Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Sequence analysis

Digging hidden feature correlation from domain knowledge base to enhance the prediction of human protein subcellular localization

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

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China
[2] Department of Computer Science, Shanghai Jiao Tong University, Shanghai, 200240, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein subcellular localization has been an important research topic in computational biology over the last two decades. Till now, a variety of computational methods have been proposed to deal with large scale data sets of proteins with unknown locations. The statistical machine learning-based approaches are a major branch of the existing predictors, and tremendous studies have shown that features extracted from biological domain knowledge can be very useful for improving the prediction accuracy. However, the domain knowledge, such as Gene Ontology and functional domain, 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 feature representation protocol denoted as HCM (Hidden Correlation Modeling). The HCM method is featured by considering structural hierarchy of the domain knowledge base and the correlations between annotation terms, so as to create more compact and discriminative feature vectors. Experimental results on four benchmark data sets show that HCM improves prediction accuracy by around 5% compared with conventional GO-based method. Moreover, the HCM-driven predictor, Hum-mPLoc 3.0, substantially enhances the accuracy for predicting multi-locational proteins by 23% on the DBMLoc data set and 16% on a newly built benchmark set.

Availability: www.csbio.sjtu.edu.cn/bioinf/Hum-mPLoc3/

Contact: hbshen@sjtu.edu.cn or yangyang@cs.sjtu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The knowledge of subcellular localization is crucial for understanding protein functions, regulation mechanisms and protein-protein interactions. Since it is often laborious and costly to identify a protein's cellular compartment using wet-lab experiment, in-silico prediction tools are of great necessity in addressing large scale data sets of proteins with unknown locations. According to SWISS-PROT (Boeckmann *et al.*, 2003) released in February 2016, among over 550,000 proteins, only 11.6% have experimentally defined and verified subcellular localization annotations, while the vast majority proteins have uncertain location annotations. Automatic tools, which allow computational prediction for the proteins of unknown locations, have been largely developed for the last decade. Especially, thanks to the abundance of protein sequences and annotation data in public databases, plentiful information, e.g. amino acid statistics, homologs with known locations, evolutional sequence information, signal peptides, physicochemical properties and prior biological knowledge, has been incorporated into the computational tools and led to powerful prediction tools. For instance, common online accessible 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) and Euk-mPLoc (Chou and Shen,

2007a) for eukaryotic proteins; and Hum-mPLoc (Chou and Shen, 2006), HSLPred (Garg *et al.*, 2005) for human proteins, etc. These web-servers have provided great convenience for biological wet-lab scientists, and faced with ever-increasing need from proteomics and related fields. Take our previously released predictor, Hum-mPLoc 2.0 (Shen and Chou, 2009), as an example, the amount of calls per year has raised from nearly 20,000 in 2010 to over 80,000 in 2015 (Supplementary Fig. S1).

Generally, computational methods for the identification of protein subcellular localization can be categorized into three types: i) homolog search-based, ii) sorting signal-based, and iii) 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 straight-forward protocol, but its performance significantly depends on the homologous targets detected (Wan *et al.*, 2013). Further, the twilight-zone phenomena also significantly challenges the hypothesis in this protocol (Gardy *et al.*, 2003), i.e. the proteins share high sequence identity could have very different structures or functions.

Targeting signal-based prediction is a type of biophysically transparent decision model. In this approach, protein sorting signals (usually in the N-terminal) are firstly identified, and then classified into different location signals according to prior knowledge, e.g. mitochondrial, chloroplast, and secretory pathway signals (PSORT, 1997; Horton *et al.*, 2007). However, the sorting signal-based method does not work for some cell compartments where our sorting knowledge has not reached, nor the query proteins with missing leading sequences.

