Multi-Label Learning With Fuzzy Hypergraph Regularization for Protein Subcellular Location Prediction

Jing Chen*, Yuan Yan Tang, *Fellow, IEEE*, C. L. Philip Chen, *Fellow, IEEE*, Bin Fang, *Senior Member, IEEE*, Yuewei Lin, and Zhaowei Shang

***Abstract*—Protein subcellular location prediction aims to predict the location where a protein resides within a cell using computational methods. Considering the main limitations of the existing methods, we propose a hierarchical multi-label learning model FHML for both single-location proteins and multi-location proteins. The latent concepts are extracted through feature space decomposition and label space decomposition under the nonnegative data factorization framework. The extracted latent concepts are used as the codebook to indirectly connect the protein features to their annotations. We construct dual fuzzy hypergraphs to capture the intrinsic high-order relations embedded in not only feature space, but also label space. Finally, the subcellular location annotation information is propagated from the labeled proteins to the unlabeled proteins by performing dual fuzzy hypergraph Laplacian regularization. The experimental results on the six protein benchmark datasets demonstrate the superiority of our proposed method by comparing it with the state-of-the-art methods, and illustrate the benefit of exploiting both feature correlations and label correlations.**

***Index Terms*—Dictionary learning, hypergraph regularization, multi-label learning, protein subcellular localization.**

I. Introduction

PROTEINS are basically important for organisms' physiological actions. Proteomics research is an attractive field in the post-genomic era. The number of newly found proteins is dramatically increasing in the last two decades, however, in which the instances whose functions we have known only cover a small part of protein databases. Proteins perform their appropriate functions only when they are located in the correct subcellular locations, which is a key functional characteristic of protein [1]. Thus protein subcellular location prediction is of great significance to the functional analysis of proteins and drug discovery. However, the traditional way to determine subcellular location of proteins is performed by biochemical experimental tests. As we know, it is time-consuming and costly. With the explosion of newly found proteins, the gap between the new proteins and the knowledge of their subcellular locations is becoming sharply wide. To tackle this problem, it is extraordinarily desirable to develop automated methods to predict subcellular locations of proteins accurately.

In the past two decades, many efforts were paid in attempts to predict proteins' subcellular locations. The pioneering investigations, such as [2], [3] and [4], originally suggested the feasibility of constructing computational models by using protein composition sequence information for protein subcellular location prediction. Motivated by these early works, several automated subcellular location predictors were then proposed for various organisms' proteins [5]–[8]. For the details of these works, we shall refer readers to the two comprehensive reviews [9], [10]. In recent years, this field has been attracting increasing attentions, and fast advances have been published. These researches mainly focus on how to effectively represent a protein (e.g., feature extraction) and how to construct prediction models (e.g., classifier construction).

For feature extraction, most researches extract the following two types of discriminative feature representations: amino acid sequence based and high-level information based. The former only involves the amino acid sequences of proteins, which can be further divided into three groups: 1) sorting-signal-based features, 2) composition-based features, and 3) homology-based features. The sorting-signal-based feature is an earlier type of protein feature representation [9], [11]–[13], where N-terminal sorting signals are used to discriminate proteins residing in different subcellular components. The composition-based method statistically analyzes the composition information embedded in the entire range of amino acid sequences, such as amino acid compositions (AAC) [4], [14], amino acid pair compositions (dipeptide) [15], gapped amino acid pair compositions [16], and pseudo amino acid compositions (PseAAC) [7]. In particular, Chou's pseudo amino acid composition is one of the

Manuscript received February 27, 2014; revised May 28, 2014; accepted July 09, 2014. Date of publication July 31, 2014; date of current version November 25, 2014. This work is supported by the Research Grants MYRG205(Y1-L4)-FST11-TYY, MYRG187(Y1-L3)-FST11-TYY, and Chair Prof. Grants RDG009/FST-TYY of University of Macau as well as Macau FDC grants T-100-2012-A3 and 026-2013-A. This research project is also supported by the National Natural Science Foundations of China (61273244, 60873092, 90820306 and 61100114) and the Fundamental Research Funds for the Central Universities (CDJXS10182216). This work is also sponsored by the National Key Technology R&D Program of China (No.2012BAI06B01) and Major Program of National Natural Science Foundation of China (No.61190122). *Asterisk indicates corresponding author.*

*J. Chen is with the Faculty of Science and Technology, University of Macau, Taipa, Macau, and also with Chongqing University, Chongqing 400030, China (e-mail: chenjingmc@gmail.com).

Y. Y. Tang is with the Faculty of Science and Technology, University of Macau, Taipa, Macau and Chongqing University, Chongqing 400030, China (e-mail: yytang@umac.mo).

C. L. P. Chen is with the Faculty of Science and Technology, University of Macau, Taipa, Macau (e-mail: philipchen@umac.mo).

