Bioinformatics Advance Access published September 11, 2015

Bioimage infomatics

Human Cell Structure-driven Model Construction for Predicting Protein Subcellular Location from Biological Images

Wei Shao, Mingxia Liu and Daoqiang Zhang*

School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Associate Editor: Prof. Robert Murphy

ABSTRACT

Motivation: The systematic study of subcellular location pattern is very important for fully characterizing the human proteome. Nowadays, with the great advances in automated microscopic imaging, accurate bioimage based classification methods to predict protein subcellular locations are highly desired. All existing models were constructed on the independent parallel hypothesis, where the cellular component classes are positioned independently in a multi-class classification engine. The important structural information of cellular compartments is missed. To deal with this problem for developing more accurate models, we proposed a novel cell structure-driven classifier construction approach (SC-PSorter) by employing the prior biological structural information in the learning model. Specifically, the structural relationship among the cellular components is reflected by a new codeword matrix under the error correcting output coding (ECOC) framework. Then, we construct multiple SC-PSorter based classifiers corresponding to the columns of the ECOC codeword matrix using a multi-kernel support vector machine (SVM) classification approach. Finally, we perform the classifier ensemble by combining those multiple SC-PSorter based classifiers via majority voting.

Results: We evaluate our method on a collection of 1636 immunohistochemistry images from the Human Protein Atlas (HPA) database. The experimental results show that our method achieves an overall accuracy of 89.0%, which is 6.4% higher than the state-of-the-art method.

Availability: The dataset and code can be downloaded from https://github.com/shaoweinuaa/

Contact: dqzhang@nuaa.edu.cn

1. INTRODUCTION

One important task in the research of proteomics is to explore the natural function of proteins in performing and regulating the activities of an organism at cell level (Breker and Schuldiner, 2014). It is widely recognized that the function of a protein is closely associated with its corresponding cellular compartments (Chebira, et al., 2007). Proteins can only find their correct interacting molecules at the right place. Thus, subcellular location can provide important clues for understanding the function of a protein. With the breakthrough of genome sequencing and bioimaging techniques, traditional time-consuming and expensive wet-lab experimental approaches cannot catch up with the speed of newly known proteins (Zhang, et al., 2006). Hence, finding an automatic computational way to determine the subcellular locations of proteins has been becoming a focus in computational biology (Glory and Murphy, 2007). From the perspective of machine learning, this task can be transformed into a multi-class or multi-label classification problem. This is a two-step framework, where we first need to figure out a proper feature representation way for encoding the protein data, which then will be fed into a trained machine learning model for label decision. There are two major research categories depending on how the protein data are represented, i.e. 1D-amino sequence and 2D-image (Xu, et al., 2013).

On one hand, if a protein is represented in amino acid sequence, PseAAC (Chou, 2001), PSSM (Jeong, et al., 2011; Pierleoni, et al., 2011), and gene ontology(Chi, 2010) are among the applied sequence-based features. In the 2nd step, various machine learning algorithms have been proposed. For instance, researchers in (Yoon and Lee, 2012) adopted a boosting framework to accomplish the classification task, and in (Wang and Li, 2013), a random label selection (RALS) method was presented to learn the label correlations from training dataset to guide the classification for multi-label proteins.

On the other hand, accompanied with the explosive increments in genomic data, we witnessed great advances in automated microscopic imaging in recent years (Peng, et al., 2012). Due to the intuitive characteristics of images compared to amino acid sequence, bioimage based protein subcellular distribution pattern analysis has attracted much attention. For example, it's found that image-based analysis can be successfully used to detect protein biomarkers, which will dynamically change their subcellular locations in the cancerous tissues (Kumar, et al., ; Xu, et al., ; Xu, et al., 2013).

If proteins are represented with 2D-images, e.g., through fluorescent or immunohistochemistry microscopy, the most widely used image features can be grouped into two categories, i.e., global and local features. For global feature, the DNA feature (Boland and Murphy, 2001) is designed to characterize DNA distribution in a cell image. Since there is high co-occurrence of protein and DNA in a protein image, we can infer the relative position of protein according to the DNA distribution. Besides, Haralick feature based on db wavelet is another global feature to describe image texture such as inertia and isotropy, which is demonstrated to be robust to cell rotations and translations (Murphy, et al., 2003) (Newberg and Murphy, 2008). As to local feature, LBP (Ojala, et al., 1996) feature is the most frequently applied descriptor to characterize the spatial structure of images involving flat areas, edges and spots. Some extensions are also reported. Yang et al. (Yang, et al., 2014) constructed a mixed local feature set by adding two extensive forms of LBP, i.e., LTP (Tan and Triggs, 2010) and LQP (Nanni, et al., 2010). Luis et al.(Coelho, et al., 2013) applied the SURF feature to handle the classification problem in cell images.

Considering different features will have their own advantages, a common strategy is to fuse multiple types of features. For instance, different features are concatenated as a long vector to perform the subsequent classification task (Xu, et al., 2013)(Newberg and Murphy, 2008)(Yang, et al., 2014). Intuitively, since single type of features cannot reflect all the information of a protein image, fusing multiple types of features together is expected to be a more promising way.

For learning algorithms design, Boland et al.(Boland, et al., 1998) applied neural networks to classify 4 protein types; Yang et al.(Yang, et al., 2014) proposed a probability-based support vector machine (SVM) to predict the subcellular location of proteins in human reproductive system. By considering a high ratio of human proteins co-exist at different locations, Xu et al. (Xu, et al., 2013) designed a multi-label classification classifier. Other efforts include Chebira et al. (Chebira, et al., 2007) used a multi-resolution approach, and Logistic regression algorithm with latent variables was proposed in (Li, et al., 2012).