The machine learning-based predictors are a class of flexible models, which require the so-called training data set to learn the classification rules by statistical learning algorithms. Thus, the quality of training data is one of the keys for this type of methods, and is closely related to the quality of learned statistical rules. Benefitting by more and more reliable annotations on subcellular localization of protein databases, the classification model can be trained more sufficiently through the collection of a large-scale training data. The other important issue in machine learning-based model is how to represent 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 of great importance to the final performance. In the existing machine learning tools for predicting subcellular localization, various features have been used, including:

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, 2007, 2008), Position Specific Scoring Matrix (PSSM) (Xie *et al.*, 2005; Pierleoni *et al.*, 2006; Chou and Shen, 2007b; Nanni *et al.*, 2013);

ii) The peptide-based features, such as the 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) database (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 has also brought 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 (Shen and Chou, 2009; Blum *et al.*, 2009). As GO database is expanded and updated regularly, the dimensionality will keep increasing with our expanded knowledge about the 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. 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, but 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 status or frequency of GO terms. It is well known that GO terms are organized by a hierarchical structure in three directed acrylic 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. Till now, a lot of methods for defining the semantic similarity between GO terms have been proposed, such as information content-based (Resnik *et al.*, 1999; Lin, 1998; Jiang and Conrath, 1997) 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 term correlation information. This motivates us to investigate the hidden correlation between GO terms for better measuring the similarity between two high-dimensional but sparse GO feature vectors. We propose a new protocol, called HCM (Hidden Correlation Modeling), to dig the hidden correlation between the annotation features of proteins. In order to deal with the lack of GO annotation for some query proteins, we also incorporate statistical residue features, as well as the peptide-based functional domain features, which are extracted from CDD (Conserved Domain Database).

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, but endowed with entirely new design on feature representation. In this new system, binary relevance (BR) multi-label classifier is trained to handle multi-locational proteins (Boutell *et al.*, 2004). The performance of HCM and Hum-mPLoc 3.0 has been compared with other feature extraction methods and predictors on a new test set and three other benchmark sets. The results show that HCM improves prediction accuracy by around 5% compared with conventional GO-based method, and Hum-mPLoc 3.0 notably enhances the accuracy for predicting multi-locational proteins by 23% on the DBMLoc data set and 16% on the new test set.

2 Materials and Methods

2.1 Data sets

In contrast to the vast number of computational tools for protein subcellular localization, predictors specialized for human proteins are few. Considering that knowledge of subcellular localization for human proteins is of rapidly increasing need for target identification in the drug discovery process, this study mainly focuses on human proteins. We constructed a new benchmark data set for human proteins, named HumB, by collecting all human proteins from SWISS-PROT released on January 2012. For

the sake of a high data quality, the proteins which have no subcellular localization annotation or have uncertain annotation with keywords like "by similarity", "potential" and "probable" were excluded. The data set covers 12 common locations in human cells, including centrosome, cytoplasm, cytoskeleton, endoplasmic reticulum, endosome, extracell, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome and plasma membrane. Proteins located in other compartments were filtered out, while proteins with multiple locations within these 12 categories were included. In addition, redundant sequences were removed using PISCES (Wang and Dunbrack, 2003) where the cutoff value of sequence identity is 25%. Finally, the benchmark data set 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 locations 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, whose samples were collected from a SWISS-PROT release of May 2015. Proteins already 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 controlled below 25%. For the sake of accurate assessment, we only considered the protein locations supported by experimental evidence, i.e. only the human proteins whose CC field contain "ECO:269" were collected (Evidence Codes Ontology, ECO, is a controlled vocabulary of terms that describe 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 designed for predicting human proteins' locations, the idea behind Hum-mPLoc 3.0 can be applied to other species. To demonstrate the generalization ability of the new predictor, and compare with the existing cutting-edge prediction tools, we used not only the human proteins in our data sets, but also several well-established data sets published by other researchers, including animals proteins in the BacelLo data set (Pierleoni *et al.*, 2006), animal proteins in the Höglund data set (Höglund *et al.*, 2006) and the DBMLoc data set (Zhang *et al.*, 2008). The details of these three sets are in Supplementary Materials.

2.2 Method

This study aims to develop a machine learning-based predictor for subcellular localization of human proteins. Fig. 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 features are given in Sections 2.2.1, 2.2.2 and 2.2.3, respectively.

