Bioinformatics, 33(6), 2017, 843–853
doi: 10.1093/bioinformatics/btw723
Advance Access Publication Date: 20 December 2016
Original Paper

OXFORD

Sequence analysis

Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features

Hang Zhou[1,2,†], Yang Yang[3,4,†] and Hong-Bin Shen[1,2,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, [2]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China, [3]Department of Computer Science and Engineering, Shanghai Jiao Tong University and [4]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai 200240, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on July 28, 2016; revised on October 31, 2016; editorial decision November 8, 2016; accepted on November 17, 2016

Abstract

Motivation: Protein subcellular localization prediction has been an important research topic in computational biology over the last decade. Various automatic methods have been proposed to predict locations for large scale protein datasets, where statistical machine learning algorithms are widely used for model construction. A key step in these predictors is encoding the amino acid sequences into feature vectors. Many studies have shown that features extracted from biological domains, such as gene ontology and functional domains, can be very useful for improving the prediction accuracy. However, domain knowledge usually results in redundant features and high-dimensional feature spaces, which may degenerate the performance of machine learning models.
Results: In this paper, we propose a new amino acid sequence-based human protein subcellular location prediction approach Hum-mPLoc 3.0, which covers 12 human subcellular localizations. The sequences are represented by multi-view complementary features, i.e. context vocabulary annotation-based gene ontology (GO) terms, peptide-based functional domains, and residue-based statistical features. To systematically reflect the structural hierarchy of the domain knowledge bases, we propose a novel feature representation protocol denoted as HCM (Hidden Correlation Modeling), which will create more compact and discriminative feature vectors by modeling the hidden correlations between annotation terms. Experimental results on four benchmark datasets show that HCM improves prediction accuracy by 5–11% and F_1 by 8–19% compared with conventional GO-based methods. A large-scale application of Hum-mPLoc 3.0 on the whole human proteome reveals proteins co-localization preferences in the cell.
Availability and Implementation: www.csbio.sjtu.edu.cn/bioinf/Hum-mPLoc3/
Contacts: hbshen@sjtu.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein subcellular localization is crucial for understanding protein functions, regulation mechanisms and protein-protein interactions. However, it is often laborious and costly to identify a protein's cellular compartment using wet-lab experiments, thus in-silico prediction tools are highly desired when working with large scale datasets of proteins with unknown locations. According to our statistics on the SWISS-PROT database (Boeckmann *et al.*, 2003) released on February 2016, only 10.4% of the 550 552 proteins have experimentally verified localization annotations (Supplementary Fig. S1). The vast number of proteins with unknown or uncertain locations requires reliable and efficient prediction methods.

Many automatic localization prediction tools have been developed. Popular online predictors include BaCelLo (Pierleoni *et al.*, 2006), YLoc (Briesemeister *et al.*, 2010), MultiLoc (Höglund *et al.*, 2006), GOASVM (Wan *et al.*, 2013), WoLF PSORT (Horton *et al.*, 2007), CellPLoc (Chou and Shen, 2010), HSLPred (Garg *et al.*, 2005), etc. These prediction algorithms and tools have provided great convenience for wet-lab scientists from proteomics and related fields.

Protein subcellular localization information has been widely used to assist disease gene discovery and drug target identification (Bakheet and Doig, 2009; Lahti *et al.*, 2012). For instance, the role that the Hippo/YAP pathway played in the development of pediatric hepatocellular carcinoma, was studied by examining the expression and subcellular localization of protein YAP in tumors (LaQuaglia *et al.*, 2016). Drug targets have also been found to favor certain subcellular localizations (Bakheet and Doig, 2009). An easy-to-use prediction tool with high accuracy will be very helpful to these wet-lab and clinical studies. Our previously released web server, HumPLoc 2.0, was specially designed for predicting human protein localizations. The number of times it is used per year has risen from 20 000 in 2010 to over 80 000 in 2015 (Supplementary Fig. S2). This indicates the importance of a further enhancement of the prediction power based on new technology and more refined annotation databases, to provide better prediction services.

Generally, computational methods for the identification of protein subcellular localization can be grouped into two categories, i.e. homolog search-based and machine learning-based approaches. The homology search-based approach can be considered as a nearest neighbor predictor, where the distance between two proteins is usually measured by their sequence identity. By searching the query protein against a large pool of annotated sequences, this method finds the top K closest proteins, and transfers their annotations to the query protein (Nair and Rost, 2002). This is a quite straightforward protocol, but its performance significantly depends on the homology targets detected (Wan *et al.*, 2013). Furthermore, the twilight-zone phenomena (Gardy *et al.*, 2003), i.e. the proteins that share high sequence identity could have very different structures or functions, would also result in exceptions of this protocol.

The machine learning-based predictors are a class of flexible models in the protein subcellular location predictions. They require the so-called training dataset to learn the classification rules by statistical learning algorithms. Thus, the quality of the training data is closely related to the quality of learned statistical rules. Benefitting from more and more reliable annotations on subcellular localization of protein databases, the classification model can be trained more sufficiently through a collection of large-scale training data. The other important issue in machine learning-based models is how to encode protein sequences, since most algorithms require numerical feature vectors as input. How to extract discriminative features from raw protein sequences as well as associated prior knowledge is crucial to the final performance. Existing machine learning tools for predicting subcellular locations use various features as follows:

(i) The residue-based statistical characteristics, such as the k-mer frequencies (Cedano *et al.*, 1997; Emanuelsson *et al.*, 2000; Park and Kanehisa, 2003), pseudo-amino-acid composition (Chou and Shen, 2006; Shen and Chou, 2007a, 2008) and Position Specific Scoring Matrix (PSSM) (Chou and Shen, 2007; Nanni *et al.*, 2013; Pierleoni *et al.*, 2006; Xie *et al.*, 2005);

(ii) The peptide-based features, such as sorting signals (usually in the N-terminal) (Emanuelsson *et al.*, 2000; Horton *et al.*, 2007; Psort, 1997; Petsalaki *et al.*, 2006; Savojardo *et al.*, 2015; Small *et al.*, 2004), functional domains (Chou and Cai, 2002; Marchler-Bauer *et al.*, 2005) and sequence motifs (Scott *et al.*, 2004);

(iii) The context vocabulary annotation-based features, such as the Gene Ontology (GO) terms (Ashburner *et al.*, 2000; Chou and Cai, 2003).

Since GO terms contain high-level abstraction of domain knowledge, they often result in higher accuracy than the residue- or peptide-based features when sufficient annotations are available. However, the large-size of annotation data brings new algorithmic challenges. For example, by using a Bernoulli event model for each GO term, i.e. binary coding for presence/absence of a GO term, the GO-based methods often result in an extremely high dimensional feature space, in which tens of thousands of GO terms are included (Blum *et al.*, 2009; Shen and Chou, 2009). As the GO database is expanded and updated regularly, the dimensionality will keep increasing with our expanded knowledge about proteins. The high dimensional feature vectors increase the complexity of the following learning process and also influence the prediction performance considering the potential noise in the annotation database. It is interesting to note that although the entire GO database is huge, each protein actually contains only a few terms. According to our statistics, proteins which have at least one GO term in the SWISS-PROT database are annotated by 6 GO terms on average. This will give us a sparse feature vector, which has thousands of dimensions but only approximately 6 useful components. Different methods have been proposed to handle such high-dimensional but very sparse feature vectors. For instance, YLoc (Briesemeister *et al.*, 2010) only selects the GO terms and PROSITE patterns which are typical for particular subcellular locations. Thus, it reduces unnecessary features and makes the results more interpretable, though it may suffer information loss. The WegoLoc (Chi and Nam, 2012) assigns a weight for each GO term and it can highlight the useful GO terms.