Although much progress has been achieved in designing different statistical classifiers, to the best of our knowledge, none of the existing image-based classifiers takes the biological cell structure information into consideration, which has already been demonstrated to be effective in solving biological sorting problems (Lin, et al., 2011). The basic hypothesis of existing predictors is to parallelly consider every cellular component class regardless of their organizations in the cell. It's expected that better performance will be achieved when we incorporate the cell nature component organization structure into the model construction.

To enable the learned model to incorporate the subcellular component organization structure, we propose a new classifier learning approach by utilizing the error correcting output coding (ECOC) framework (Dietterich, et al., 1995). This new approach can decompose the multi-class problem into several binary classification problems according to the prior human cell structural information. The final decision will then be made by combining the results of these binary classifiers. In the new structure-driven learning approach, we firstly construct a codeword matrix to reflect the biological structure of cellular compartments with ECOC. Then, for each binary classifier corresponding to the columns of the ECOC codeword matrix, we use kernel combination method to fuse different types of features rather than the direct combination strategy. Finally, we perform the classifier ensemble by combining multiple classifiers via majority voting. The experimental results show that our method performs much better than several state-of-the-art methods, because the proposed approach has incorporated the cell structure prior knowledge into model generation.

2. MATERIALS AND METHODS

2.1 Dataset

Starting from 2005, the researchers use the antibody-based technique to functional study of the human proteome and build the well-known Human Protein Atlas (HPA) database (Uhlen, et al., 2005). In the recent 13th version of HPA database, 86%of human genome is involved. Specifically, 16975 genes with 24028 antibodies have been covered to 46 different normal human tissues and 20 different cancer types.

In this study, we have generated a collection of 1636 immunohistochemistry images with high validation and objective scores (Ponten, et al., 2008) from HPA as our benchmark dataset. It contains 21 proteins related to 46 normal human tissues. And each image belongs to one of the seven most frequently appeared subcellular locations, namely, cytoplasm, golgi apparatus, mitochondrion, vesicles, nucleus, endoplasmic reticulum and lysosome. Table 1 summarizes the distribution of our dataset.

Table 1. The distribution of the benchmark dataset.

Category	*Size*
cytoplasm	391
golgi apparatus	228
mitochondrion	319
vesicles	139
nucleus	183
endoplasmic reticulum	216
lysosome	160
total	1636

2.2 Overview of Our Method

Figure 1 shows the flowchart of our method, which consists of four major steps. Firstly, we extract and select features from the given protein images. Then, we use ECOC method to transform the multi-class classification problem into a series of binary classification sub-problems according to a pre-defined codeword matrix. Here, the codeword matrix is comprised of 14 bits. The first 6 bits are designed according to the biological structure of cellular compartments, which can bring more prior information to the learning process. And the other 8 checking bits are used to strengthen the error-correcting ability of this ECOC-based model. Next, since db wavelet was employed to get the multi-resolution global feature (i.e., Haralick feature), we can construct 10 different SC-PSorter models based on different sets of features extracted from 10 vanishing moments of db wavelets. Moreover, for each SC-PSorter model, we construct 14 multi-kernel based SVM classifiers corresponding to the columns of the ECOC codeword matrix. Finally, we perform the classifier ensemble by combining those 10 SC-PSorter based classifiers via majority voting.

2.3 Feature Extraction and Selection

For protein images, different types of features (i.e., global and local features) are expected to provide complementary information (Yang, et al., 2014). Therefore, we are encouraged to use both of the two types of features to describe protein images. Specifically, as to global feature, we select Haralick feature with 10 different vanishing moments from 1 to 10, then for each vanishing moment, an 836-dimensional feature can be obtained. In addition, a 4-dimensional global DNA feature is also incorporated due to their values in inferring the relative position of protein. As to local feature, we prefer to choose the most widely used LBP feature, which is constituted by a 256-dimensional vector. Hence, for every protein image, we can get a 1096-dimensional descriptor for each db wavelet if we directly combine them together. After that, to reduce computational time cost and avoid overfitting, we select the most distinguishing features by applying the stepwise discriminant analysis (SDA) method (Huang, et al., 2003).

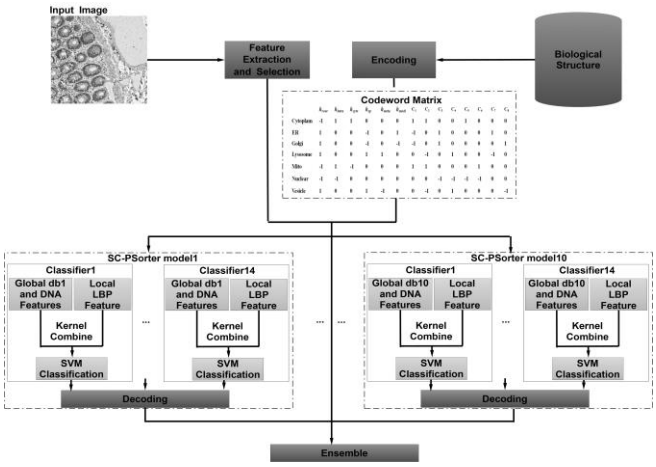

Fig. 1. The flowchart of our proposed method.

2.4 Error Correcting Output Coding (ECOC)

In this paper, our goal is to determine the subcellular location that a protein image belongs to. Since there are seven subcellular locations, this problem can be regarded as a multi-class classification problem. Nowadays, multi-class classification is an important issue in many machine learning domains, such as text classification(Nigam, et al., 2000), and medical analysis(Lu, et al., 2005). There are two main lines to deal with such multi-class learning problems, including "direct multi-class representation" and "(indirect) decomposition design". The first line aims to design multi-class classifiers directly, such as neural network (Boland, et al., 1998), and multi-class support vector machines (SVM)(Yang, et al., 2014). In contrast, the second line endeavors to firstly transform the original multi-class problem into several binary classification problems, and then to combine the results of these binary classifiers for making final decision. As a typical indirect decomposition way to deal with multi-class problems, error correcting output coding (ECOC) (Dietterich, et al., 1995)(Liu, et al., 2015) is one of the representative methods. Specifically, there are three main steps in ECOC-based classification system, including 1) encoding, which decomposes the original problem into several binary classification problems; 2) binary classifier learning; and 3) decoding, which makes a final decision based on the outputs of those binary classifiers. In the following, we will introduce each step in detail.

