of both single- and multi-location *homo sapien* proteins. Specifically, given a query protein, a GO frequency vector is constructed by exploiting the information in the ProSeq-GO database. By using the one-vs-rest LASSO classifiers, 87 out 8,000+ GO terms are selected. Based on these 87 essential GO terms, a depth-dependent GO hierarchical information-based method is proposed to incorporate the information from other non-essential GO terms into the feature vectors, which are again presented to multi-label LASSO classifiers for classification.

The key contributions of this paper can be summarized as follows: (1) Our experimental results are consistent with biological annotations, i.e., the key GO terms have higher weights in determining the corresponding subcellular location; (2) not only GO terms from cellular component category contributes to the prediction, but also those from the categories of molecular functions and biological processes; (3) by using mLASSO-Hum, we can obtain a sparse solution, and through this sparse solution, we can easily see which GO terms play more significant roles in indicating whether a query protein belongs to a certain subcellular location or not; (4) by incorporating the depth-dependent GO transferring information, the performance of mLASSO-Hum is significantly better than existing state-of-the-art multi-label human-protein predictors.

Experimental results on a recent human benchmark dataset demonstrate that mLASSO-Hum performs significantly better than existing state-of-the-art multi-label human-protein predictors. To enhance the impacts of computational methods [89, 90], we have provided a web-server for mLASSO-Hum available online at `http://bioinfo.eie.polyu.edu.hk/mLASSOHumServer/`.

Acknowledgment

This work was in part supported by HKPolyU Grant G-YN18 and G-YL78.

References

[1] G. Lubec, L. Afjehi-Sadat, J. W. Yang, J. P. John, Searching for hypothetical proteins: Theory and practice based upon original data and literature, Prog. Neurobiol 77 (2005) 90–127.

[2] V. Krutovskikh, G. Mazzoleni, N. Mironov, Y. Omori, A. M. Aguelon, M. Mesnil, F. Berger, C. Partensky, H. Yamasaki, Altered homologous and heterologous gap-junctional intercellular communication in primary human liver tumors associated with aberrant protein localization but not gene mutation of connexin 32, Int. J. Cancer 56 (1994) 87–94.

[3] M. D. Kaytor, S. T. Warren, Aberrant Protein Deposition and Neurological Disease, J. Biol. Chem. 274 (1999) 37507–37510.

[7] From the QuickGO [88] server, we can see that "GO:0005886" *plasma membrane* is a child term of "GO:0016020" (*membrane*); hence, "GO:0016020" which is a SENeg term for *mitochondrion* means that the presence of "GO:0005886" will also indicate that the query protein does not locates in *mitochondrion*.

[4] Y. Chen, C. F. Chen, D. J. Riley, D. C. Allred, P. L. Chen, D. V. Hoff, C. K. Osborne, W. H. Lee, Aberrant Subcellular Localization of BRCA1 in Breast Cancer, Science 270 (1995) 789–791.

[5] X. Lee, J. C. J. Keith, N. Stumm, I. Moutsatsos, J. M. McCoy, C. P. Crum, D. Genest, D. Chin, C. Ehrenfels, R. Pijnenborg, F. A. V. Assche, S. Mi, Downregulation of placental syncytin expression and abnormal protein localization in pre-eclampsia, Placenta 22 (2001) 808–812.

[6] A. Hayama, T. Rai, S. Sasaki, S. Uchida, Molecular mechanisms of Bartter syndrome caused by mutations in the BSND gene, Histochem. & Cell Biol. 119 (10) (2003) 485–493.

[7] M. C. Hung, W. Link, Protein localization in disease and therapy, J. of Cell Sci. 124 (Pt 20) (2011) 3381–3392.

[8] 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.

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

[10] M. Mandal, A. Mukhopadhyay, U. Maulik, Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC, Medical and Biological Engineering and Computing (2015) 331–344.

[11] X. Wang, W. Zhang, Q. Zhang, G. Z. Li, MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier, Bioinformatics (2015) 1–7.

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

[13] 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.

[14] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, Journal of Theoretical Biology 364 (2015) 284–294.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] S. Wan, M. W. Mak, S. Y. Kung, Protein subcellular localization prediction based on profile alignment and Gene Ontology, in: 2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'11), 2011, pp. 1–6.

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

[22] S. Wan, M. W. Mak, S. Y. Kung, Adaptive thresholding for multi-label SVM classification with application to protein subcellular localization prediction, in: 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'13), 2013, pp. 3547–3551.

[23] 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.

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

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

[26] S. Wan, M. W. Mak, S. Y. Kung, GOASVM: Protein subcellular localization prediction based on gene ontology annotation and SVM, in: 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'12), 2012, pp. 2229–2232.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

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

[33] 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`

[34] 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.

[35] 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.

[36] K. C. Chou, Y. D. Cai, Using GO-PseAA predictor to predict enzyme subclass, Biochemical and Biophysical Research Communications 325 (2) (2004) 506–509.

[37] Y. D. Cai, G. P. Zhou, K. C. Chou, Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition, Journal of Theoretical Biology 234 (1) (2005) 145–149.

[38] K. C. Chou, Y. D. Cai, Using GO-PseAA predictor to identify membrane proteins and their types, Biochemical and Biophysical Research Communications 327 (3) (2005) 845–847.

[39] Y. D. Cai, K. C. Chou, Predicting 22 protein localizations in budding yeast, Biochemical and Biophysical Research Communications 323 (2) (2004) 425–428.