In this study, we encode feature vectors by GO correlation information instead of using the presence or frequency of GO terms. It is well known that GO terms are organized by a hierarchical structure in three directed acyclic graphs (DAGs), i.e. biological process (BP), molecular function (MF) and cellular component (CC). The terms are correlated by paths consisting of different types of edges (i.e. relationships) in the DAGs. Many methods that define the semantic similarity between GO terms have been proposed, such as information content-based (Jiang and Conrath, 1997; Lin, 1998; Resnik *et al.*, 1999) and graph-based methods (Wu *et al.*, 2005; Wang *et al.*, 2007; Zhang *et al.*, 2006). However, to the best of our knowledge, very few predictors of protein subcellular localization take into account the correlation between GO terms. This motivates us to incorporate the hidden correlation between GO terms to get a better similarity measure between two high-dimensional but sparse GO feature vectors. We propose a new protocol, called HCM (Hidden Correlation Modeling), to exploit the hidden correlation between the annotation features of proteins. In order to

deal with the lack of GO annotation for some query proteins due to the incompleteness of the GO database, we also incorporate the statistical residue features, as well as the peptide-based functional domain features which are extracted from Conserved Domain Database (CDD). With these new advantages in feature representation, we have constructed a new predictor, called Hum-mPLoc 3.0, which is named after our previously developed predictor for human protein localization predictions, but endowed with an entirely new feature representation.

2 Material and methods

2.1 Datasets

In this study, we mainly focus on human proteins, considering the predictors specific to human proteins are still relatively few when comparing to the rapidly increasing need for the targets annotations. We constructed a new benchmark dataset for human proteins, named HumB, by collecting all human proteins from SWISS-PROT released on January 2012. To ensure high data quality, we excluded the proteins that have no subcellular locations or have uncertain annotation with keywords like 'by similarity', 'potential' and 'probable'. Moreover, we used PISCES (Wang and Dunbrack, 2003) to remove redundant sequences, with the identity cutoff of 25%, i.e. to cluster similar sequences and get representative proteins automatically outputted by PISCES. Then we extracted their localization information from SWISS-PROT. In this study, we focused on 12 major compartments in human cells, including centrosome, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracellular, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome and plasma membrane. Finally, the benchmark dataset includes 3129 human proteins, 2306 of which have single subcellular location and the rest are multi-locational proteins. Intuitively, each location can be regarded as a class label, and a protein with more than one location is a multi-labeled sample. HumB has a total of 4229 labels, and each protein has 1.35 labels on average.

Besides the benchmark set HumB, an independent test set named HumT was also prepared for performance evaluation, from a May 2015 SWISS-PROT release. Proteins annotated with experimentally verified subcellular locations in the release of January 2012 were removed. In other words, HumT has no overlap with HumB. Moreover, in order to reduce bias, sequence similarity between HumB and HumT was limited to below 25%. To ensure the quality of assessment, we only considered the protein locations supported by experimental evidence, i.e. only the human proteins whose CC field contains 'ECO:269' were collected (Evidence Codes Ontology, ECO, is a controlled vocabulary of terms that describes the source of the information and ECO:269 represents a type of experimental evidence). Finally, HumT includes 379 human proteins and 541 labels. (Data distributions of HumB and HumT are shown in Supplementary Table S1).

Although we mainly focus on the prediction of human protein localization in this study, the HCM method can be extended to other species. In order to compare with the existing cutting-edge prediction tools (not limited to human protein predictors), we also tested the HCM model on several other well-established datasets published by other researchers, including animals proteins in the BacelLo dataset (Pierleoni *et al.*, 2006), animal proteins in the Höglund dataset (Höglund *et al.*, 2006) and the DBMLoc dataset (Zhang *et al.*, 2008). Note that to fairly compare with the methods for other species, we re-trained the HCM model on these three datasets for model comparison. The details of these three sets are given in the Supplementary Materials.

2.2 Methods

This study aims to develop a machine learning-based predictor for subcellular localization of human proteins. Figure 1 shows the overall architecture of the new predictor, including two major parts, feature extraction and classifier construction. The feature vectors produced by the new feature presentation protocol, HCM, cover both residue statistics and biological prior knowledge. Details on each type of feature are given in Sections 2.2.1, 2.2.2 and 2.2.3, respectively.

2.2.1 Residue-based statistical features

The statistical properties of residues are the building blocks in the feature vectors of a subcellular location predictor, especially when annotation data is not available. Here, the residue-based features include the 20-D amino acid composition (*AAC*) and evolutional information represented by the Position Specific Scoring Matrix (PSSM). The matrix, S_{PSSM} (Eq. (1)) for each protein sequence, is constructed by using PSI-BLAST to search SWISS-PROT with an E-value cutoff of 0.001 (Altschul *et al.*, 1997),

$$S_{\mathrm{PSSM}}=\begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,20} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,20} \\ \cdots & \cdots & \ddots & \cdots \\ S_{L,1} & S_{L,2} & \cdots & S_{L,20} \end{bmatrix}, \tag{1}$$

where $S_{i,j}$ represents the probabilistic score that the jth ($1\le j\le 20$) amino acids occurring at the ith ($1\le i\le L$) position of the sequence, and L represents the length of the protein sequence.

In order to condense the matrix into a feature vector with fixed length, each column is averaged into a single value. Note that residues at different positions of the sequence usually have various mutation rates, thus each row is first normalized to reduce potential bias. The z-score normalization is adopted here (Eq. (2)),

$$S^0_{i,j}=\frac{S_{i,j}-\frac{1}{N}\sum_{k=1}^{N}S_{i,k}}{\sqrt{\frac{1}{N-1}\sum_{u=1}^{N}\left(S_{i,u}-\frac{1}{N}\sum_{k=1}^{N}S_{i,k}\right)^2}}, \tag{2}$$

where $S^0_{i,j}$ represents the normalized score and N represents the number of different amino acids, i.e. N is equal to 20. Then for each column, an average score is calculated as Eq. (3),

$$\overline{S^0_j}=\frac{1}{L}\sum_{i=1}^{L}S^0_{i,j}. \tag{3}$$

After these two operations, the S_{PSSM} is transformed into a 20-D vector in Eq. (4),

$$\overline{S_{\mathrm{PSSM}}}=\left[\overline{S^0_1},\overline{S^0_2},\overline{S^0_3},\ldots,\overline{S^0_{20}}\right]. \tag{4}$$

Then, AAC and the normalized PSSM vector are combined into a 40-D vector, which catches not only amino acid frequency information of the protein itself, but also the residue statistics from its functional related homologs. Furthermore, considering that localization information is often implied in the N-terminal and C-terminal of amino acid sequences (Pierleoni *et al.*, 2006), we extracted sequence features of multiple segments from both terminals, specifically, the first 10, 20, …, 60 residues of N-terminal, and the last 10, 20, …, 100 residues of C-terminal. For each segment, a 40-D vector is created using the method described above (AAC + PSSM). By