2.2.1 Residue-based feature

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

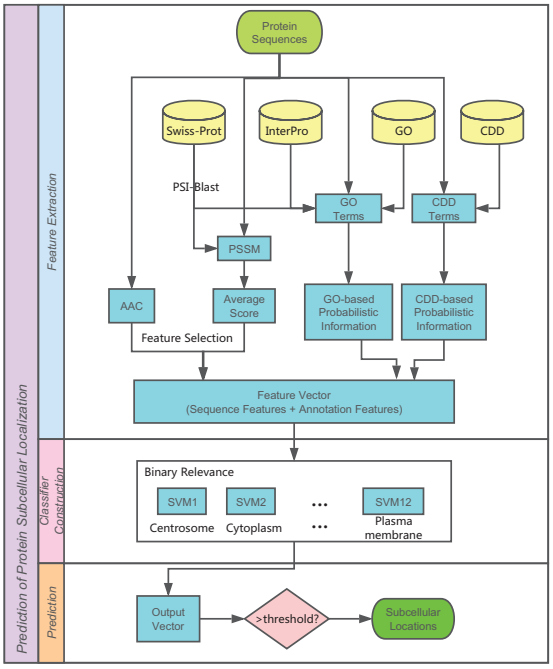


Fig. 1. Flowchart of the new predictor

$$S_{\text{PSSM}} = \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,20} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ S_{i,1} & S_{i,2} & \cdots & S_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ S_{L,1} & S_{L,2} & \cdots & S_{L,20} \end{pmatrix}, \tag{1}$$

where $S_{i,j}$ represents the original score that the residue at the ith position of the sequence mutates to the jth amino acids ($1 \leq j \leq 20$) during the evolution process, and L represents the length of the protein sequence.

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

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

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 Equ. 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-Dim vector in Equ. 4,

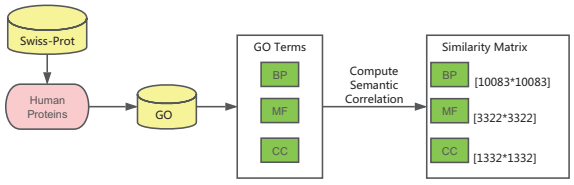
$$\overline{S_{\text{PSSM}}} = [\overline{S^0_1}, \overline{S^0_2}, \overline{S^0_3}, \ldots, \overline{S^0_{20}}]. \tag{4}$$

Then, AAC and the normalized PSSM vector are combined as a 40-Dim vector, which catches not only amino acid frequency information of the protein itself, but also the residue statistics from its functional related homologs.

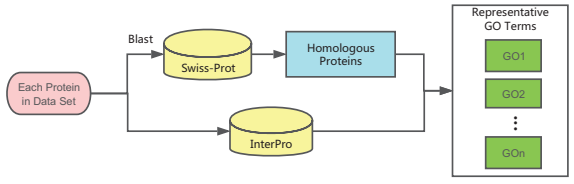
Furthermore, considering that localization information is often implied in the N-terminal and C-terminal of amino acid sequences (Pierleoni *et al.*, 2006), we extract sequence features of multiple segments from both terminals, specifically, the first $10, 20, \ldots, 60$ residues of N-terminal, and the last $10, 20, \ldots, 100$ residues of C-terminal. For each segment, a 40-Dim vector is created using the method described above. By concatenating all these 40-Dim vectors (for the full sequence and 16 segments), the total dimensionality becomes $680(40 \times 17)$, which may contain some redundant information. Thus, the Correlation-based Feature Selection (Hall and Smith, 1999) (CFS) method is adopted and finally lead to a 43-Dim feature vector.

2.2.2 GO annotation-based feature

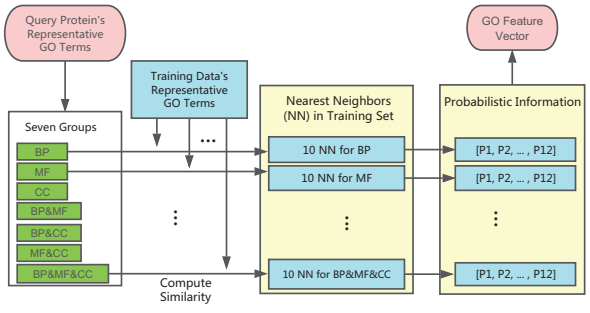
The feature extraction procedure consists of three parts as shown in Fig. 2. Before digging the correlation information, we need a matrix of pairwise GO similarities. It is extremely costly to construct such a matrix with a total of over 40,000 terms in GO database. Thus we only use GO terms annotated 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, including 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.



(a) Construct similarity matrices



(b) Search representative GO terms



(c) Produce feature vectors

Fig. 2. Flowchart of GO feature extraction

In Fig. 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 as the following two steps.