B. Fang and Z. Shang are with the college of Computer Science, Chongqing University, Chongqing 400030, China (e-mail: fb@cqu.edu.cn; szw@cqu.edu.cn).

Y. Lin is with University of South Carolina, Columbia, SC 29208 USA. (e-mail: ywlin.cq@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNB.2014.2341111

most commonly-used feature representations, which is often incorporated with other protein attributes, for example, physical chemistry properties [17] and evolutionary information [18]. The homology-based feature is based on the assumption that homologous sequences are more likely to have the same subcellular location [19]–[21]. The high-level information based methods construct features based on high-level semantic annotations from external knowledge bases, such as texts from PubMed titles and abstracts on proteins [22], and Gene Ontology (GO) terms from annotation databases [23], [24].

For classifier construction, so far, many computational methods, such as K-nearest neighbor [25], [26], support vector machines [27], neural networks [28], [29], and hidden Markov models [30], [31] were widely used for protein subcellular location prediction. In particular, among these methods, the support vector machine is shown to achieve better performance in most cases. Furthermore, several novel machine learning paradigms, such as ensemble learning [32], semi-supervised learning [33], multi-task learning [34], transfer learning [35], and boosting learning [36], have been applied for this field. As a typical classification problem from the point view of pattern recognition, some recognition techniques including dimension reduction [37], feature fusion [38], and feature selection [39] etc., have also been employed in this field, which substantially improve the performance of protein subcellular location prediction.

However, the main limitations of these existing intelligent techniques could be summarized as the following two points.

- The traditional methods assume that each protein resides at only one subcellular location, thus they handle a single-label problem. However, we need to note that some proteins may simultaneously exist in, or move between two or more different subcellular locations. In fact, proteins of this kind should draw our special attentions because they may have some valuable biological functions for both basic research and drug discovery [40], [41]. Besides, the recent research [42] by Millar *et al.* has shown an increasing number of proteins with multiple locations in the cell. So it is necessary to take multi-location or multiplex proteins into account when constructing the subcellular location predictors.
- The traditional prediction models are usually constructed based on the direct mapping from extracted features to annotation labels. In other words, these methods construct the simple "flat" models. In fact, the hierarchical multi-layer prediction models have been widely and effectively evaluated in many other biological pattern recognition fields [43], [44]. Some recent researches indicate that a hierarchial structure is substantially beneficial for exploring relations embedded in data features and annotation labels of biological systems [45]–[47]. It would be promising to consider a hierarchical prediction model for protein subcellular location prediction.

For the first issue, in fact, there have been several studies to address subcellular location prediction for multi-location proteins recently [48]–[52]. Chou *et al.* proposed a series of methods, such as iLoc-Euk [53], iLoc-Plant [54], iLoc-Virus [55], iLoc-Hum [56], iLoc-Gpos [57], iLoc-Gneg [58] and iLoc-Animal [59], to deal with the multi-location problem for eukaryotic, plant, virus, human, Gram-positive, Gram-negative, and animal proteins, respectively. All these methods follow the general steps including feature vectors construction, multiple K-nearest neighbor learners training and prediction results integration. In addition, some novel machine learning paradigms, such as transfer learning [60], semi-supervised learning [37], and binary relevance-based multi-label learning [61], have been also employed to cope with the multi-location problem. We shall refer readers to the recent comprehensive review for the details of the multi-label prediction in molecular biosystems [62]. Among the existing researches, most methods directly transform the multi-label problem into multiple independent traditional single-label classification tasks. Some studies suggest us that the straight solutions of this type are usually not optimal. Furthermore, the traditional methods ignore the possible correlations embedded in multi-label problems. In fact, each subcellular location is not isolated physiologically. From our general experiences, there could be relations among samples with different labels, among samples within the same label, and among involved labels, which few studies considered in subcellular location multi-label prediction. To better solve this problem, in this work, we integrate intra-label similarity and inter-label diversity, which involves both feature space and label space, into the proposed multi-label learning scheme.

For the other issue, Nair and Rost have constructed a tree structure called LOCtree by hand to mimic the cellular sorting process for protein subcellular location prediction [63]. They suggest that this simple hierarchical structure performs better than those traditional "flat" methods. Pierleoni *et al.* proposed a similar tree structure called BaCelLo to incorporate the relationships between subcellular locations, and placed more emphasis on performance balance among all the locations [64]. Both these methods predefined the hierarchical structures by utilizing the priori knowledge of the localization mechanisms. As pointed in [65], the problem of this predefined architecture is that a prediction mistake at a top node could not be corrected at lower nodes. Bulashevska and Eils constructed a hierarchical prediction model through a learning process based on labeled sequence data, where the relationships between subcellular locations were not explored explicitly [65]. The method of Yang *et al.* explored the interdependences between subcellular locations and incorporated them into the learned hierarchical prediction model [66]. We note that the inter-location relationships were represented simply through the different pathways on the prediction tree.