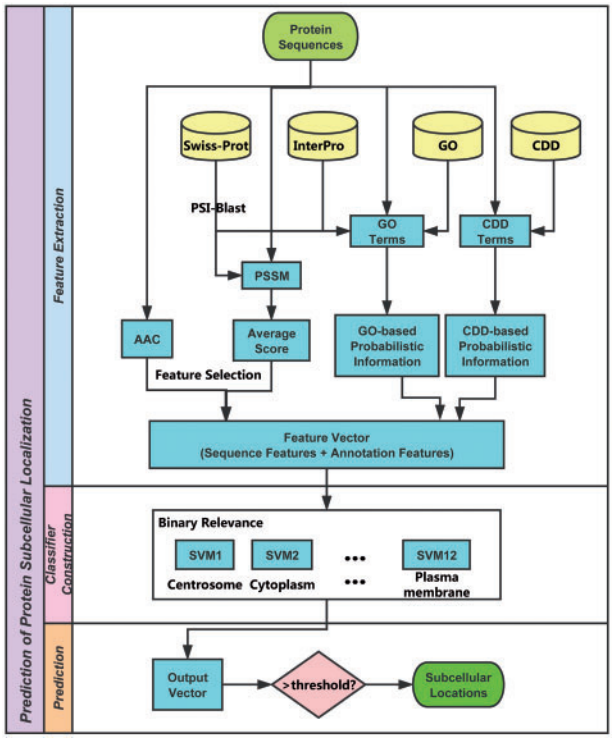

Fig. 1. Flowchart of the new predictor Hum-mPLoc 3.0

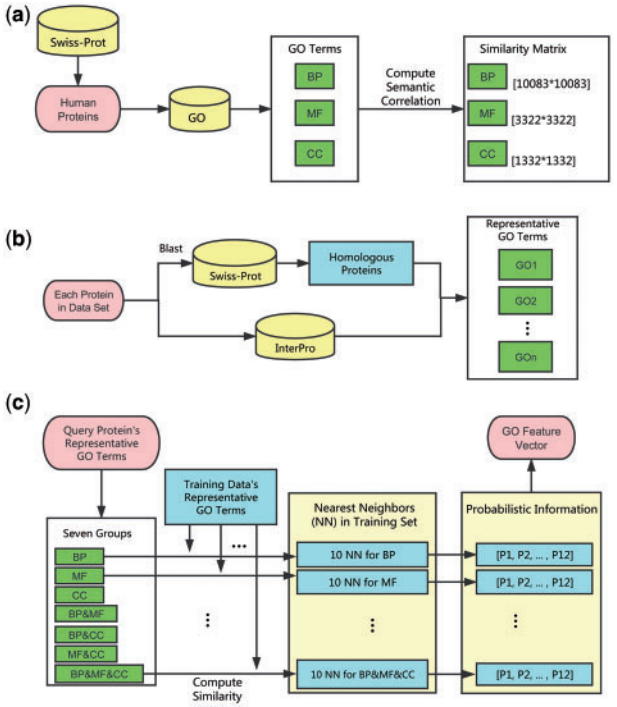

Fig. 2. Flowchart of GO-based feature extraction

concatenating all these 40-D vectors (for the full sequence and 16 segments), the total dimensionality becomes 680 (40×17), which may contain some redundant information. Thus, the Correlation-based Feature Selection (Hall and Smith, 1999) (CFS) method is adopted to select the important features and finally leads to a 43-D feature vector.

2.2.2 Nearest neighbor-guided GO annotation correlation features

The feature extraction procedure consists of three parts as shown in Figure 2. Before exploiting the correlation information, we need a matrix of pairwise GO similarities. Considering it is very costly to construct such a matrix with tens of thousands of terms in GO database, we only use GO terms annotated for human proteins searched in SWISS-PROT. The GO annotation contains both experimentally supported and computationally inferred GO terms. Here, only the first type is considered to ensure the quality of annotations, which includes 10083 BP, 3322 MF and 1332 CC terms. By using an improved information content-based measure (Yang *et al.*, 2012), the three similarity matrices are constructed for BP, MF and CC, respectively.

In Figure 2(b), it can be observed that, instead of using proteins' own GO terms, we retrieve the representative GO terms for each protein from its homologous proteins. This is based on the consideration that many proteins have no or scarce GO annotation (Shen and Chou, 2009). Mei (2012) and Wan *et al.* (2013) adopted the same strategy in their studies. Specifically, the homologs, i.e. the proteins which have more than 50% sequence identity and 60% positives with the query protein, are searched by BLAST in SWISS-PROT. The GO terms are extracted from both SWISS-PROT and InterPro database (Zdobnov and Apweiler, 2001).

Given the correlation matrices of GO terms and representative GO terms for each protein, the GO features are produced according to the following two steps.

1. Search nearest neighbors for query proteins

Intuitively, a query protein would have a high probability of having the same subcellular locations as its most similar proteins according to their GO annotation. Here we identify the query protein's nearest neighbors in the training set based on the similarity measured by semantic correlation between GO terms. Specifically, the similarity between the query protein and the kth training protein is defined as the square root of the sum of squared correlation between each GO term of the query protein and the GO terms set of the training protein, K, as shown in Eq. (5),

$$Sim_k = \sqrt{\sum_{i=1}^{n} Cor(x_i, K)^2}, \tag{5}$$

where n is the number of representative GO terms of the query protein and $x_i$is its ith representative component, and K is the GO terms set of the kth training protein. $Cor(x_i, K)$ is defined in Eq. (6),

$$Cor(x_i, K) = \max_{1 \le i \le m} Cor(x_i, y_j), \tag{6}$$

where y_js are GO terms, $K = \{y_1, y_2, \ldots, y_m\}$, and $Cor(x_i, y_j)$ is an component in the correlation matrices of GO terms.

According to the similarity measurement (Eq. (5)), the query protein's nearest neighbors in the training set can be identified. Since BP, MF and CC are three respective DAGs in GO database, they may play different roles in measuring the similarity between gene products. Therefore, we divided the representative GO terms of each protein into 7 groups, i.e. BP, MF, CC, BP&MF, MF&CC, BP&CC and BP&MF&CC. Similarity scores are computed and the top 10 nearest neighbors are selected in each of the 7 groups.

2. Generate probabilistic information

In this step, feature vectors are represented by probabilistic information. Let pro_a denotes the probability of the query protein being in the location a. Initially, pro_a is defined as the ratio between the sum of similarities with the nearest neighbors localized at a and the

sum of similarities with all the 10 nearest neighbors, as shown in Eq. (7),

$$pro_a = \frac{\sum_{j\in I_{N_a}} sim_j}{\sum_{i\in I_N} sim_i}, \tag{7}$$

where I_N is the index set of all the nearest neighbors of the query protein ($|I_N| = 10$), and I_{Na} is the index set of the nearest neighbors which are located at a ($I_{Na} \subseteq I_N$). However, due to the incompleteness of GO annotation, some proteins may have no or few neighbors in the training set. Therefore, we tackle this problem with a smoothing technique by adding a Bayesian prior shown in Eq. (8). The prior is equal to the proportion of proteins locating at a, which gives us:

$$pro_a = \frac{\sum_{j\in I_{N_a}} sim_j + \frac{num_a}{num}}{\sum_{i\in I_N} sim_i + 1}, \tag{8}$$

where num_a and num are the numbers of proteins locating at a and the total number of proteins in the training set, respectively.