2.4.1 *Encoding*

In the encoding procedure, a codeword matrix $\mathbf{M}_{k\times l}$ is employed to decompose the original multi-class problem into several binary sub-problems. Here, the *r*-th $(\mathrm{r}=1,2\ldots\mathrm{k})$ row of $\mathbf{M}$ (i.e., $\mathbf{M}_r$) represents the codeword of the *r*-th class, while each column of $\mathbf{M}$ denotes the new class label vector for each of original classes. The elements in each column of the codeword matrix can be set as -1,0, and 1 in ternary ECOC encoding methods, and -1 and -1 in binary ECOC encoding methods (Pujol, et al., 2006). Below we briefly introduce two typical ECOC encoding strategies including One-Vs-All coding and the Forest coding (Escalera, et al., 2007).

1) One-Vs-All Coding

In this approach, k different binary classifiers are built, each of which learns to distinguish one class versus the others. In the codeword matrix, all of the diagonal elements are set as 1, while the others as -1. In Eq. (1), we show a codeword matrix $\mathbf{M}_{k\times k}$ using the widely used one-vs-all coding strategy, which transforms the *k*-class classification problem into *k* one-vs-all binary classification problems:

$$\mathbf{M}=\begin{pmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ \ldots & \ldots & \ldots & \ldots \\ -1 & -1 & -1 & 1 \end{pmatrix}_{k\times k} \quad (1)$$

2) Forest Coding

In this data-dependent coding strategy, the codeword matrix is completely determined by the partition of the dataset by using the decision tree algorithm. Here, when building decision tree, each node corresponds to the best bi-partition of the set of classes by maximizing the mutual information between different types of samples. The process is recursively applied until sets of single classes corresponding to the tree leaves are obtained.

2.4.2 Binary Classifier Learning

The second step is to train multiple binary classifiers based on the codeword matrix $\mathbf{M}$. Specifically, a binary classifier is corresponding to a specific column of the codeword matrix, where the samples labeled as 1 are used as positive instances, and samples labeled as -1 are regarded as negative instances. It is worth noting that those instances labeled as 0 will not be used for training the classifier in ternary encoding methods. Given a codeword matrix $\mathbf{M}$ that contains *l* columns, we then learn a total of *l* binary classifiers. In the literature, these binary classifiers are usually directly taken from many existing classifiers (e.g. SVM) (Escalera, et al., 2007).

2.4.3 Decoding

In the decoding stage, given a new instance, we firstly calculate the output vector from the multiple binary classifiers, i.e., $\mathrm{H}(\mathbf{z})=(\mathrm{h}_1(\mathbf{z}),\mathrm{h}_2(\mathbf{z}),\ldots,\mathrm{h}_l(\mathbf{z}))$. Then, we need to find a decoding method that can transform the output vector into a specific target class label, through which the original multi-class classification problem can be solved ultimately. Currently, there are many decoding methods, such as Hamming distance (HD) decoding, Euclidean distance (ED) decoding, and linear-loss weighted (LLW) decoding (Escalera, et al., 2010). Among various decoding strategies, Hamming distance is one of simple and effective decoding method. Accordingly, we use the Hamming distance to perform ECOC decoding in this paper. Then, the predicted class label y of a testing instance $\mathbf{z}$ can be estimated through the following

$$y=\arg\min_r \sum_{i=1}^{l}\left|h_i(\mathbf{z})-\mathbf{M}_{r,i}\right| \; r=1,\ldots,k \quad (2)$$

where the testing sample $\mathbf{z}$ is assigned to label *r* if the output vector $\mathrm{H}(\mathbf{z})$ is much closer to the *r*-th row of the codeword matrix $\mathbf{M}$ in terms of Hamming distance, when comparing with the other rows of $\mathbf{M}$.

2.5 ECOC Coding with Biological Structural Information

As mentioned in Section 2.4, ECOC-based methods transform the multi-class classification problem into a series of binary classification sub-problems according to a pre-defined codeword matrix. Different designs of codeword matrix may lead to different partitions of original classes, which will affect the classification performance. Hence, the design of the codeword matrix is important for ECOC-based methods. On the other hand, it is highly recognized that the biological structural information (Lin, et al., 2011) plays a crucial role in determining protein subcellular location. Intuitively, such structural information can be used to guide the codeword matrix design, which can bring more prior information to the learning process and boost the learning performance. Accordingly, we design a codeword matrix according to the hierarchical structure of cellular compartments. In Table 2, we illustrate the proposed codeword matrix by taking advantage of the cellular compartments structure shown in Figure 2.

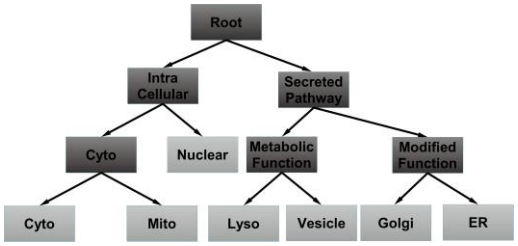

Fig. 2. Biological structure of cellular compartments.