Towards this background, in this work, we deal with the task of subcellular location prediction of multi-location proteins. The proposed multi-label learning method is constructed on the three-layer hierarchical model as Fig. 1. This model consists of the three layers: feature layer, label layer and latent layer. The latent layer acts as the link between the feature layer and the label layer. The extracted latent concepts perform as the dictionary items which are commonly used in document analysis. Two normal graphs are constructed within the feature layer and the label layer, respectively. In the feature layer, the original features are decomposed onto the latent concepts. A fuzzy hypergraph is used to regularize the consistency between the original features and the intermediate latent codes. The other hypergraph is constructed to regularize the consistency

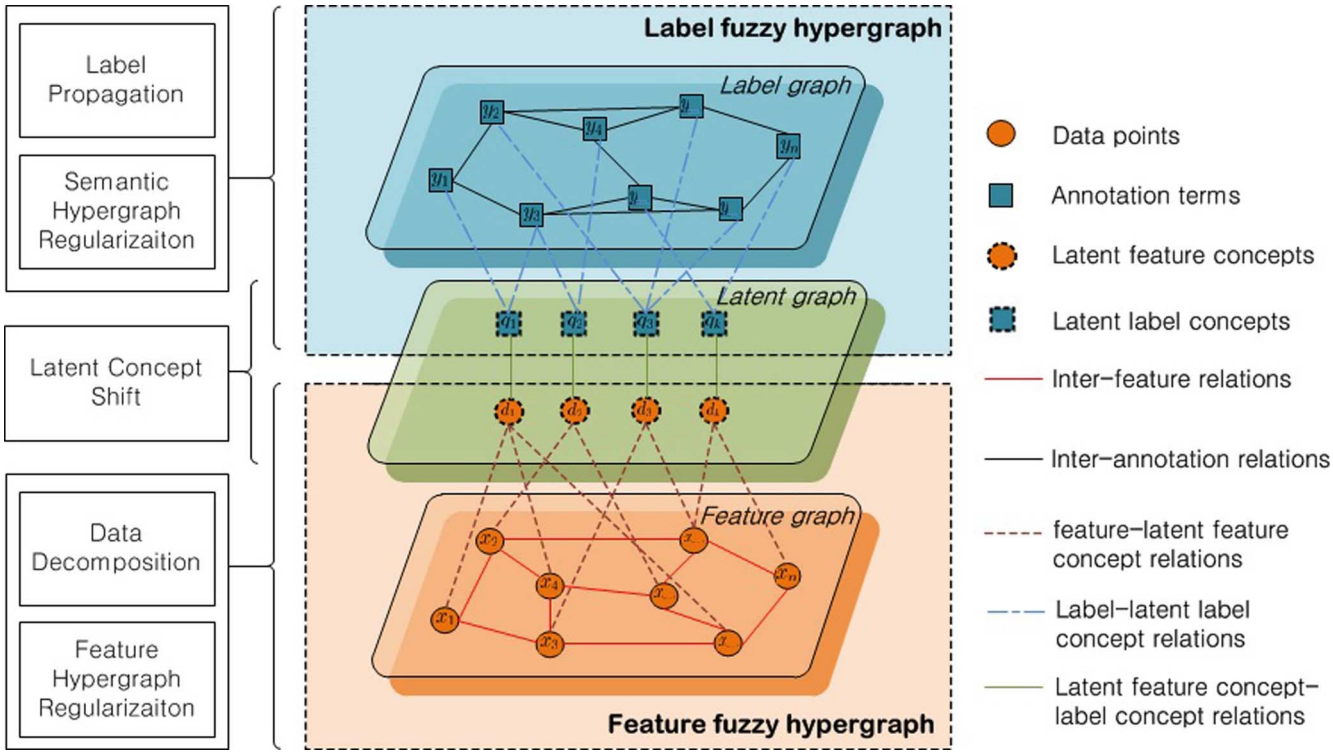

Fig. 1. The diagram of the proposed three-layer model.

between the annotation lists and the latent codes. Above all, the annotation information is propagated from the labeled proteins to the unlabeled proteins by the dual fuzzy hypergraph regularized multi-label learning.

As pointed out in a recent comprehensive review [67] and carried out in a series of follow-up publications [68]–[75], to establish a practical statistical predictor for a protein system, we need to consider the following procedures: i) construct or select a valid benchmark dataset to train and test the predictor; ii) formulate the protein or biological samples with an effective mathematical expression or model that can truly reflect their intrinsic correlation with the attribute to be predicted; iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; v) establish a user-friendly web-server for the predictor that is accessible to the public. The following describes how to deal with these steps.