For each of the 7 GO groups, a 12-D vector was calculated consisting of the probabilistic information for 12 locations. Finally, a feature vector with 84-D was generated. In order to produce the probabilistic information for training proteins, a 10-fold cross validation was conducted.

2.2.3 Nearest neighbor-guided functional domain correlation features

Besides the statistical properties of single residues, conserved peptides are also helpful to identify subcellular localization. We use the Conserved Domain Database (CDD) (v3.12) from ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/, and produce CDD-based features and their hidden correlations are also modeled.

The first step, again, is to construct a correlation matrix. For all the 3129 human proteins in HumB, CDD terms are searched by RPS-BLAST with an E-value cutoff of 0.001, and we only use the Superfamily information as CDD features, resulting in 2313 CDD terms in total.

Unlike the GO terms, CDD terms have no semantic structure. Here we adopt a symmetrical uncertainty method (Hall, 1999) to construct the pairwise correlation matrix for CDD features. Firstly, a 3129×2313-D binary matrix is constructed. The element at the ith row and jth column represents whether the ith protein contains the jth CDD term. Then two quantities of entropy are defined. The first one, $H(f_i^{cdd})$, is the entropy of the ith CDD feature (Eq. (9)),

$$H(f_i^{cdd}) = -\sum_{m\in\{0,1\}} p(f_i^{cdd} = m) \times \log p(f_i^{cdd} = m), \tag{9}$$

where $p(f_i^{cdd}=1)$ denotes the probability of the ith term being present in the training set. For example, the CDD term, cl21453, occurs 51 times in the training set, so the probability of this CDD features is 0.0163 (51/3129). The second one, $H(f_i^{cdd}, f_i^{cdd})$ is the differential entropy of the ith feature and jth feature (Eq. (10)),

$$\begin{aligned} &H(f_i^{cdd}, f_j^{cdd}) \\ &= -\sum_{n\in\{0,1\}}\sum_{m\in\{0,1\}} p(f_i^{cdd} = m \;\&\; f_j^{cdd} = n) \\ &\times \log p(f_i^{cdd} = m \;\&\; f_j^{cdd} = n). \end{aligned} \tag{10}$$

The correlation between two CDD terms is defined in Eq. (11),

$$S_{i,j}^{cdd} = \frac{2 \times (H(f_i^{cdd}) + H(f_j^{cdd}) - H(f_i^{cdd}, f_j^{cdd}))}{H(f_i^{cdd}) + H(f_j^{cdd})}. \tag{11}$$

The following steps are the same as GO feature extraction, i.e. step 1: use PRS-BLAST to extract CDD terms for query proteins; step 2: compute similarities of query proteins with training proteins based on the matrix $[S_{i,j}^{cdd}]_{2313\times2313}$; step 3: find top 10 nearest neighbors from the training set, and step 4: generate probabilistic information as a 12-D feature vector corresponding to the 12 locations studied in this study.

Finally, the residue-based features, GO and CDD features are combined into a 139 ($43 + 12 \times 7 + 12$) dimensional feature vector.

2.2.4 Multi-label classification

For the classification system, there are 12 class labels corresponding to 12 subcellular locations. We used support vector machines (Cortes and Vapnik, 1995) as classifiers, and adopted the binary relevance strategy (Boutell *et al.*, 2004) to construct 12 binary classifiers. Parameters γ and C were optimized via 10-fold cross validation.

In the test phase, the output for each test sample is a 12-D score vector. Each dimension of the vector represents the confidence of being in a certain subcellular location. The subcellular locations whose corresponding scores are positive are assigned to the test proteins, i.e. the threshold score is 0. If all the scores are negative, the subcellular location with the maximal score in the vector will be assigned.

2.2.5 Evaluation criteria

In this study, we used customized ACC and F_1 to evaluate the multi-label classification performance (Briesemeister *et al.*, 2010). Different from conventional accuracy and F_1 definition, the ACC is the average of individual accuracies for all test samples, and F_1 is the average of F_1 values of all locations. (Equations are in Supplementary Materials).

3 Experimental results

3.1 Comparison of different feature coding methods

In order to assess the performance of HCM, we compared it with four other feature extraction methods, namely SEQ + GO_7, SEQ + GO_1, SEQ + GO_0 and SEQ. Details are given below.

- SEQ+GO_7: Residue and GO features, i.e. HCM without CDD features.
- SEQ+GO_1: The same as SEQ+GO_7, except that GO terms from BP, MF and CC are used as a whole set, while HCM and SEQ+GO_7 consider 7 groups of GO terms.
- SEQ+GO_0: Residue and conventional GO features. The GO features are binary values, i.e. 1 for presence and 0 for absence. In order to avoid a high-dimensional feature space and conduct a fair comparison, the binary values are also converted to probabilistic information. The similarity between a query protein and the kth protein in training set is defined as $sim_k = 1 - hit_k/\sqrt{num_{query} \times num_k}$, where hit_k denotes the number of common GO features of these two proteins, num_{query} and num_k are the numbers of GO terms of the query protein and the kth protein, respectively. Similarly, the top 10 nearest neighbors are used to calculate the probability information as the features.
- SEQ: residue features only, i.e. the first part of HCM.

All of the above methods are tested on the aforementioned four datasets. For BaCelLo, Höglund and HumB, prediction accuracies were evaluated by using their reported test sets. The DBMLoc data has no separated test set, thus the accuracy was obtained via nested 5-fold cross-validation. The results are shown in Figure 3.