As shown in Table 2, we derive six binary classifiers from this codeword matrix. Starting from roots, we use classifier h_{root} to distinguish between 3 intra-cellular compartments and the other 4 secreted pathway-based compartments. Then, for the splits under inter-cellular compartments, we apply h_{intra} to discriminate nuclear and cytoplasm internal node, a union of proteins in cytoplasm and mitochondrion. Similar to h_{intra}, h_{cyto} is another classifier that is applied to characterize the differences between cytoplasm and mitochondrion. Pointing to the right sub-tree of root node, since the main function of vesicle is to uptake and transport of materials within the cytoplasm, and lysosome is capable of breaking down all kinds of biomolecules, they can be categorized into metabolic functional compartments. Moreover, the main functions of golgi apparatus and ER are modifying the proteins for cell secretion, so classifier h_{sp} is applied to distinguish the cellular compartments having either metabolic or modified functions under the node of Secreted Pathway. Last but not least, h_{meta} and h_{modi} are also constructed to classify the compartments within the nodes of Metabolic Function and Modified Function shown in Figure 2. In this work, we will mainly use this coding pattern to predict protein subcellular location under ECOC framework. Moreover, in Lin et al.'s work (Lin, et al., 2011), they also utilize the biological structure information and then build a tree-based classifier to predict the sequence-based protein subcellular location. Here, we will compare our proposed ECOC-based method with Lin et al.'s method in Section S. 5 in the supplementary material.

2.6 ECOC Coding by Adding Checking Bits

From the detailed analysis in supplementary Section S.4, it is worth noting that, although the codeword matrix shown in Table 2 follows the biological structure shown in Figure 2, it has no error-correcting ability for the Hamming distances between pairs of codewords are too short (e.g. there is only 1 bit difference between the codewords under the nodes of Cytoplasm, Metabolic Function and Metabolic Function). So, in order to strengthen the error-correcting ability of the codeword matrix shown in Table 2, we add 8 checking bits for each codeword of cellular compartment (shown in Table 3) to enlarge the hamming distances between pairs of codewords.

As can be seen from Table 3, the newly added 8 checking bits are used for distinguishing different nodes or cellular compartments (e.g. c_1 is used to distinguish between Modified Function and Cytoplasm based proteins), which cannot reflect the hierarchical structure of cellular compartments in Figure 2. So, after adding these 8 checking bits, each cellular compartment is represented by 14 bits, and the Hamming distances between pairs of codewords are accordingly enlarged (e.g. the hamming distance between the codewords under the nodes of Cytoplasm is enlarged from 2 to 4 if we add the 8 checking bits).

Table 2. Corresponding coding matrix to the biological structure of cellular compartments.

	h_{root}	h_{intra}	h_{cyto}	h_{sp}	h_{meta}	h_{modi}
Cytoplasm	-1	1	1	0	0	0
ER	1	0	0	-1	0	1
Golgi	1	0	0	-1	0	-1
Lyso	1	0	0	1	1	0
Mito	-1	1	-1	0	0	0
Nuclear	-1	-1	0	0	0	0
Vesicle	1	0	0	1	-1	0

Table 3. The added 8 checking bits to enlarge the hamming distances between pairs of codewords.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
Cytoplam	1	1	0	0	1	0	0	0
ER	-1	0	1	0	0	0	1	0
Golgi	-1	0	1	0	0	0	0	1
Lysosome	0	-1	0	1	0	0	-1	0
Mito	1	1	0	0	0	1	0	0
Nuclear	0	0	-1	-1	-1	-1	0	0
Vesicle	0	-1	0	1	0	0	0	-1

2.7 Kernel Combination

As the second procedure in ECOC framework, binary classifier learning is also important for multi-class classification. To better make use of different kinds of features, we adopt a kernel combination method (Zhang, et al., 2011) (Wang, et al., 2008) to design each of multiple binary classifiers in ECOC. Specifically, for each dichotomy, we fuse both the global features (i.e., Haralick

feature and DNA feature) and the local features (i.e., LBP) by a multi-kernel based SVM classifier (Zhang, et al., 2011) (Wang, et al., 2008).

Suppose that we are given n protein images. Let $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ denote the global and the local feature of the i-th sample, respectively, and their corresponding labels belong to $\{-1, 1\}$. Multi-kernel based SVM solves the following primal problem

$$\begin{aligned} &\arg\min_{\mathbf{w}^{(m)},\beta_m,\varepsilon_i} \frac{1}{2}\sum_{m=1}^{2}\beta_m\left\|\mathbf{w}^{(m)}\right\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ &s.t. \quad y_i\left(\sum_{m=1}^{2}\beta_m((\mathbf{w}^{(m)})^T\phi^m(\mathbf{x}_i^m)+b)\right)\geq 1-\varepsilon_i \\ &\varepsilon_i \geq 0 \ (i=1,2...,n) \end{aligned} \tag{3}$$

where $\mathbf{w}^{(m)}$, β_m and ϕ^m represent the normal vector to hyperplane, the weight value, and the kernel-induced mapping function of m-th type of feature, respectively. For a testing sample, its corresponding label can be obtained by the following

$$f(x) = sign\left(\sum_{i=1}^{n} y_i\alpha_i \sum_{m=1}^{2}\beta_m k^m(\mathbf{x}_i^m, \mathbf{x}^m)+b\right) \tag{4}$$

Here, k^m is the kernel function of the m-th type of feature induced by mapping function ϕ^m, with element. $k^{(m)}(\mathbf{x}_i^m, \mathbf{x}_j^m) = \phi^m(\mathbf{x}_i^m)^T\phi^m(\mathbf{x}_j^m)$. From Equation (3) and (4), we can see that the multi-kernel based SVM is an extension of single-kernel based SVM, where the kernel matrix for multi-kernel based SVM is a linear combination of its single kernel matrix on different types of features. The weight value $\beta_m(m=1,2, \beta_1+\beta_2=1)$ is applied to balance the importance between the global and the local features, which is determined by a coarse-grid search method (Zhang, et al., 2011). Specifically, we first equally split training samples into ten subsets, and utilize nine of them to train a series of models to determine a $\beta_m(m=1,2)$ that can achieve the highest classification accuracy on the remaining one subset. Then, after getting the optimal $\beta_m(m=1,2)$, we can train a model for all of the training samples, thus the final effectiveness of this model can be evaluated by classification accuracy on the testing samples.