The remainder of this paper is organized as follows: in Section II, the problem of subcellular location prediction for multi-location proteins is represented formally and then a multi-label prediction method is proposed. In Section III, we introduce the six protein benchmark datasets, formulate a protein with the two types of effective mathematical expressions, perform cross-validation tests on the six datasets to evaluate the effectiveness of the proposed predictor, report and discuss the experimental results. Finally, Section IV summarizes our work and presents some future directions.

II. The Proposed FHML Method

A. Problem Formulation

For a multi-label protein subcellular location prediction task, consider a protein database $\mathcal{D} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$ of n protein sequences and an annotation vocabulary $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_k\}$ of k subcellular location labels. Each protein $\mathcal{I}_i$ is represented by its original feature vector $x_i \in \mathbb{R}^m$ for $i = 1, 2, \ldots, n$. Then we have the protein dataset $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m \times n}$. Among these proteins in the database, l proteins are annotated with one or more subcellular location labels of the vocabulary $\mathcal{V}$, and other u proteins are not annotated. Here, $l + u = n$. Without loss of generality, we assume that the first l proteins are labeled in advance by the label indicator matrix $\widetilde{Y}_L = [\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_l] \in \{0, 1\}^{k \times l}$. Each $\widetilde{y}_i$ is a k-dimensional vector. The value of 1 indicates that the protein $\mathcal{I}_i$ resides at the corresponding subcellular location and the value of 0 indicates $\mathcal{I}_i$ has no probability to exist in that location. We denote the output real-valued label score matrix as $Y = [Y_L\ Y_U]$, and the final 0–1 label matrix as $Y^* = [Y_L^*\ Y_U^*]$.

B. Latent Encoding

Considering the limitation of the existing methods, we explore the intrinsic relations embedded both in feature space and label space, i.e., feature correlation and label correlation. To deal with this issue, we extract latent feature concepts for feature space and latent label concepts for label space. And then based on the extracted latent concepts, we construct an individual hypergraph in feature space and label space, respectively, to capture the embedded intrinsic high-order relations. Then the three-layer hierarchical model could be constructed to deal with the multi-label problem.

At first, we decompose the original feature vectors onto their latent feature concepts by dictionary learning under the nonnegative matrix factorization framework (NMF) [76]. Motivated by the nonnegative matrix factorization originally applied in face recognition, we know that the extracted latent feature concepts are relevant to the parts of the original holistic

features. Thus, the obtained new feature representation is localized on the latent feature concepts. For the protein dataset $X = [x_1, x_2, \ldots, x_n]$, formally, we reconstruct it by using the linear combination of latent feature concepts as $X = DZ$, where $D = [d_1, d_2, \ldots, d_r] \in \mathbb{R}^{m \times r}$ is the latent feature concept basis matrix and $Z = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{r \times n}$ is the new feature representation over the latent basis. In this work, the basis matrix D and the coefficient matrix Z are both constrained as nonnegative matrices. Each d_i acts as a latent feature concept and the nonnegative column vector z_j is used as the weight coefficient vector of the jth protein belongs to each latent feature concept. D and Z could be obtained under the dictionary learning framework as follows,

$$\begin{aligned} \min_{D,Z} \; & \|X - DZ\|_F^2 \\ s.t. \quad & D, Z \geq 0, \\ & \mathbf{1}^T Z = \mathbf{1}^T. \end{aligned} \tag{1}$$

The constraint $\mathbf{1}^T Z = \mathbf{1}^T$ enforces each column z_j to be a normalized weight vector. Here, we call the new feature representation Z as latent codes.

In addition, we also decompose the annotation vectors onto the latent label concepts. For the protein dataset X, we denote the corresponding subcellular location annotation matrix as Y. For Y, we define the prediction model from the latent codes to the annotation vectors as follows:

$$Y = QZ \tag{2}$$

where $Q \in \mathbb{R}^{k \times r}$ and $Q \geq 0$. Thus, the column vectors $Q = [q_1, q_2, \ldots, q_r]$ are regarded as the latent label concepts, and Q is used as the codebook in label space. Here, we assume $Q = PD$, where $P \in \mathbb{R}^{k \times m}$ and $P \geq 0$. Then, P is the relation matrix which shifts the latent components from feature space to label space.

Herein, the Y can be predicted by $Y = PDZ$. In addition, the predicted labels of labeled data should be enforced to be consistent with original labels. Mathematically, we should first optimize the following objective function:

$$\begin{aligned} \min_{P,D,Y} \; & \lambda_1 \|Y - PDZ\|_F^2 + \lambda_2 \|Y_L - \tilde{Y}_L\|_F^2 \\ s.t. \quad & P, D, Y \geq 0 \\ & \mathbf{1}^T Y = \mathbf{1}^T. \end{aligned} \tag{3}$$

The last constraint $\mathbf{1}^T Y = \mathbf{1}^T$ normalizes each annotation vector to avoid the scaling problem. Moreover, this normalization constraint ensures that we can substitute the standard inner for the cosine similarity.