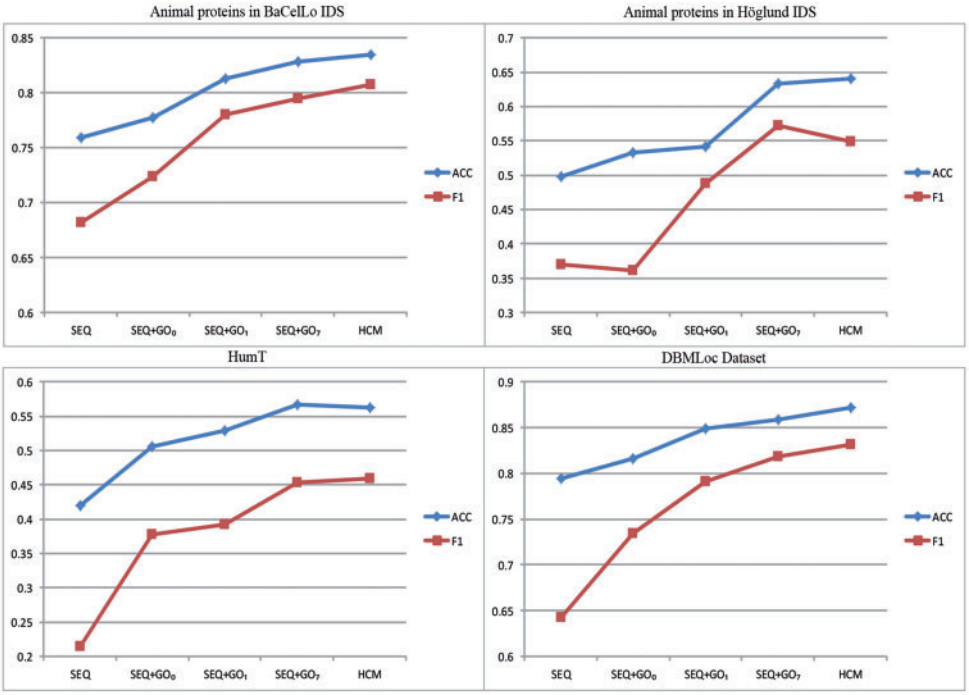

Fig. 3. Prediction accuracies of different types of features for four datasets

Generally, both *ACC* and F_1 are monotonically increasing from SEQ to HCM. The method containing only residue features is apparently not capable of providing reliable predictions. Especially for HumT, the accuracy is below 50%. Compared with SEQ + GO_0, SEQ + GO_1 performs better on all four datasets, with an increase of 2–5% for ACC and 2–13% on F_1. As for SEQ + GO_7, it achieves better performance than SEQ + GO_1. This may be due to two reasons. One is that the computation of GO set similarity on the whole set does not fully utilize the correlation between different categories of GO terms. The other reason is that reusing GO sets strengthens the impact of GO correlation-based features, making the GO information dominate the feature vector, which is beneficial for the classification in most cases. The last method, HCM, obtains a slight improvement over SEQ + GO_7 by adding CDD features. In summary, all three types of features in the proposed HCM method contribute to the discrimination of protein locations.

3.2 Prediction performance is affected by GO similarity definition and annotation coverage

In this section, three different variants of GO-based protein similarity definitions are assessed, and we also discuss the impact of GO annotation coverage on performance.

3.2.1 Strategies for computing the similarity between proteins

There are multiple ways to obtain a measure of similarity between two genes according to the similarities of their GO terms, such as the maximum (MAX) and best match average (BMA) approaches (Yu *et al.*, 2010). MAX chooses the maximum similarity of two GO terms from two genes. It only considers the most similar pair of GO terms, but fails to cover the overall similarity of two GO sets, thus it may be incapable of dealing with multi-locational proteins. For instance, protein of P42772 is annotated by GO:0005737 and GO:0005634, which suggests two locations, cytoplasm and nucleus, while the MAX strategy only leads to one location of nucleus. The BMA strategy first gets the maximum similarity for each GO term in one set to all GO terms in the other set, and then calculates an overall average value of these maximum values. BMA treats each GO term with equal weight, but since GO terms have different information content, the GO terms with more information content should have higher weights.

In the HCM pipeline of this study, the method (Yang *et al.*, 2012) that we used to calculate pairwise GO similarity tends to output high values for GO pairs located at bottom of the DAGs, where the bottom GOs will have high information content. In other words, the similarity value itself can be used as a weight. Thus, we adopted the Euclidean distance-based metric to measure the similarity between proteins. For instance, the homologous proteins Q5T6F0 and Q6PEY1 both contain and only one GO term, GO:0005515 (protein binding), which is a very general molecular function term with low information content. In this case, although these two genes have exactly the same GO sets, they will not be assigned a high similarity value when using this method.

We thus compared our metric implemented in HCM with MAX and BMA on the new HumT test set, where only the GO similarity calculation is different while all other steps are kept the same. The results show that HCM has the highest *ACC* and F_1. As for *ACC*, HCM is 1% higher than BMA and 3% higher than MAX. Also, HCM predicts the most proteins with all correct labels among the three methods (Supplementary Fig. S3). These results suggest the importance of defining the similarity between two GO sets because each protein corresponds to a set of GO terms. Our experimental results also indicate that the similarity definition which considers the DAG hierarchical structure of GO knowledge base is a better choice in this study.

3.2.2 Impact of GO annotation coverage

Although GO-based features have been widely demonstrated to have a positive impact on protein subcellular localization prediction,

previous studies have also shown that due to the incompleteness of the GO database, not all the proteins can be represented by the GO features and the performance will be affected by the GO feature coverage (Shen and Chou, 2009). To systematically investigate the impact of coverage on the performance, we tested several potential conditions of GO feature collection.

First, we conducted experiments to assess the performance with and without CC terms (Supplementary Fig. S4). The results show that without any CC terms, the BP and MF terms can improve both *ACC* and F_1 by 6.3% compared with the model using only the residue-based features. Moreover, using all the three types of GO terms can improve *ACC* by 8.6% and F_1 by 10.3% compared with using CC alone. These results indicate that BP and MF terms also contribute an important role in the identification of subcellular localization. Similar observations had been reported in previous studies. For example, Wan *et al.* found that over half of the most essential GO terms for subcellular localization come from the MF and BP categories (Wan *et al.*, 2015).

Second, we noticed that current annotations are affiliated with evidence codes in the GO database. There are two types of evidence codes assigned by curators, namely the experimental evidence codes and the computational analysis evidence codes. In the previous experiments, in order to ensure the reliability of annotation data, we only used the GO terms with experimental evidence. In order to test the effect of the source of GO terms, we also implemented another version using all available GO terms assigned by curators, including both experiment-supported and computationally inferred ones. By using all GO annotations, the GO feature coverage of proteins will be increased accordingly. Take the proteins in the HumT set as an example, the percentage of proteins with GO feature representations increases from 60.7% to 93.4% (by adding computationally inferred GOs, but excluding the test proteins' own GOs). Interestingly, the *ACC* and F_1 also further increased by 5% and 18%, respectively (Supplementary Fig. S5). These results suggest that the performance is closely related with the GO feature coverage, and current computational tools can also infer reliable GO annotations, which provide an important supplement for the incomplete experimental GO annotations. Since the existing predictors tend to use all the available GO terms for model construction, we adopted the same strategy for comparing results in the following experiments.

Finally, as shown above, even when we use all the available GO information, there are still some proteins which cannot be represented by any GO features, including 6.6% (25 protein samples) of the HumT dataset. Among 20 197 human proteins in SWISS-PROT database, 1041 (5.2%) proteins cannot be represented in the GO feature space. For these proteins, one common strategy is to use other features as a complement. For example, the functional domain CDD features and the residue-based statistical features (denoted as SEQ) are adopted in this paper. We thus further examined the performance difference in cases with and without effective GO features in the HumT set. For the 354 proteins in HumT which can be represented with GO features, the *ACC* and F_1 are 63.4% and 64.7% respectively. For the remaining 25 proteins (in 7 locations), their *ACC* and F_1 are 64.0% and 49.7% (Supplementary Table S6). Since they do not have GO features, we use a Bayesian prior as a complement when creating probabilistic information (Eq. (8)) together with the CDD and SEQ features. When only using the SEQ + CDD features, the *ACC* and F_1 of the 25 proteins will further decrease to 62.0% and 46.6%, respectively. These results suggest that GO features are important for localization prediction, especially for the minor classes (subcellular locations with relatively fewer samples). For instance, the F_1 score is improved by approximately 16% when incorporating the GO features compared to the models without GO features.