1. Search nearest neighbors for query proteins.
   Intuitively, a query protein would have 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 term set of the training protein, $\mathcal{K}$, as shown in Equ. 5,

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

   where the correlation between x_i (the ith representative GO term of the query protein) and $\mathcal{K}$ is defined in Equ. 6,

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

   where y_js are GO terms, and $\mathcal{K} = \{y_1, y_2, \cdots, y_m\}$.

   According to the similarity measurement, query protein's nearest neighbors in the training set can be identified. Because 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 divide 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 10 nearest neighbors are selected in each of the 7 groups, respectively.
2. Generate probabilistic information.
   In this step, feature vectors are expressed by probabilistic information. Let pro_a denote the probability of the query protein being in the location a. Initially, pro_a is defined as the ratio that the sum of similarities between the query protein and its nearest neighbors which locate at a to the sum of similarities between the query protein and all of its 10 nearest neighbors, as shown in Equ. 7,

$$pro_a = \frac{\sum_{j \in \mathcal{I}_{N_a}} sim_j}{\sum_{i \in \mathcal{I}_N} sim_i}, \tag{7}$$

   where $\mathcal{I}_N$ is the index set of all the nearest neighbors of the query protein ($|\mathcal{I}_N| = 10$), and $\mathcal{I}_{N_a}$ is the index set of the nearest neighbors which locate at a.

   However, due to the lack 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 Equ. 8. The prior is equal to the proportion of proteins locating at a.

$$pro_a = \frac{\sum_{j \in \mathcal{I}_{N_a}} sim_j + \frac{num_a}{num}}{\sum_{i \in \mathcal{I}_N} sim_i + 1}, \tag{8}$$

   where num_a and num are the number 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-Dim vector is consisted of the probabilistic information for all 12 locations. Finally, a feature vector with 84-Dim is generated. In order to produce probabilistic information for training proteins, 10-fold cross validation is conducted.

2.2.3 Peptide-based feature

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

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

Unlike GO terms, CDD terms have no semantic structure. Here we adopt a symmetrical uncertainty method (Press *et al.*, 1996) to construct the pairwise correlation matrix for CDD features. Firstly, a 3129×2313-Dim binary matrix is constructed. The element on 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 (Equ. 9),

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

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, $H(f_j^{cdd}, f_i^{cdd})$ is the differential entropy of the ith feature and jth feature (Equ. 10),

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

The correlation between two CDD terms is defined in Equ. 11,

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

The following steps are the same as GO feature extraction. We use PRS-BLAST to extract CDD terms for query proteins, compute similarities of query proteins with training proteins based on the matrix $[S_{ij}^{cdd}]_{2313\times2313}$, find nearest neighbors, and generate probabilistic information as a 12-Dim feature vector corresponding to the 12 locations.

Finally, residue-based features, GO and CDD features are combined as a total of 139 (43+12×7+12) dimensional feature vector.

2.2.4 Multi-label classification

For the classification system, there are 12 class labels corresponding to 12 subcellular locations. By using support vector machines (Cortes and Vapnik, 1995) as classifiers and the binary relevance strategy (Boutell *et al.*, 2004), 12 binary classifiers are trained with optimal γ and C parameters searched by 10-fold cross validation. In the test phase, the output for each test sample is a 12-Dim 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 of 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, ACC and F_1, are used to evaluate the performance (Briesemeister *et al.*, 2010). These two measures are customized for multi-label classification, thus they are 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 Compare different types of features

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 $\mathrm{Sim}_k = 1-\mathrm{hit}_k/\sqrt{\mathrm{num}_{\mathrm{query}}\times\mathrm{num}_k}$, where hit_k denotes the number of common GO features of these two proteins, $\mathrm{num}_{\mathrm{query}}$ and num_k are the numbers of GO terms of the query protein and the kth protein, respectively. Ten nearest neighbors are used to calculate the probability information.
- SEQ: residue features only, i.e. the first part of HCM.

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

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 prediction. Especially on HumT, the accuracy is below 50%. Compared with SEQ+GO$_0$, SEQ+GO$_1$ performs better on all four data sets, with an increase of $2\%\sim5\%$ on ACC and $2\%\sim13\%$ on F_1. As for SEQ+GO$_7$, it seems that the 7 groups contain redundant information, but SEQ+GO$_7$ achieves apparently better performance than SEQ+GO$_1$. It may be due to two reasons. One reason 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, has slight improvement than 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 Compare variants of GO-based feature extraction methods

In this section, some variants of GO-based features in HCM are assessed. The variants include methods using different similarity definition, and using different types or sources of GO terms. The results are discussed as below.