[40] K. C. Chou, H. B. Shen, Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites, Journal of Proteome Research 6 (2007) 1728–1734.

[41] K. C. Chou, H. B. Shen, Recent progress in protein subcellular location prediction, Analytical Biochemistry 370 (1) (2007) 1–16.

[42] K. C. Chou, H. B. Shen, Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms, Nature Protocols 3 (2008) 153–162.

[43] A. Garg, M. Bhasin, G. P. S. Raghava, SVM-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search, J. of Biol. Chem. 280 (2005) 14427–14432.

[44] K. C. Chou, H. B. Shen, Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization, Biochem Biophys Res Commun 347 (2006) 150–157.

[45] 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.

[46] 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.

[47] 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.

[48] J. C. Mueller, C. Andreoli, H. Prokisch, T. Meitinger, Mechanisms for multiple intracellular localization of human mitochondrial proteins, Mi-

tochondrion 3 (2004) 315–325.

[49] K. C. Chou, Impacts of bioinformatics to medicinal chemistry, Medicinal Chemistry 11 (2015) 218–234.

[50] K. C. Chou, H. B. Shen, A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple site: Euk-mPLoc 2.0, PLoS ONE 5 (2010) e9931.

[51] H. B. Shen, K. C. Chou, A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0, Analytical biochemistry 394 (2) (2009) 269–274.

[52] K. C. Chou, H. B. Shen, Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization, PLoS ONE 5 (2010) e11335.

[53] 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.

[54] H. B. Shen, K. C. Chou, Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites, J. Biomol. Struct. Dyn. 26 (2010) 175–186.

[55] H. B. Shen, K. C. Chou, Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins, Protein and Peptide Letters 16 (12) (2009) 1478–1484.

[56] 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.

[57] 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.

[58] K. C. Chou, Z. C. Wu, X. Xiao, iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins, PLoS ONE 6 (3) (2011) e18258.

[59] 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 & Peptide Letters 19 (2012) 4–14.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] J. He, H. Gu, W. Liu, Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites, PLoS ONE 7 (6) (2011) e37155.

[66] S. Wan, M. W. Mak, B. Zhang, Y. Wang, S. Y. Kung, An ensemble classifier with random projection for predicting multi-label protein subcellular localization, in: 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2013, pp. 35–42. doi:10.1109/BIBM.2013.6732715.

[67] L. Q. Li, Y. Zhang, L. Y. Zou, C. Q. Li, B. Yu, X. Q. Zheng, Y. Zhou, An ensemble classifier for eukaryotic protein subcellular location prediction using Gene Ontology categories and amino acid hydrophobicity, PLoS ONE 7 (1) (2012) e31057.

[68] S. Briesemeister, J. Rahnenführer, O. Kohlbacher, YLoc—an interpretable web server for predicting subcellular localization, Nucleic Acids Research 38 (Suppl 2) (2010) W497–W502.

[69] W. Chen, P. M. Feng, E. Z. Deng, H. Lin, K. C. Chou, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Analytical Biochemistry 462 (2014) 76–83.

[70] Y. Xu, X. Wen, L.-S. Wen, L.-Y. Wu, N.-Y. Deng, K.-C. Chou, iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS ONE 9 (8) (2014) e105018.

[71] H. Lin, E. Z. Deng, H. Ding, W. Chen, K. C. Chou, iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition, Nucleic Acids Research 42 (21) (2014) 12961–12972.

[72] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS ONE 10 (2015) e0121501.

[73] J. Jia, Z. Liu, X. Xiao, iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC, Journal of Thoretical Biology 377 (2015) 47–56.

[74] K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review), Journal of Theoretical Biology 273 (2011) 236–247.

[75] Z. Lu, L. Hunter, GO molecular function terms are predictive of subcellular localization, in: In Proc. of Pac. Symp. Biocomput. (PSB'05), 2005, pp. 151–161.

[76] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, H. Shatkay, SherLoc2: A high-accuracy hybrid method for predicting subcellular localization of proteins, Journal of Proteome Research 8 (2009) 5363–5366.

[77] K. C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Molecular BioSystems 9 (2013) 1092–1100.

[78] X. Wang, G. Z. Li, A multi-label predictor for identifying the subcellular locations of singleplex and multiplex eukaryotic proteins, PLoS ONE 7 (5) (2012) e36317.

[79] K. Nakai, Protein sorting signals and prediction of subcellular localization, Advances in Protein Chemistry 54 (1) (2000) 277–344.

[80] 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.

[81] R. Tibshirani, Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[82] 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.

[83] Y. Lu, Y. Zhou, W. Qu, M. Deng, C. Zhang, A Lasso regression model for the construction of microRNA-target regulatory networks, Bioinformatics 27 (17) (2011) 2406–2413.

[84] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of statistics 32 (2) (2004) 407–499.

[85] 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.

[86] 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.

[87] T. Hastie, R. Tibshirani, J. Friedman, The element of statistical learning, Springer-Verlag, 2001.

[88] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, R. Apweiler, QuickGO: a web-based tool for Gene Ontology searching, Bioinformatics 25 (22) (2009) 3045–3046.

[89] B. Liu, F. Liu, X. Wang, J. Chen, Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Research (2015) doi:10.1093/nar/gkv1458.

[90] P. Du, S. Gu, Y. Jiao, PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets, International Journal of Molecular Sciences 15 (3) (2014) 3495–3506.