3.3 Comparison with state-of-the-art predictors

Table 1 compares Hum-mPLoc 3.0 with four existing subcellular location prediction methods specialized for human proteins, including YLOC+ (Briesemeister *et al.*, 2010) (YLoc+ can only predict 9 locations), iLoc-Hum (Chou *et al.*, 2012), WegoLoc (Chi and Nam, 2012) and mLASSO-Hum (Wan *et al.*, 2015). During the experiments, we directly submitted the proteins in the HumT dataset to the above online servers and got prediction responses. It's worth noting that the HumB and HumT datasets used to construct and test Hum-mPLoc 3.0 do not have any overlap as we stated before, but we did not remove the potential overlap between submitted HumT set and the training sets of the other four tools, in order to evaluate the performance on the same set. For a baseline comparison, we also listed the performance of a model denoted as SEQ + CDD, which does not use the GO features in Hum-mPLoc 3.0 in Table 1.

As can be seen from this table, compared to iLoc-Human, the *ACC*s are 0.63 (Hum-mPLoc 3.0) versus 0.41 (iLoc-Human), and F_1 values are 0.65 versus 0.32, respectively. When comparing to WegoLoc, the *ACC* of Hum-mPLoc 3.0 is better by 0.13 (0.63 versus 0.50) and F_1 is better by 0.21 (0.65 versus 0.44). When comparing to mLASSO-Hum, the *ACC*s are comparable (0.63 versus 0.65), while the F_1 of Hum-mPLoc 3.0 is 9% better (0.65 versus 0.56). The reason may be that Hum-mPLoc 3.0 works better on minor classes, like lysosome and peroxisome, thus resulting in a high averaged F_1. Hum-mPLoc 3.0 obtains the highest F_1 value on 6 of the 12 locations, and mLASSO-Hum performs the best on 5 locations. It is also worth mentioning that the mLASSO-Hum method searches the closest protein with GO subcellular location annotation from the database for the query protein, and we found that 216 (57% of the HumT dataset) sequences submitted to mLASSO-Hum may be predicted by using the GO terms from the submitted proteins themselves, based on the responses of the server (E-value = 0.0). The reason may be that many proteins in the HumT dataset have been covered by the training database used in mLASSO-Hum. Interestingly, although the SEQ + CDD model does not incorporate the GO terms, it works no worse than YLoc+ or iLoc-Human, which contains GO features. This suggests that the functional domains also play a substantial role in predicting protein subcellular locations.

In order to further test the generalization and application ability of the HCM driven predictor on other species, we tested our protocol on three other well established datasets including proteins from animal, plant and eukaryote. We re-trained three HCM-driven models on these three datasets, and used the same test sets or nested 5-fold cross validations as by the other existing predictors we are comparing to. Table 2 shows the comparison of the HCM driven predictor with some state-of-the-art predictors, including YLoc (Briesemeister *et al.*, 2010), MultiLoc (Höglund *et al.*, 2006) and BaCelLo's method (Pierleoni *et al.*, 2006). As shown in the table, the new method has substantially improved *ACC* and F_1 for all 3 datasets. For the two mono-locational datasets, BacelLo and Höglund, the *ACC*s of the new method are 7% higher than the best *ACC*s obtained by other predictors, respectively. For the

Table 1. Comparison of human protein subcellular location predictors on HumT dataset[a]

Location	YLoc+[b]			iLoc-Human[c]			WegoLoc[d]			mLASSO-Hum[e]			SEQ+CDD[f]			Hum-mPLoc 3.0[g]		
	pre	rec	F1	pre	rec	F1	pre	rec	F1	pre	rec	F1	Pre	rec	F1	pre	rec	F1
Centrosome	–	–	–	0	0	0	0.75	0.14	0.23	0.59	0.59	0.59	0	0	0	0.75	0.55	0.63
Cytoplasm	0.55	0.85	0.67	0.5	0.54	0.52	0.69	0.53	0.6	0.93	0.51	0.66	0.63	0.65	0.64	0.76	0.73	0.74
Cytoskeleton	–	–	–	0	0	0	0.32	0.34	0.33	0.9	0.22	0.35	1	0.07	0.14	0.8	0.68	0.74
ER	0.71	0.12	0.21	0	0	0	0. 73	0.2	0.31	0.74	0.49	0.59	0.9	0.22	0.35	0.83	0.37	0.51
Endosome	–	–	–	0	0	0	0. 25	0.07	0.11	0.38	0.2	0.26	0	0	0	0.58	0.47	0.52
Extracellular	0.39	0.85	0.54	0.62	0.62	0.62	0.67	0.77	0.71	0.16	0.69	0.26	0.32	0.54	0.4	0.5	0.46	0.48
Golgi apparatus	0.1	0.05	0.07	0.6	0.3	0.4	0.6	0.15	0.24	0.72	0.65	0.68	0.29	0.1	0.15	0.69	0.45	0.55
Lysosome	0	0	0	0.5	0.13	0.2	0.2	0.13	0.15	0.55	0.75	0.63	0.5	0.13	0.2	0.71	0.63	0.67
Mitochondrion	0.65	0.43	0.52	0.95	0.33	0.49	0.79	0.73	0.76	0.83	0.88	0.85	0.78	0.53	0.63	0.78	0.75	0.76
Nucleus	0.41	0.57	0.48	0.54	0.7	0.61	0.65	0.64	0.64	0.85	0.7	0.76	0.47	0.74	0.57	0.75	0.71	0.73
Peroxisome	0.07	0.5	0.13	1	0.5	0.67	0.5	1	0.67	0.29	1	0.44	0	0	0	1	1	1
Plasma membrane	0.41	0.44	0.42	0.42	0.33	0.37	0.44	0.53	0.48	0.58	0.56	0.57	0.52	0.27	0.36	0.65	0.44	0.52
ACC	0.45			0.41			0.50			0.65			0.47			0.63		
F_1	0.34			0.32			0.44			0.56			0.29			0.65		

[a]ER: Endoplasmic reticulum. pre denotes precision, and rec denotes recall.
[b]http://abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi.
[c]http://www.jci-bioinfo.cn/iLoc-Hum.
[d]http://www.btool.org/WegoLoc ((the multiplex threshold was set to 1, which is the best on HumT dataset after trying different values.).
[e]http://bioinfo.eie.polyu.edu.hk/mLASSOHumServer, where 216 of the 379 submitted proteins were predicted using their own GO terms according to the response of the server.
[f]SEQ + CDD uses only sequence and CDD features, without GO features.
[g]Query proteins' own GO terms have been removed.

Table 2. Comparison of seven predictors on three datasets[a]

	ACC/F_1		
	BaCelLo	Höglund	DBMLoc
YLoc-LowRes	0.79/0.75	–	–
YLoc-HighRes	0.74/0.69	0.56/0.34	–
YLoc+	0.58/0.67	0.53/0.37	0.64/0.68
MultiLoc2-LowRes	0.73/0.76	–	–
MultiLoc2-HighRes	0.68/0.71	0.57/0.41	–
BaCelLo	0.64/0.66	–	–
HCM-driven predictor	0.86/0.84	0.64/0.59	0.87/0.84

[a]YLoc+ and HCM-driven predictor can deal with multiple-locational proteins; Results of BaCelLo, YLoc and MultiLoc were extracted from (Pierleoni *et al*., 2006; Briesemeister *et al*., 2010; Höglund *et al*., 2006), respectively.

multi-locational dataset DBMLoc, the improvement is more significant, where the *ACC* and F_1 are 23% and 16% greater compared with YLoc+.