3.2.1 Strategies for computing similarity between proteins

There are multiple choices to obtain a similarity between two genes according to the similarities of their GO terms. Two typical methods are MAX and BMA (best match average) (Yu *et al.*, 2010). The former method chooses the maximum similarity of two GO terms from two genes respectively. This method only considers the most similar pair of GO

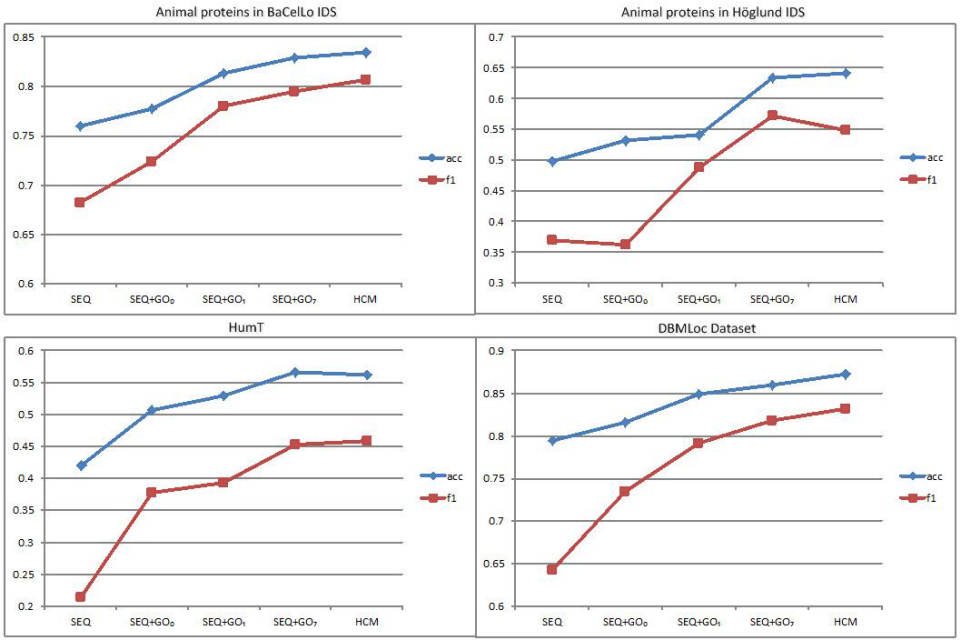


Fig. 3. Prediction accuracies of different types of features on four data sets

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, P42772 is annotated by GO:0005737 and GO:0005634, which suggest two locations of cytoplasm and nucleus, while the MAX strategy only leads to one location. The BMA strategy has been more widely used. It firstly 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 thses maximum values. BMA treats each GO with equal weight, but since GO terms have different information content, we think the GO terms with higher information content should have more weights.

In 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. (Bottom GOs have high information content.) In other words, the similarity value itself can be used as weight. Thus, we adopted the Euclidean distance-based metric to measure the similarity between proteins. For instance, homologous proteins of the pair of Q5T6F0 and Q6PEY1 both contain and only contain the GO term, GO:0005515, which has a low information content. Because this term represents the function of protein binding, and a lot of proteins have this function. In this case, although these two genes have exactly the same GO sets, they are not assigned a high similarity value using our method.

Furthermore, to verify the superiority to other strategies, we compared our metric implemented in HCM with MAX and BMA on the new test set. 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. S2).

3.2.2 Sources of GO terms

We also conducted experiments to assess the performance of GO features with and without CC terms (Supplementary Fig. S3). The results show that without any CC term, the BP and MF terms can improve accuracy by around 10% compared with residue-based features. Also, BP and MF terms can achieve comparable performance to CC terms with close ACC and slightly worse F_1. Moreover, using all the three types of GO can improve both ACC and F_1 by nearly 10% compared with using CC alone.

Table 1. Comparison of seven predictors on four data sets

	ACC/F_1			
	BaCelLo	Höglund	HumT	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	0.45/0.34	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	-	-	-
Hum-mPLoc 3.0	0.83/0.81	0.64/0.55	0.61/0.55	0.87/0.83

YLoc-LowRes, MultiLoc2- LowRes and BaCelLo only predict globular proteins;
YLoc+ and Hum-mPLoc 3.0 can deal with multiple-locational proteins;
YLoc-HighRes and MultiLoc2-HighRes can predict nine subcellular locations.