2.8 Ensemble Classification Method

As can be observed from Figure 1, db wavelet has 10 vanishing moments from db1 to db10. Accordingly, we construct 10 SC-PSorter based classification models, with each one corresponding to a specific type of vanishing moments. Inspired by(Xu, et al., 2013) (Yang, et al., 2014) (Liu, et al., 2015), we adopt a majority voting strategy to combine those SC-PSorter based models together. Specifically, for a testing protein image $\mathbf{z}$, if the i-th (i=1, 2, …, 10) SC-PSorter model predicts that it belongs to the location $c(1 \leq c \leq 7)$, the vote for the c-th compartment is added by one. Then, $\mathbf{z}$ is in the location with the largest vote based on all of the 10 SC-PSorter models.

3. Experimental Results

3.1 Experimental Settings

In previous works (i.e. Xu, et al., 2013 and Yang, et al., 2014), researchers use images from the same protein for training and testing via cross-validation. Following these works, in our experiment, we use the same cross-validating strategy. Specifically, we equally divide the images in each protein into 5 disjoint subsets, with four subsets used for training and the remaining subset used for testing. For all the proposing and comparing methods in this paper, the SVM classifier is implemented by using LIBSVM toolbox(Chang, et al., 2011), with a RBF kernel and the parameter σ is tuned from 0.9 to 2.1 at a step size of 0.1 by using grid search on the training data. Also, for each feature f_i in the training set, a common feature normalization scheme is adopted, that is, the normalized feature $f_i^{'} = f_i / f_i^{\max}$, where $f_i^{\max}$ is the maximum value of the i-th feature in the training set. Also, the $f_i^{\max}$ value will be used to normalize its corresponding feature in the test set.

3.2 Results for Combining Different Types of Features

To evaluate the efficacy of using different types of features, we firstly perform experiments by only using one type of features(i.e., the global feature and local feature) and combining different types of features together (i.e., direct combine and kernel combine) to predict the targets of proteins. Here, we choose the one-vs-all coding strategy, with experimental results shown in Figure 3. As can be seen from Figure 3, on one hand, direct combination of different types of features will not lead to the improvements of prediction accuracies. In most cases, the classification accuracies for this combination strategy are between using global or local feature only. On the other hand, using kernel combination method to fuse different types of features is a much more effective way, where classification accuracies consistently outperform those methods based on one single type of features (i.e., global feature or local feature). However, even for the kernel combination strategy, the classification accuracies cannot achieve to 70% for all of the 10 db wavelets, which reminds us to replace the simple one-vs-all coding strategy with other coding strategies for further improving.

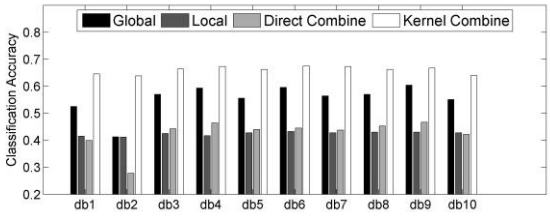

Fig. 3. Classification accuracies by using single type of features and combining different types of features together.

3.3 Results for Different Coding Strategies

In the second groups of experiments, we test the classification performances for three different coding strategies, namely, one-vs-all, forest and biological structural based coding strategies (i.e., following the codeword matrix shown in Table 2), which are denoted as OVA-PSorter, F-PSorter, and S-PSorter, respectively. Here, we also use kernel combination method to fuse global features (including both Haralick features and DNA features) and local features (i.e., LBP features) due to its superior classification performances in the first group of experiments. Figure 4 shows the classification accuracies by using all of the 10 db wavelets.

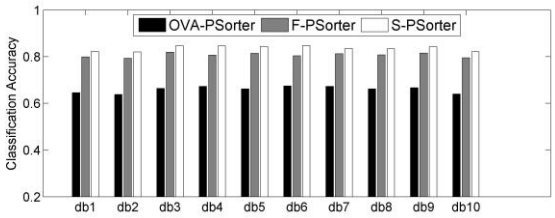

Fig. 4. Performance comparison among different coding strategies.

As can be seen from Figure 4, our proposed S-PSorter method consistently outperforms the other two methods (i.e., OVA-PSorter and F-PSorter), which shows the advantage of using the biological structure of cellular compartments to design the codeword matrix. On the other hand, Figure 4 indicates that F-PSorter achieves consistently better classification accuracies than OVA-PSorter. This is because F-PSorter constructs codeword matrix by maximizing the mutual information between different classes rather than the simple one-vs-all coding strategy used in OVA-PSorter. Moreover, we also compare the computational efficiency of different coding strategies in supplementary Section S.8

3.4 Further Improvement by Adding Checking Bits

As discussed in Section 2.6, we also add 8 checking bits (shown in Table 3) to the codeword matrix of S-PSorter model to strengthen its error-correcting ability. Here, we denote two ECOC-based methods, whose codeword matrices are derived by adding these 8 checking bits to the S-PSorter based codeword matrix and only using these 8 checking bits, as SC-PSorter and C-PSorter, respectively. Figure 5 presents the individual classification accuracies of the above two methods when comparing with S-PSorter method for all of the 10 db models.

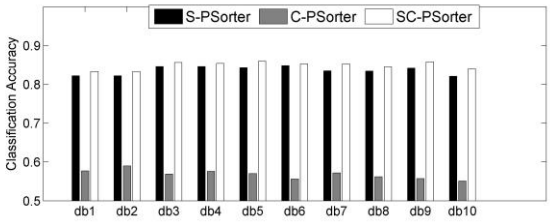

Fig. 5. Performance comparisons among S-PSorter, C-PSorter and SC-PSorter methods.