The performance improvement compared with other predictors may be mainly because the proposed HCM method can catch the hidden correlations, thus making the samples cluster in a more condensed space, and it uses the renewed annotation database. For example, YLoc adopts conventional binary coding to express GO features, and to avoid high dimensionality, it only considers the annotation-based features which are directly related with protein localization to certain compartments. However, the annotation terms without any indication of localization may also help, such as the BP and MF terms, as we showed before in Section 3.2. Besides, the annotation databases have been updated rapidly, and our method uses the latest version of gene ontology and conserved domain databases, which have more coverage than the old versions used by previous predictors.

3.4 Large-scale prediction on the whole human proteome

We applied Hum-mPLoc 3.0 to all 20197 human proteins in SWISS-PROT released on Feb., 2015 (Supplementary Fig. S6 shows the percentages of their localization annotation). Supplementary Figure S7 shows two pie charts of the distributions for 12 subcellular locations. One is from the set of human proteins with experimentally verified locations, and the other is from our predicted results on the whole human proteome. Intuitively, the two distributions are very similar, e.g. cytoplasm 29.7% versus 24.2%, nucleus 25.1% versus 24.1%, and plasma membrane 16.8% versus 16.6%. In the predicted results of the total 20197 proteins, we found that 16717 proteins have only one location (82.8%), 3104 proteins have two locations (15.4%), 335 proteins have three locations (1.7%), and 41 proteins have four locations (0.2%). Interestingly, by examining the co-localization patterns, we found that the most frequently co-occurred pairs are nucleus and cytoplasm (1718 times), followed by cytoskeleton and cytoplasm (869 times). The detailed times of co-occurrence for each pair of locations are listed in Supplementary Table S7. We performed hierarchical clustering on this matrix, and depict a heat map and a cluster dendrogram in Supplementary Figure S8. From the hierarchical tree, several clusters can be observed: <centrosome, cytoskeleton, cytoplasm, nucleus>, <extracellular, plasma membrane>, <ER, endosome, Golgi apparatus>, indicating an interesting relation and organization of the cellular compartments. We further checked frequent triples and quadruples of locations. Top ranked triples include: <centrosome, cytoskeleton, cytoplasm> (155), <cytoskeleton, cytoplasm, nucleus> (39), and <ER, endosome, Golgi apparatus> (32). The most frequent quadruple is <ER, endosome, Golgi apparatus, plasma membrane>, which takes up 35 of the 41 proteins that co-localize at 4 compartments. These combinations are consistent with the clusters yielded by the hierarchical tree in Supplementary Figure S8. These large-scale prediction results are also available at the Hum-mPLoc 3.0 website (www.csbio.sjtu.edu.cn/bioinf/Hum-mPLoc3/).

4 Discussion

4.1 Pitfall of GO features

Our results as well as other studies have shown that GO features are important for the protein subcellular localization prediction due to the fact that they represent a high-level knowledge of proteins. Two problems have limited their high efficacy in the real-world model construction: (i) the incompleteness of GO annotation database and (ii) the very sparse characteristics of GO feature vector. For instance, of the current whole human proteome, 5.2% proteins (1041/20197) have no GO features and 25.7% proteins (5181/20197) are associated with less than 5 GO terms (more than 14000 terms are observed in the database). Our results show that the HCM model proposed in this paper can efficiently deal with the high-dimensional and sparse feature learning problem in a much lower feature space. At the same time, our results also show that incorporating the CDD and SEQ features also plays an indispensable role for the prediction task, especially when the annotation data is incomplete or unreliable. The human protein NPIP, for example, can only be represented by one GO term in this study, GO:0005505, which cannot give many informative clues for the prediction. Thus, the CDD and SEQ features play the leading role for predicting its subcellular locations in the model. Another example is protein MTO1, with the following GO terms: GO:0044822, GO:0008033, GO:0050660 and GO:0002098. By using these GO terms, a prediction result of cytoplasm is obtained, but with low confidence, which is in fact a wrong answer; while CDD + SEQ-based features correctly predict that the protein is located in the mitochondrion with high confidence. These examples suggest that functional domain and residue-based features are very essential in the prediction of protein subcellular localization, and can function against bias induced by incomplete and sparse GO annotation data.

4.2 Usage of cross-species GO terms

In this study, we convert the similarity between two GO sets to the similarity between their annotated proteins, and find nearest neighbors for each query protein. The neighbors are searched in a cross-species manner, i.e. the neighbors include proteins from other species. We conducted an experiment on the HumT dataset using only human proteins as neighbors. The results show an obvious drop in accuracy, with a 8% drop in *ACC* and 14% drop in F_1 (Supplementary Fig. S5). This indicates that the homologs in different species also share some useful common attributes. Take FRY_HUMAN for an example, we found GO:0005737 from its homologous protein FRY_DROME and this GO term helps inferring the correct subcellular location, which is the cytoplasm, indicating an interesting cross-species knowledge transfer.

4.3 Subcellular locations coverage and future development

Driven by the new feature presentation protocol of HCM, Hum-mPLoc 3.0 has achieved notable improvement compared with the previous version Hum-mPLoc 2.0. On the independent HumT dataset, both *ACC* and F_1 increase around 10%. The major updates include: (i) taking into consideration feature correlation and the hierarchical structure of GO terms; (ii) extracting residue features from different segments of N- and C-terminals and (iii) use of the latest versions of gene ontology, conserved domain database and SWISS-PROT database.

One of the important future directions of current Hum-mPLoc 3.0 is how to further improve its prediction coverage and depth. The current server covers 12 human major subcellular locations, and if a protein in fact locates outside the covered 12 location classes, the output from Hum-mPLoc 3.0 server may not make any sense. To quantitatively measure the coverage of these 12 subcellular locations in the known human proteome, we collected all the 8389 human proteins with experimentally verified subcellular localizations from the SWISS-PROT database. Among them, approximately 8.3% are annotated by membrane proteins with keywords like 'membrane', 'single-pass type i membrane protein', 'multi-pass membrane protein', but are not plasma membrane (GO:0005886) proteins. This group of proteins is undergoing a clear assignment to detailed cellular compartments. We further found that 90.4% proteins fall within the 12 locations of this study. We did find that the remaining 1.3% proteins annotated with 'cell junction', 'flagellum' and 'cell projection', etc. Due to there being too few samples in these locations, we did not incorporate them into the current Hum-mPLoc 3.0 server. In addition, we checked a specialized subcellular localization database, LOCATE, that houses mouse and human proteins with annotations extracted from databases and literatures (Sprenger, *et al.*, 2008). Among the 34728 human protein entries in LOCATE, about 1.1% proteins fall out of the range of the 12 locations, which is consistent with the statistics in the SWISS-PROT database. As the number of proteins in these uncovered locations increases, we will keep updating our model accordingly.

Another future direction for updating the Hum-mPLoc 3.0 is to incorporate the sub-subcellular location prediction modules. Some cell compartments can be further grouped into functional units. For instance, the nucleus can be further classified as nuclear speckle, nucleolus, nuclear matrix, etc (Shen and Chou, 2007b). We plan to add sub-subcellular location prediction modules into our Hum-mPLoc 3.0 to strengthen its prediction depth.