These results demonstrate that BP and MF terms are also informative in identifying protein subcellular localization.

In order to ensure the reliability of annotation data, we only used the GO terms with experimental evidence, but we also implemented another version using all GO terms including both experiment-supported and computationally-inferred ones. Experimental results show that the latter version indeed has more coverage and increases by 5% on ACC and 18% on F_1 (Supplementary Fig. S4), but whether the inferred GOs would bring bias is unclear. Thus we keep both of these two versions in our prediction tool.

3.3 Compare with other predictors

Table 1 shows a comparison of the HCM driven predictor, Hum-PLoc 3.0, 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 on all data sets. For the two mono-locational data sets, BacelLo and Höglund, the ACCs of the new method are 4% and 7% higher than the best ACCs obtained by other predictors,

Table 2. Comparison of YLoc+ and Hum-mPLoc 3.0 on HumT

Location	YLoc+			Hum-mPLoc 3.0a			Hum-mPLoc 3.0b		
	pre	rec	F1	pre	rec	F1	pre	rec	F1
Centrosome	-	-	-	0.33	0.05	0.08	0.75	0.55	0.63
Cytoplasm	0.55	0.85	0.67	0.70	0.75	0.73	0.76	0.73	0.74
Cytoskeleton	-	-	-	0.83	0.24	0.38	0.8	0.68	0.74
ER*	0.71	0.12	0.21	0.72	0.32	0.44	0.83	0.37	0.51
Endosome	-	-	-	1	0.07	0.13	0.58	0.47	0.52
Extracellular	0.39	0.85	0.54	0.70	0.54	0.61	0.5	0.46	0.48
Golgi apparatus	0.1	0.05	0.07	0.75	0.3	0.43	0.69	0.45	0.55
Lysosome	0	0	0	0.5	0.13	0.20	0.71	0.63	0.67
Mitochondrion	0.65	0.43	0.52	0.84	0.68	0.75	0.78	0.75	0.76
Nucleus	0.41	0.57	0.48	0.58	0.74	0.65	0.75	0.71	0.73
Peroxisome	0.07	0.5	0.13	1	0.5	0.67	1	1	1
Plasma membrane	0.41	0.44	0.42	0.55	0.38	0.45	0.65	0.44	0.52

∗ Endoplasmic reticulum.
Hum-mPLoc 3.0a: use only experimental-supported GO terms.
Hum-mPLoc 3.0b: use all GO terms.
pre denotes precision, and rec denotes recall.

respectively. For the two multi-locational data sets, DBMLoc and HumT, the improvement is more significant. Compared with YLoc+, the ACC increases by 16% and 23%, and the F_1 increases by 21% and 15%, on HumT and DBMLoc, respectively. In addition, to further investigate the prediction performance on each location, we compare precision, recall and F_1 obtained by YLoc+ and two versions of Hum-PLoc 3.0 on all 12 locations (YLoc+ can only predict 9 locations) in Table 2. Hum-PLoc 3.0 has an overall enhancement in differentiating the 12 subcellular locations. It achieves the increase of over 20% on five locations, ER, Golgi apparatus, lysosome, mitochondrion and peroxisome. Especially, for the three small classes, lysosome, peroxisome and Golgi apparatus, Hum-PLoc 3.0 has very significant improvement, 20%∼54% on F_1 by Hum-PLoc 3.0a. In addition, Hum-PLoc 3.0b which uses all GO terms has generally higher accuracies than Hum-PLoc 3.0a except on extracellular, where the F_1 drops 13%.

The great improvement compared with YLoc may be mainly due to the hidden correlation-based method as well as renewed annotation database. YLoc adopts conventional binary coding to express GO features, and to avoid high dimensionality, it only considers the annotation-based features which directly related with protein localization on certain compartment. However, the annotation terms without any indication of localization may also help. Besides, the annotation databases have been updated rapidly, our method uses the latest version of gene ontology and conserved domain database, which have more coverage than the old versions used by previous predictors.