As can be seen from Figure 5, on one hand, SC-PSorter consistently outperforms S-PSorter on all of the 10 db wavelets. This is because we enlarge the hamming distances between pairs of codewords, and thus a few mistakes in some bits can be corrected by the decoding procedure. On the other hand, Figure 5 also shows the classification accuracies of C-PSorter method are consistently inferior to the other methods, which is because these 8 checking bits are just designed to distinguish different nodes or cellular compartments, and they do not reflect the hierarchical structure of cellular compartments shown in Figure 2. (Detailed classification results reported in supplementary Section 4)

3.5 Ensemble Results With Different Coding Strategies

As shown in Figure 4 and Figure 5, for each method (i.e., OVA-PSorter, F-PSorter, S-PSorter and SC-PSorter), the classification accuracies from its individual 10 classifiers are different, which motivates us to utilize an ensemble strategy for better fusing the complementary individual decisions. Here, we use the majority voting strategy introduced in section 2.8 to combine different classifiers together. Table 4 compares the classification accuracies of the best individual classifier and the ensemble model for these 4 methods (i.e., OVA-PSorter, F-PSorter, S-PSorter and SC-PSorter)

As can be seen from Table 4, we can always obtain better classification accuracies when performing the classifier ensemble via majority voting. Table 4 also indicates that the classification accuracy of our proposed SC-PSorter method is improved from the best individual classifier of 0.860 to 0.890, which is the best overall classification accuracy among all of the four methods.

Table 4. Comparison between individual and ensemble classification for four different coding strategies.

	Best independent classifier	*Ensemble prediction*
OVA-PSorter	0.675	0.679
F-PSorter	0.819	0.853
S-PSorter	0.848	0.874
SC-PSorter	0.860	0.890

4. Discussion

4.1 Comparisons with Existing Works

We also compare our SC-PSorter method with several existing approaches for image-based prediction of protein subcellular location. For example, in (Xu, et al., 2013) the authors construct k SVM models where k is the number of classes. In training the i-th SVM model, examples belonging to the i-th class are seen as positive samples, while the other examples as negative samples. For a query protein image $\mathbf{z}$, the output vector consists of k probabilities related to k different subcellular locations. We say $\mathbf{z}$ belongs to the class with the largest probability in the output vector. Different from (Xu, et al., 2013), the authors in (Yang, et al., 2014) construct $k(k-1)/2$ SVM classifiers where each one is trained from two different classes. Then, the testing sample $\mathbf{z}$ is fed into these $k(k-1)/2$ SVMs and these classifiers also output a probability denoting which class it belongs to. Here, the SVM classifier is implemented with RBF kernel and the parameter σ is also tuned from 0.9 to 2.1 at a step size of 0.1 by using grid search on the training data.

Figure 6 presents the individual classification accuracies of the above two methods when comparing with our proposed SC-PSorter method. As can be seen from Figure 6, our proposed SC-PSorter method consistently achieves better classification accuracies than the other two approaches. Here, the better classification accuracies of our method are mainly owing to the following two aspects: 1) Since the prior biological information is crucial in computational biology, we use the biological structural information to guide the learning procedure; 2) We apply the multiple kernel combination strategy to fuse different types of

features, which is regarded as a much more effective and flexible way than the direct combination strategy applied in the other two methods.

We also compare the classification accuracies of the best individual classifier and the ensemble model for the above three methods. As can be seen from Table 5, for these three methods, we obtain better classification accuracies when performing the classifier ensemble via majority voting than the best individual classifier. On the other hand, Table 5 also indicates that, by performing the ensemble strategy, the SC-Porter method achieves the best classification accuracy among all of the three methods. This result again validates the advantage of our proposed SC-PSorter method for prediction of protein subcellular location.

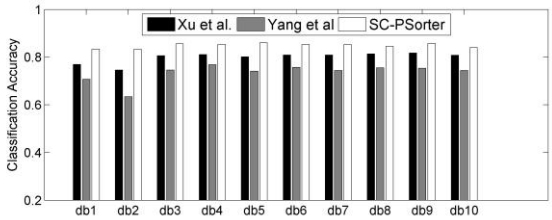

Fig. 6. Classification accuracies achieved by SC-PSorter and the other two methods.

Table 5. Comparison between individual and ensemble classification for three different methods.

	Best independent classifier	*Ensemble prediction*
Xu et al.	0.817	0.826
Yang et al.	0.768	0.772
SC-PSorter	0.860	0.890

4.2 Diversity Analysis

For the purpose of understanding how our proposed ensemble SC-PSorter works, we endeavor to apply the kappa measure (Rodriguez and Kuncheva, 2006) to plot the diversity-error diagram, which evaluates the level of agreement between the outputs of two individual classifiers. In Figure 7, we show the diversity-error diagrams of ensemble SC-PSorter and the methods in (Xu, et al., 2013) and (Yang, et al., 2014) for the task of predicting protein subcellular locations. In our experiment, each ensemble contains ten individual classifiers, with each corresponding to a specific classifier using different global db features. The value on the x-axis of a diversity-error diagram denotes the kappa diversity of a pair of classifiers in the ensemble, while the value on the y-axis is the averaged individual error of a pair of classifiers. Since a small value of kappa diversity indicates better diversity and a small value of averaged individual error indicates a better accuracy, the most desirable pairs of classifiers will be close to the bottom left corner of the graph. As shown in Figure 7, our proposed ensemble SC-PSorter achieves much lower kappa value as well as much lower classification errors than the methods in (Xu, et al., 2013). At the same time, our proposed SC-PSorter method is not as diverse as the method in (Yang, et al., 2014), but apparently, it has more accurate base classifiers than the method in (Yang, et al., 2014). It seems that our proposed method can achieve a better trade-off between accuracy and diversity than the compared two methods. That is, it builds a classifier ensemble based on the reasonable diverse but markedly accurate individual components.