C. Dual Fuzzy Hypergraph Laplacian Regularization

We can view the above decomposition in this way: the sample i is related to the latent feature concept j with the weight z_{ji} when $z_{ji} \neq 0$, and the sample i is unrelated to the latent feature concept j when $z_{ji} = 0$. Herein, each protein sequence feature vector could be reconstructed by some latent feature concepts; on the other hand, each latent feature concept covers a subset of samples. The z_{ji} acts as a membership degree of the protein i to the latent feature concept j. The decomposition of label space could be explained in the similar way. Naturally, the latent feature concept i could viewed to be belonged to itself group completely. We define the latent code of the latent feature concept i as the column vector z_i^D with 1 in i-th entry and 0 elsewhere. Then, the latent codes of the latent feature concepts can be define as $Z_D = \{z_i^D\}_{r \times r}$, and the latent label concepts share the same codes Z_D.

This viewpoint motivates us to employ a hypergraph to represent these relations, in which a hyperedge covers a subset of vertices. We construct fuzzy hypergraphs in feature space and label space, respectively. Each latent concept corresponds to a hyperedge, and the instances (i.e., feature vectors in feature space, annotation vectors in label space) connected to the latent concept belong to its corresponding hyperedge. Here, the instance i is connected to the latent concept j if its weight z_{ji} is non-zero. In feature space, we construct a fuzzy hypergraph $G_F = (V_F, E_F, W_F)$, where V_F is the set of vertices associated to protein features, E_F is the set of hyperedges associated to latent feature concepts and W is the fuzzy degrees of vertices to hyperedges. Here, let $W_F = Z$. In this way, all the protein samples are organized by using latent feature concepts on the fuzzy hypergraph. In label space, we also construct a fuzzy hypergraph $G_S = (V_S, E_S, W_S)$, where V_S is the set of vertices associated to protein annotation vectors, E_S is the set of hyperedges associated to latent label concepts and W_S is the fuzzy degrees of vertices to hyperedges. Here, let $W_S = Z$. In this way, all the protein annotations are also organized by the fuzzy hypergraph.

To capture the embedded intrinsic correlation, we perform a novel regularization on this fuzzy hypergraph. The regularization is based on the assumption that the proteins in the same feature hyperedge have similar latent codes and the similar latent codes yield similar annotations. This type of intrinsic relations could be preserved by performing hypergraph Laplacian regularization.

Following the star expansion algorithm, we transform the initial fuzzy hypergraphs G_F and G_S into the two bipartite graphs $\hat{G}_F = (\hat{V}_F, \hat{E}_F)$ and $\hat{G}_S = (\hat{V}_S, \hat{E}_S)$ with the adjacency matrices as $\hat{W}_F$ and $\hat{W}_S$ by introducing a new vertex for each hyperedge. Then we could transform the dual fuzzy hypergraph Laplacian regularization into the traditional graph Laplacian regularization. The vertex set $\hat{V}_F$ consists of the initial vertices correspond to protein feature vectors and the new vertices correspond to latent feature concepts, i.e., $\hat{V}_F = \hat{X} = [X, D]$. The weight of each edge in G_F is inherited from the fuzzy membership degree of each vertex in the hypergraph G_F, i.e., the weight of each edge is defined as the inner of the two joint vertices. The similarity matrix $\hat{W}_F$, whose entry $\hat{W}_{ij} = \hat{x}_i^T \hat{x}_j$ measures the similarity between a vertex pair $(\hat{x}_i, \hat{x}_j)$, i.e.,

$$\hat{W}_F = \hat{X}^T \hat{X} = \begin{bmatrix} X^T X & X^T D \\ D^T X & D^T D \end{bmatrix}. \tag{4}$$

We define the degree matrix as $\hat{D}_F$, which is a diagonal matrix with $\hat{D}_{ii}^F = \sum_j \hat{W}_{ij}^F$.

On the other hand, the vertex set of the bipartite graph $\hat{G}_S$ in label space is $\hat{Y} = [Y, Q] = [y_1, y_2, \ldots, y_n, q_1, q_2, \ldots, q_r]$. The pairwise similarity is measured by the inner $\hat{y}_i^T \hat{y}_j$ for the pair

$(\hat{y}_i, \hat{y}_j)$. Thus, these pairwise similarity measures constitute the following similarity matrix $\hat{W}^S$ with $\hat{y}_i^T \hat{y}_j$ to be the entry W_{ij}^S.

$$\hat{W}_S = \hat{Y}^T \hat{Y} = \begin{bmatrix} Y^T Y & Y^T Q \\ Q^T Y & Q^T Q \end{bmatrix}. \tag{5}$$

Thus, we can extend the optimization problem (1) by adding a graph Laplacian regularization term as follows:

$$\begin{aligned} \min_{D,Z} \|X - DZ\|_F^2 + \lambda_3 tr(\hat{Z}\hat{L}_F\hat{Z}^T) \\ s.t. \quad D, Z \geq 0 \\ \mathbf{1}^T Z = \mathbf{1}^T. \end{aligned} \tag{6}$$

Define the Laplacian matrix $\hat{L}_F = \hat{D}_F - \hat{W}_F$.