4 Discussion

In the proposed method, feature vectors consist of both residue-based and prior knowledge-based features. The latter often shows higher quality and results in better predictions. However, statistical residue features also play an indispensable role for the prediction task, especially when the annotation data is incomplete or unreliable. The human protein NPIP, for example, its homologous proteins have very few annotations, with only one GO term, GO:0005505. Thus residue-based features play the leading role. Another example is protein MTO1, with GO terms of GO:0044822, GO:0008033, GO:0050660 and GO:0002098. By using these GO terms, a prediction result of cytoplasm is obtained with low confidence, which is a wrong answer; while residue-based features predict the correct protein location of mitochondrion with high confidence. These examples suggest that residue-based features are very essential in the prediction of protein subcellular localization, and can function against bias induced by incomplete annotation data.

Most of the existing feature extraction methods adopt binary coding to represent the presence status of annotation terms, while the newly proposed HCM incorporates correlations of annotation features, because of the fact that a highly similar pair of proteins usually has not exactly the same but highly correlated GO terms or conserved domains. Although the semantic similarity between GO terms has been a well-studied subject that various metrics have been proposed, it has seldom been used in the construction of features vectors for protein classification problems. 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 HumT using only human proteins as neighbors. The results show obvious drop on accuracies, with 5.8% on ACC and 7.5% of F_1 (Supplementary Fig. S4). It indicates that the homologs in different species also share the same attributes. Take FRY_HUMAN for example, we found GO:0005737 from its homologous protein FRY_DROME and this GO term inferred that this protein locates in cytoplasm.

Driven by the new feature presentation protocol of HCM, we release the latest version of our human protein multi-localization predictor, Hum-mPLoc 3.0. Compared with the previous version Hum-mPLoc 2.0, it achieves notable improvement. Specifically, both ACC and F_1 increase around 10% on HumT data set. The major updates on developing the method include: i) consider feature correlation and hierarchy structure of GO; ii) extract residue features from different segments of N- and C-terminals; iii) use latest version of gene ontology, conserved domain database and SWISS-PROT database.

5 Conclusion

Identification of protein subcellular localization is very crucial for understanding protein functions. Researchers in the field of bioinformatics have put a lot of efforts in the design of computational tools for this task, but there are still some challenges on improving the performance of the current predictors to truly assist related protein research in biological labs. Two major challenges are: i) How to effectively incorporate prior biological knowledge into the prediction tools; ii) How to deal with proteins with multiple locations.

Benefiting by the rapid accumulation of various annotation data, the predictors using biological 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. Moreover, most methods directly regard each knowledge term's presence status or frequency as a feature, but neglect the structural properties of knowledge base or relationship between terms. Therefore, the domain knowledge has not been utilized sufficiently, and the resulted feature representation has high dimensionality and low efficiency.In the study, we dig hidden correlation between each pair of annotation terms from the gene ontology and conserved domain database, and proposed the HCM feature extraction method. HCM can address the lack of annotation. Besides the consideration of correlation, another difference between the proposed method and other methods is that we generate probabilistic information from query protein's nearest neighbors to express feature vectors. The dimensionality is much condensed compared with conventional GO-based feature representation.

The HCM-driven predictor, named Hum-mPLoc 3.0, is designed to handle multi-locational human proteins. In order to systematically evaluate its performance, a series of experiments have been conducted

on previous published data sets as well as new data sets collected by ourselves. We compared the performance of each type of features in HCM, and also compared Hum-mPLoc 3.0 with existing established predictors. These experiments show substantial enhancement on prediction accuracy obtained by the new predictor, demonstrating that the new predictor has high-quality features, and sufficient capability to predict proteins with multiple locations. With the rapid increasing need for computational identification of protein subcellular localization, the latest Hum-mPLoc, will continue to contribute to proteomic researches and provide much more accurate predictions.

Funding

This work has been supported by the Science and Technology Commission of Shanghai Municipality (No. 16JC1404300), and the Shanghai Municipal Natural Science Foundation (No. 16ZR1448700).

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., *et al.* (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, **25**(17), 3389–3402.

Ashburner, M., Ball, C. A., Blake, J. A., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, **10**(1), 1.

Boeckmann, B., Bairoch, A., Apweiler, R., *et al.* (2003). The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, **31**(1), 365–370.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. (2004). Learning multi-label scene classification. *Pattern recognition*, **37**(9), 1757–1771.