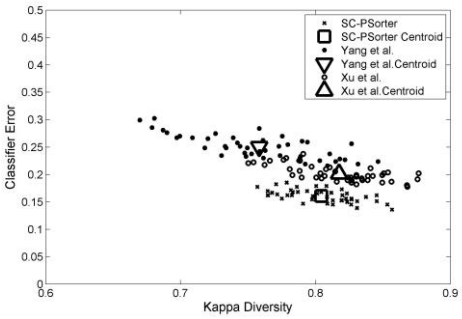

Fig. 7. The diversity-error diagrams of classifiers in the task of determining protein subcellular locations.

4.3 Slight Variations on Tree Structure

We design another two variants of the hierarchy of cellular compartments (i.e. T1, T2 shown in Figure S1 and Figure S2 in supplementary Section S.1, respectively) , which are constructed by making slight variations on the tree structure in Figure 2. Specifically, on one hand, for the tree representation T1, we neglect the hierarchical structure of 4 cellular compartments (i.e. Lysosome, Vesicle, Golgi, ER) under the node of Secreted Pathway. On the other hand, for the tree representation T2, we misuse Cytoplasm (which is originally under the node of Intra Cellular) as a metabolic functional compartment (under the node of Secreted Pathway).

The classification results in Table S3 in the supplementary Section S.1 shows that the slight variations on the tree structure in Figure 2 will lead to decreases in the classification accuracies for all of the 10 db models when comparing with S-PSorter method. Also, as can be seen from Table S4 in the supplementary Section S.1, the ensemble classification results for these two tree representations (i.e. T1 and T2) are 0.866 and 0.842, respectively, which are still higher than those of previously published methods (i.e., Xu, et al., 2013 and Yang, et al., 2014), although a bit lower than our original S-PSorter method. These results suggest that the proposed tree representation in Figure 2 reflects the true hierarchy of subcellular compartments.

4.4 Prediction on Unseen Proteins

As mentioned in Section 3.1, we use images from the same protein to evaluate the performance of SC-PSorter method. However, a strict test, i.e., recognizing subcellular patterns in new protein, is also very important. So we have added one more experiment to compare SC-PSorter method with the other two methods (i.e., Xu, et al., 2013 and Yang, et al., 2014) for predicting proteins that are not included in the training set **(**detailed in supplementary Section 3**).**

As can be seen in Table S6, our proposed SC-PSorter method consistently achieves better classification accuracies than the other two approaches for all of the 10 db wavelets. Moreover, when performing the classifier ensemble via majority voting, the classification accuracies for SC-PSorter，Xu et al. (2013) and Yang et al. (2014) methods are 0.809, 0.684 and 0.603, respectively.

This result again validates the advantage of our proposed SC-PSorter method for the prediction of protein subcellular location.

5. Conclusion

In this paper, we develop and test a novel prediction model, SC-PSorter, for determining image based protein subcellular locations. Specifically, we firstly devise a novel codeword matrix by considering the biological structural information under the ECOC framework, and then for each binary classifier corresponding to the columns of the ECOC codeword matrix, we adopt kernel combination method to fuse different types of features. Finally, we develop a classifier ensemble by combining multiple SC-PSorter based classifiers via majority voting.

In this study our method has been shown effective in case of each protein corresponding to only one location. However, as a matter of fact, nearly 20% percentage of human proteins co-exist more than two locations (Zhu, et al., 2009), and thus we will design a new method to solve this multi-label based protein classification problem. Also, since different biomarker may provide complementary information for the prediction of protein subcellular location (Breker and Schuldiner, 2014), we will add non-image data (e.g., amino acid sequence) to our image-based predictor for further performance improvement.

ACKNOWLEDGEMENTS

We thank Prof. Hongbin Shen for his helpful suggestions, and the anonymous reviewers for valuable comments.

Funding: This work was supported in part by the National Natural Science Foundation of China (61422204; 61473149); Jiangsu Natural Science Foundation for Distinguished Young Scholar (BK20130034); NUAA Fundamental Research Funds (NE2013105).

REFERENCES

Boland, M.V., Markey, M.K. and Murphy, R.F. (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images, *Cytometry*, **33**, 366-375.

Boland, M.V. and Murphy, R.F. (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells, *Bioinformatics*, **17**, 1213-1223.

Breker, M. and Schuldiner, M. (2014) The emergence of proteome-wide technologies: systematic analysis of proteins comes of age, *Nat Rev Mol Cell Biol* , **15**, 453-464.

Chang, C.C. and Lin, C.J. (2011) LIBSVM: A Library for Support Vector Machines, *Acm T Intel Syst Tec*, **2**.

Chebira, A., Barbotin, Y. and Charles, J. (2007) A multiresolution approach to automated classification of protein subcellular location images, *BMC Bioinformatics*, **8**, 5.

Chi, S.M. (2010) Prediction of protein subcellular localization by weighted gene ontology terms, *Biochem Bioph Res Co*, **399**, 402-405.

Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins*, **43**, 246-255.

Coelho, L.P., Kangas, J.D., Naik, A.W., Osuna-Highley, E., Glory-Afshar, E., Fuhrman, M., Simha, R., Berget, P.B., Jarvik, J.W. and Murphy, R.F. (2013) Determining the subcellular location of new proteins from microscope images using local features, *Bioinformatics*, **29**, 2343-2349.

Dietterich, T.G., Bakiri, G. and (1995) Solving Multiclass Learning Problems via Error-Correcting Output Codes, *Artif Intell* , **2**, 24.

Escalera, S., Pujol, O. and Radeva, P. (2007) Boosted Landmarks of Contextual Descriptors and Forest-ECOC: A novel framework to detect and classify objects in cluttered scenes, *Pattern Recogn Lett*, **28**, 1759-1768.

Escalera, S., Pujol, O. and Radeva, P. (2010) On the Decoding Process in Ternary Error-Correcting Output Codes, *IEEE Trans Pattern Anal Mach Intell*, **32**, 120-134.