In label space, the optimization problem (3) is extended by adding a graph Laplacian regularization term as follows:

$$\begin{aligned} \min_{P,D,Y} \quad & \lambda_1 \|Y - PDZ\|_F^2 + \lambda_2 \|Y_L - \tilde{Y}_L\|_F^2 \\ & + \lambda_4 tr(\hat{Z}\hat{L}_S\hat{Z}^T) \\ s.t. \quad & P, D, Y \geq 0 \\ & \mathbf{1}^T Y = \mathbf{1}^T \end{aligned} \tag{7}$$

where the Laplacian matrix $\hat{L}_S = \hat{D}_S - \hat{W}_S$.

D. Multi-Label Learning Formulation

By integrating all of the above two folds, the semisupervised multilabel learning problem for protein subcellular location prediction is formulated as a dual fuzzy hypergraph regularized nonnegative data factorization problem in the following form:

$$\begin{aligned} \min_{P,D,Z,Y} \quad & \|X - DZ\|_F^2 + \lambda_1 \|Y - PDZ\|_F^2 \\ & + \lambda_2 \|Y_L - \tilde{Y}_L\|_F^2 + \lambda_3 \mathrm{Tr}(\hat{Z}\hat{L}_F\hat{Z}^T) \\ & + \lambda_4 \mathrm{Tr}(\hat{Z}\hat{L}_S\hat{Z}^T) \\ s.t. \quad & P, D, Z, Y \geq 0 \\ & \mathbf{1}^T Z = \mathbf{1}^T, \quad \mathbf{1}^T Y = \mathbf{1}^T. \end{aligned} \tag{8}$$

The parameters λ_1, λ_2, λ_3 and λ_4 are used to balance the contribution of each objective terms to the solution.

E. Solution

The cost function is not convex with respect to D, Z, P and Y together. Thus, it is not realistic to find the global minima. However, the cost function is strictly convex with respect to each matrix variable block respectively. So, here we adopt the common method which is to iteratively optimize the objective function by alternatively minimizing over one matrix variable while keeping the other three blocks fixed. Together with the strict convexity of the objective function, we can deduce that each subproblem has a unique minimum. Here, for the nonnegative constraint, we employ the multiplicative iterative algorithm used for NMF. For the sum-to-one constraint, an effective technique in [77] is employed here. We use the matrices $\bar{X}$ and $\bar{D}$ to take the place of X and D as inputs, which are defined as

$$\bar{X} = \begin{bmatrix} X \\ \delta \mathbf{1}^T \end{bmatrix}, \qquad \bar{D} = \begin{bmatrix} D \\ \delta \mathbf{1}^T \end{bmatrix} \tag{9}$$

where δ adjusts the effect of the sum-to-one constraint. The larger δ forces each column of X or D to keep the sum-to-one constraint better but slows down the convergence rate. Similarly, We use the following equation to replace the original decomposition assumption $Y = PDZ$.

$$\begin{bmatrix} I \\ \delta \mathbf{1}^T \end{bmatrix} Y = \begin{bmatrix} PDZ \\ \delta \mathbf{1}^T \end{bmatrix}. \tag{10}$$

Here, we denote $\bar{I} = \begin{bmatrix} I \\ \delta \mathbf{1}^T \end{bmatrix}$ and $\bar{S} = \begin{bmatrix} PDZ \\ \delta \mathbf{1}^T \end{bmatrix}$.

After the real-valued label score matrix Y is obtained, we need a cut-off threshold to transform the score matrix into the 0–1 matrix Y^*. Thus, we obtain the final predicted label subset for each protein. In this work, we employ the S-Cut technique to optimize the threshold based on the Hamming distance between the actual label matrix $\tilde{Y}_L$ and the predicted label matrix Y_L^* of the labeled proteins. The detailed steps is summarized in Algorithm – FHML.

Algorithm — FHML

Input: protein dataset X, annotated label matrix $\tilde{Y}_L$
Initialization: Randomly choose D^0, Z^0, P^0 and Y^0 as nonnegative matrices.