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Yloc[a]an interpretable web server for predicting subcellular localization. *Nucleic acids research*, **38**(suppl 2), W497–W502.

Cedano, J., Aloy, P., Perez-Pons, J. A., and Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *Journal of molecular biology*, **266**(3), 594–600.

Chi, S.-M. and Nam, D. (2012). Wegoloc: accurate prediction of protein subcellular localization using weighted gene ontology terms. *Bioinformatics*, **28**(7), 1028–1030.

Chou, K.-C. and Cai, Y.-D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, **277**(48), 45765–45769.

Chou, K.-C. and Cai, Y.-D. (2003). A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochemical and biophysical research communications*, **311**(3), 743–747.

Chou, K.-C. and Shen, H.-B. (2006). Hum-ploc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochemical and biophysical research communications*, **347**(1), 150–157.

Chou, K.-C. and Shen, H.-B. (2007a). Euk-mploc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *Journal of proteome research*, **6**(5), 1728–1734.

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

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, **300**(4), 1005–1016.

Gardy, J. L., Spencer, C., Wang, K., *et al.* (2003). Psort-b: Improving protein subcellular localization prediction for gram-negative bacteria. *Nucleic acids research*, **31**(13), 3613–3617.

Garg, A., Bhasin, M., and Raghava, G. P. (2005). Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *Journal of biological Chemistry*, **280**(15), 14427–14432.

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*, volume 1999, pages 235–239.

Höglund, A., Dönnes, P., Blum, T., *et al.* (2006). Multiloc: prediction of protein subcellular localization using n-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**(10), 1158–1165.

Horton, P., Park, K.-J., Obayashi, T., *et al.* (2007). Wolf psort: protein localization predictor. *Nucleic acids research*, **35**(suppl 2), W585–W587.

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

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., *et al.* (2005). Cdd: a conserved domain database for protein classification. *Nucleic acids research*, **33**(suppl 1), D192–D196.

Mei, S. (2012). Predicting plant protein subcellular multi-localization by chou's pseaac formulation based multi-label homolog knowledge transfer learning. *Journal of theoretical biology*, **310**, 80–87.

Nair, R. and Rost, B. (2002). Sequence conserved for subcellular localization. *Protein Science*, **11**(12), 2836–2847.

Nanni, L., Brahnam, S., Ghidoni, S., Menegatti, E., and Barrier, T. (2013). A comparison of methods for extracting information from the co-occurrence matrix for subcellular classification. *Expert systems with applications*, **40**(18), 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**(13), 1656–1663.

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2006). Bacello: a balanced subcellular localization predictor. *Bioinformatics*, **22**(14), e408–e416.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1996). *Numerical recipes in C*, volume 2. Citeseer.

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.

Scott, M. S., Thomas, D. Y., and Hallett, M. T. (2004). Predicting subcellular localization via protein motif co-occurrence. *Genome research*, **14**(10a), 1957–1966.

Shen, H.-B. and Chou, K.-C. (2007). Hum-mploc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and biophysical research communications*, **355**(4), 1006–1011.

Shen, H.-B. and Chou, K.-C. (2008). Pseaac: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry*, **373**(2), 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. *Analytical biochemistry*, **394**(2), 269–274.

Wan, S., Mak, M.-W., and Kung, S.-Y. (2013). 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**, 40–48.

Wang, G. and Dunbrack, R. L. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, **19**(12), 1589–1591.

Wang, J. Z., Du, Z., Payattakool, R., *et al.* (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics*, **23**(10), 1274–1281.

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. (2005). Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic acids research*, **33**(9), 2822–2837.

Xie, D., Li, A., Wang, M., Fan, Z., and Feng, H. (2005). Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using svm and profile of psi-blast. *Nucleic acids research*, **33**(suppl 2), W105–W110.

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, **28**(10), 1383–1389.

Yu, G., Li, F., Qin, Y., *et al.* (2010). Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, **26**(7), 976–978.

Zdobnov, E. M. and Apweiler, R. (2001). Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, **17**(9), 847–848.

Zhang, P., Zhang, J., Sheng, H., *et al.* (2006). Gene functional similarity search tool (gfsst). *BMC bioinformatics*, **7**(1), 1.

Zhang, S., Xia, X., Shen, J., *et al.* (2008). Dbmloc: a database of proteins with multiple subcellular localizations. *BMC bioinformatics*, **9**(1), 127.