Glory, E. and Murphy, R.F. (2007) Automated Subcellular Location Determination and High-Throughput Microscopy, *Dev Cell* , **12**, 10.

Jeong, J.C., Lin, X.T. and Chen, X.W. (2011) On Position-Specific Scoring Matrix for Protein Function Prediction, *IEEE Acm T Comput Bi*, **8**, 308-315.

Huang, K., M. Velliste. and Murphy, R.F. (2003). Feature Reduction for Improved Recognition of Subcellular Location Patterns in Fluorescence Microscope Images, *Proc. SPIE*, ***4962***, *307-318*.

Kumar, A., Rao, A., Bhavani, S., Newberg, J.Y. and Murphy, R.F. Automated analysis of immunohistochemistry images identifies candidate location biomarkers for cancers, *Proc Natl Acad Sci U S A*, **111**, 18249-18254.

Li, J.Y., Xiong, L., Schneider, J. and Murphy, R.F. (2012) Protein subcellular location pattern classification in cellular images using latent discriminative models, *Bioinformatics*, **28**, I32-I39.

Lin, T.H., Murphy, R.F. and Bar-Joseph, Z. (2011) Discriminative Motif Finding for Predicting Protein Subcellular Localization, *IEEE ACM T Comput Bi*, **8**, 441-451.

Liu, M., Zhang, D., Shen, D. and the Alzheimer's Disease Neuroimaging, I. (2015) View-centralized multi-atlas classification for Alzheimer's disease diagnosis, *Hum Brain Mapp*, **36**, 1847-1865.

Liu, M., Zhang, D., Chen, S., Xue, H. (2015) Joint Binary Classifier Learning for ECOC-based Multi-class Classification, *IEEE Trans Pattern Anal Mach Intell*, **99**.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebet, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R. and Golub, T.R. (2005) MicroRNA expression profiles classify human cancers, *Nature*, **435**, 834-838.

Murphy, R.F., Velliste, M. and Porreca, G. (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images, *J Vlsi Sig Proc Syst*, **35**, 311-321.

Nair, R. and Rost, B. (2005) Mimicking cellular sorting improves prediction of subcellular localization, *J Mol Biol*, **348**, 85-100.

Nanni, L., Lumini, A. and Brahnam, S. (2010) Local binary patterns variants as texture descriptors for medical image analysis, *Artif Intell Med*, **49**, 117-125.

Newberg, J. and Murphy, R.F. (2008) A framework for the automated analysis of subcellular patterns in human protein atlas images, *J Proteome Res*, **7**, 2300-2308.

Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T. (2000) Text classification from labeled and unlabeled documents using EM, *Mach Learn*, **39**, 103-134.

Ojala, T., Pietikäinen, M. and Harwood, D. (1996) A comparative study of texture measures with classification based on featured distributions, *Pattern Recognit* , **29**, 9.

Peng, H., Bateman, A., Valencia, A. and Wren, J.D. (2012) Bioimage informatics: a new category in Bioinformatics, *Bioinformatics*, **28**, 1057.

Pierleoni, A., Martelli, P.L. and Casadio, R. (2011) MemLoci: predicting subcellular localization of membrane proteins in eukaryotes, *Bioinformatics*, **27**, 1224-1230.

Ponten, F., Jirstrom, K. and Uhlen, M. (2008) The Human Protein Atlas - a tool for pathology, *J Pathol*, **216**, 387-393.

Pujol, O., Radeva, P. and Vitria, J. (2006) Discriminant ECOC: a heuristic method for application dependent design of error correcting output codes, *IEEE Trans Pattern Anal Mach Intell*, **28**, 1007-1012.

Rodriguez, J.J. and Kuncheva, L.I. (2006) Rotation forest: A new classifier ensemble method, *IEEE Trans Pattern Anal Mach Intell* , **28**, 1619-1630.

Tan, X.Y. and Triggs, B. (2010) Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions, *IEEE Trans Image Process* , **19**, 1635-1650.

Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C.A., Amini, B., Andersen, E., Andersson, A.C., Angelidou, P., Asplund, A., Asplund, C., Berglund. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics, *Mol Cell Proteomics*, **4**, 1920-1932.

Wang, X. and Li, G.Z. (2013) Multi-Label Learning via Random Label Selection for Protein Subcellular Multi-Locations Prediction, *IEEE Acm T Comput Bi*.

Wang, Z., Chen, S. and Sun, T. (2008) MultiK-MHKS: a novel multiple kernel learning algorithm, *IEEE Trans Pattern Anal Mach Intell*, **30**, 348-353.

Xu, Y.Y., Yang, F., Zhang, Y. and Shen, H.B. Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning, *Bioinformatics*, **31**, 1111-1119.

Xu, Y.Y., Yang, F., Zhang, Y. and Shen, H.B. (2013) An image-based multi-label human protein subcellular localization predictor (iLocator) reveals protein mislocalizations in cancer tissues, *Bioinformatics*, **29**, 2032-2040.

Yang, F., Xu, Y.Y., Wang, S.T. and Shen, H.B. (2014) Image-based classification of protein subcellular location patterns in human reproductive tissue by ensemble learning global and local features, *Neurocomputing*, **131**, 113-123.

Yoon, Y. and Lee, G.G. (2012) Subcellular Localization Prediction through Boosting Association Rules, *IEEE Acm T Comput Bi*, **9**, 609-618.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D. and Alzheimer's Disease Neuroimaging, I. (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment, *Neuroimage*, **55**, 856-867.

Zhang, T.L., Ding, Y.S. and Chou, K.C. (2006) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence, *Comput Biol Chem*, **30**, 367-371.

Zhu, L., Yang, J., and Shen, H.B. (2009) Multi label learning for prediction of human protein subcellular localizations. *Protein J*., **28**, 384–390.