A. **For** $t = 0, 1, 2, \ldots, T_{\max}$, **do**

1) For given $D = D^t, P = P^t, Y = Y^t$, update the latent codes Z as:

$$Z_{ij}^{t+1} = Z_{ij}^t \times \frac{(\bar{D}^T \bar{X} + A_1^Z + A_2^Z + A_3^Z)_{ij}}{(\bar{D}^T \bar{D} Z + A_4^Z)_{ij}};$$

2) For given $Z = Z^t, P = P^t, Y = Y^t$, update the latent feature concept basis matrix D as:

$$D_{ij}^{t+1} = D_{ij}^t \times \frac{(A^D)_{ij}}{(D(A^D)^T A)_{ij}};$$

3) For given $Z = Z^t, D = D^t, P = P^t$,, update the label ranking matrix Y as:

$$Y_{ij}^{t+1} = Y_{ij}^t \times \frac{(A_1^Y + A_2^Y)_{ij}}{(A_3^Y)_{ij}};$$

4) For given $Z = Z^t, D = D^t, Y = Y^t$, update the relation matrix P as:

$$P_{ij}^{t+1} = P_{ij}^t \times \frac{(A^P)_{ij}}{(P(A^P)^T P)_{ij}};$$

5) If $\| Z^{t+1} - Z^t \| < \epsilon, \| D^{t+1} - D^t \| < \epsilon$, $\| Y^{t+1} - Y^t \| < \epsilon$, and $\| P^{t+1} - P^t \| < \epsilon$, then break.

end

B. Optimize the threshold θ and perform cut-off on Y_U

Output: The predicted label matrix Y_U^*

The above abbreviations are expanded as follows:

$$\begin{aligned} A_1^Z &= \lambda_1 D^T P^T Y \\ A_2^Z &= \lambda_3 (Z X^T X + Z_D D^T X) \\ A_3^Z &= \lambda_4 (Z Y^T Y + Z_D D^T P^T Y) \end{aligned}$$

$$A_4^Z = \lambda_1 D^T P^T P D Z$$
$$A^D = X Z^T + \lambda_1 P^T Y Z^T + \lambda_3 X Z^T Z_D + \lambda_4 P^T Y Z^T Z_D$$
$$A_1^Y = \lambda_1 \bar{I}^T \bar{S} + \lambda_2 \tilde{Y}_L \bar{I}^T$$
$$A_2^Y = \lambda_4 (Y Z^T Z + P D Z_D^T Z)$$
$$A_3^Y = \lambda_1 \bar{I}^T \bar{I} Y + \lambda_2 Y \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$
$$A^P = \lambda_1 Y Z^T D^T + \lambda_4 Y Z^T Z_D D^T.$$

III. Experimental Results and Discussion

A. Materials

We evaluate our proposed method on the six multi-location protein benchmark datasets from the well-known package Cell-Ploc 2.0 [78]. The protein subcellular location annotations in these datasets are experimentally determined. These six datasets cover eukaryotic, human, plant, gram-positive, gram-negative, and virus cells which are denoted by eukaryote, human, plant, gpos, gneg, and virus in the following discussions, respectively. Each protein in the dataset has less than 25% sequence similarity to any other in the same subcellular location group, which makes it more reliable to compare our proposed method with others. The datasets are obtained from the Online Supporting Information in [78], where the more detailed description of these six datasets can also be found.

To develop an effective predictor for a protein system, one of the keys is to describe a protein mathematically in an effective formulation that can truly reflect their intrinsic correlation with the target to be predicted [62]. For protein subcellular location prediction, there are many different representation methods, however, our focus is to show the benefit of our proposed hierarchical multi-label learning method comparing with other existing multi-label methods for protein subcellular location prediction. We only extract the two types of discriminative features, i.e., PseAAC and PSSM-ACT, for a given protein sequence, and then concatenate them serially as its original high-dimension feature vectors. The detailed description of these two types of features can be found in [38] and hence there is no need to repeat here. In this study, the PseAAC features of all proteins are generated via the server at: http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/. In this work, the PseAAC and PSSM-ACT feature vectors are both in 140-dimension.