5 Conclusion

Identification of protein subcellular localization is crucial for understanding protein function. Benefiting from the rapid accumulation of various annotation data, the predictors using domain knowledge for protein subcellular localization have significantly enhanced their accuracies. However, the prediction results are not always good, especially when the query protein lacks enough annotation data. Moreover, most methods directly regard each knowledge term's presence status or frequency as a feature, but neglect the structural properties of the knowledge base or relationship between terms. Therefore, domain knowledge has not been utilized sufficiently, and often the generated feature representation has very high dimensionality, which will result in low efficacy. In this study, we exploit the hidden correlations between each pair of annotation terms from the gene ontology and conserved domain database, and proposed the HCM feature extraction method. HCM provides a new strategy for more efficiently realizing the domain knowledge-based feature representation. Hum-mPLoc 3.0, which is constructed on the HCM pipeline, has shown promising performance for human protein subcellular location prediction.

Acknowledgements

We thank Dr. Jouko Virtanen for reading through the manuscript, and the anonymous reviewers for suggestions and comments which helped improving the quality of this paper. We thank Mr. Ya-Nan Wang for his assistance in the experiments.

Funding

This work was supported by the Natural Science Foundation of China (Nos. 61671288, 31628003), the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300) and the Natural Science Foundation of Shanghai (No. 16ZR1448700).

Conflict of Interest: none declared.

References

Altschul,S.F. *et al*. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Bakheet,T.M. and Doig,A.J. (2009) Properties and identification of human protein drug targets. *Bioinformatics*, **25**, 451–457.

Blum,T. *et al*. (2009) Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinf*., **10**, 1.

Boeckmann,B. *et al*. (2003) The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*., **31**, 365–370.

Boutell,M.R. *et al*. (2004) Learning multi-label scene classification. *Pattern Recognit*., **37**, 1757–1771.

Briesemeister,S. *et al*. (2010) Yloc-an interpretable web server for predicting subcellular localization. *Nucleic Acids Res*., **38**, W497–W502.

Cedano,J. *et al*. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol*., **266**, 594–600.

Chi,S.M. and Nam,D. (2012) Wegoloc: accurate prediction of protein subcellular localization using weighted gene ontology terms. *Bioinformatics*, **28**, 1028. 1030.

Chou,K.C. and Cai,Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem*., **277**, 45765–45769.

Chou,K.C. and Cai,Y.D. (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun*., **311**, 743–747.

Chou,K.C. and Shen,H.B. (2006) Hum-ploc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun*., **347**, 150–157.

Chou,K.C. and Shen,H.B. (2007) Memtype-2l: a web server for predicting membrane proteins and their types by incorporating evolution information through pse-pssm. *Biochem. Biophys. Res. Commun*., **360**, 339–345.

Chou,K.C. and Shen,H. (2010) Cell-ploc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Sci*., **2**, 1090–1103.

Chou,K.C. *et al*. (2012) iloc-hum: using the accumulation- label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst*., **8**, 629–641.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn*., **20**, 273–297.

Emanuelsson,O. *et al*. (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. Mol. Biol*., **300**, 1005–1016.

Gardy,J.L. *et al*. (2003) Psort-b: Improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic Acids Res*., **31**, 3613–3617.

Garg,A. *et al*. (2005) Support vector machine- based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem*., **280**, 14427–14432.

Hall,M.A. (1999). *Correlation-based feature selection for machine learning* (Doctoral dissertation, The University of Waikato).

Hall,M.A. and Smith,L.A. (1999). Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper. In: *FLAIRS conference*, vol. 1999, pp. 235–239.

Höglund,A. *et al*. (2006) Multiloc: prediction of protein subcellular localization using n-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.

Horton,P. *et al*. (2007) Wolf psort: protein localization predictor. *Nucleic Acids Res*., **35**, W585–W587.

Jiang,J.J. and Conrath,D.W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Lahti,J.L. *et al*. (2012) Bioinformatics and variability in drug response: a protein structural perspective. *J. R. Soc. Interface*, **9**, 1409–1437.

LaQuaglia,M.J. *et al*. (2016) Yap subcellular localization and hippo pathway transcriptome analysis in pediatric hepatocellular carcinoma. *Sci. Rep*., **6**, 30238.

Lin,D. (1998). An information-theoretic definition of similarity. In: *International Conference on Machine Learning*, vol. 98, pp. 296–304.

Marchler-Bauer,A. *et al*. (2005) Cdd: a conserved domain database for protein classification. *Nucleic Acids Res*., **33**, D192–D196.

Mei,S. (2012) Predicting plant protein subcellular multi-localization by Chou's Pseaac formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol*., **310**, 80–87.

Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci*., **11**, 2836–2847.

Nanni,L. *et al*. (2013) A comparison of methods for extracting information from the co-occurrence matrix for subcellular classification. *Expert Syst. Appl*., **40**, 7457–7467.

Park,K.J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.

Petsalaki,E.I. *et al*. (2006) Predsl: a tool for the n-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinf*., **4**, 48–55.

Pierleoni,A. *et al*. (2006) Bacello: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Psort,I. (1997) Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *J. Mol. Biol*., **266**, 594–600.

Resnik,P. *et al*. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

Savojardo,C. *et al*. (2015) Tppred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics*, **31**, 3269–3275.

Scott,M.S. *et al*. (2004) Predicting subcellular localization via protein motif co-occurrence. *Genome Res*., **14**, 1957–1966.

Shen,H.B. and Chou,K.C. (2007a) Hum-mploc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun*., **355**, 1006–1011.

Shen,H.B. and Chou,K.C. (2007b) Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Select*., **20**, 561–567.

Shen,H.B. and Chou,K.C. (2008) Pseaac: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem*., **373**, 386–388.

Shen,H.B. and Chou,K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mploc 2.0. *Anal. Biochem*., **394**, 269–274.

Small,I. *et al*. (2004) Predotar: A tool for rapidly screening proteomes for n-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.

Sprenger,J. *et al*. (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res*., **3636**, D230–D233.

Wan,S. *et al*. (2013) Goasvm: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of chou's pseudo-amino acid composition. *J. Theor. Biol*., **323**, 40–48.

Wan,S. *et al*. (2015) mlasso-hum: A lasso- based interpretable human-protein subcellular localization predictor. *J. Theor. Biol*., **382**, 223–234.

Wang,G. and Dunbrack,R.L. (2003) Pisces: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,J.Z. *et al*. (2007) A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**, 1274–1281.

Wu,H. *et al*. (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res*., **33**, 2822–2837.

Xie,D. *et al.* (2005) Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using svm and profile of psi-blast. *Nucleic Acids Res.*, **33**, W105–W110.

Yang,H. *et al.* (2012) Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, **28**, 1383–1389.

Yu,G. *et al.* (2010) Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, **26**, 976–978.

Zdobnov,E.M. and Apweiler,R. (2001) Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, **17**, 847–848.

Zhang,P. *et al.* (2006) Gene functional similarity search tool (GFSST). *BMC Bioinf.*, **7**, 1.

Zhang,S. *et al.* (2008) DBMLoc: a database of proteins with multiple subcellular localizations. *BMC Bioinf.*, **9**, 127.