B. Performance Measures

Performance evaluation in multi-label prediction is different from that in traditional single-label prediction. As the case study of [79] suggested, a multi-label system should be evaluated by the two paradigms: example-based and label-based. The example-based evaluation generates a score for each example (protein) and averages later over all examples. The label-based evaluation judges the quality of a multi-label system for each label (subcellular location) and then averages over all labels. Accordingly, we use example-based *F-Measure* and *Accuracy*, and label-based *Precision* and *Recall*. The definitions of these four measures are given as follows, following those presented in [79]. The detailed explanation of definitions and intuitive meanings of multi-label performance measures in molecular biosystems can also be found in the recent comprehensive review [48]. For a label (subcellular location) c, a ground-truth annotation binary vector $y_c \in \mathbb{R}^n$ denotes the membership of all proteins. The i-th "1" element y_{ci} of the vector y_c indicates the protein i belongs to the location c. Accordingly, the estimated annotation by the predictor is denoted as a binary vector z_c. If the elements of the binary vectors are treated as logical values, then $TP(c)$, $FP(c)$ and $FN(c)$ can be written as: $TP(c) = \sum_{i=1}^{n}(z_{ci} \vee y_{ci})$, $FP(c) = \sum_{i=1}^{n}(\neg z_{ci} \vee y_{ci})$, and $FN(c) = \sum_{i=1}^{n}(z_{ci} \vee \neg y_{ci})$. Meanwhile, for the i-th example, $\mathcal{Y}_i$ denotes its true label set and $\mathcal{Z}_i$ denotes its predicted label set. Then the average label-based *Precision* and *Recall*, and example-based *F-Measure* and *Accuracy* are defined as follows:

$$Precission = \frac{1}{k}\sum_{c=1}^{k}\frac{TP(c)}{TP(c)+FP(c)}, \tag{11}$$

$$Recall = \frac{1}{k}\sum_{c=1}^{k}\frac{TP(c)}{TP(c)+FN(c)}, \tag{12}$$

$$F-Measure = \frac{1}{n}\sum_{i=1}^{n}\frac{2|\mathcal{Y}_i \bigcap \mathcal{Z}_i|}{|\mathcal{Y}_i| + |\mathcal{Z}_i|} \tag{13}$$

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathcal{Y}_i \bigcap \mathcal{Z}_i|}{|\mathcal{Y}_i \bigcup \mathcal{Z}_i|}. \tag{14}$$

C. Experimental Settings

From the existing publications, the commonly-cited methods able to deal with both single-location and multi-location proteins in subcellular location prediction are multi-label K-nearest neighbor (abbreviated as mKNN) in iLoc-Euk [53] and multi-label support vector machine (abbreviated as mSVM) in ML-PLoc [80]. As we know, the general KNN and SVM are both popular methods and evaluated effectively in many classification applications. In iLoc-Euk, Chou *et al.* extended the cosine distance based KNN by introducing an accumulation-layer scale into the multi-location version, which is at present known as the best prediction method able to deal with multi-location proteins when predicting protein subcellular localization [53]. In ML-PLoc, Zhu *et al.* decomposed the multi-label prediction problem into multiple independent binary classification problem and each subproblem is solved by a SVM classifier [80]. In this study, we choose these two competitive predictors as the baseline methods to evaluate our proposed method. Since our focus in this evaluation is only on the discriminative ability of predictors, for the fairness and reliability of comparison results, the input of mKNN and mSVM is also the serially combination of PseAAC and PSSM-ACT features as our proposed method dose. Although iLoc-Euk web-server and ML-Ploc software package are publicly available, the former only accepts Gene Ontology (GO) representation or PSSM feature as its input and the latter only uses PSSM feature to make prediction. Here, we construct mKNN predictor following its original description [53], while the proteins are represented as the combination of PseAAC and PSSM-ACT features. The mSVM is extended for the same protein features as used in our method through using the core codes provided by ML-PLoc package [81]. For mKNN, $K = 3$. For mSVM, the parameters of SVM is set as $\gamma = 2^1$ and $C = 2^7$, which has been demonstrated to yield the best performance in

TABLE I
PERFORMANCE COMPARISON OF THE DIFFERENT METHODS ON THE SIX DATASETS

Dataset	Method	*Precision*	*Recall*	*F-Measure*	*Accuracy*
eukaryote	mKNN	0.5460±0.0009	0.5035±0.0010	0.7968±0.0011	0.7763±0.0013
	mSVM	0.5395±0.0121	0.5081±0.0138	0.7834±0.0114	0.7798±0.0125
	FHML	**0.5642±0.0010**	**0.5279±0.0009**	**0.8160±0.0007**	**0.7913±0.0010**
human	mKNN	0.5286±0.0025	0.4927±0.0028	0.7512±0.0033	0.7403±0.0024
	mSVM	0.5345±0.0031	0.4883±0.0026	0.7468±0.0028	0.7391±0.0028
	FHML	**0.5753±0.0027**	**0.5296±0.0023**	**0.8279±0.0029**	**0.7860±0.0022**
plant	mKNN	0.4875±0.0034	0.4423±0.0030	0.7294±0.0036	0.7026±0.0030
	mSVM	0.4908±0.0049	0.4496±0.0053	0.7375±0.0050	0.7210±0.0052
	FHML	**0.5547±0.0032**	**0.5038±0.0032**	**0.8026±0.0041**	**0.7813±0.0033**
gpos	mKNN	0.6691±0.0054	0.6758±0.0057	0.9366±0.0051	0.9378±0.0051
	mSVM	0.6740±0.0046	0.6647±0.0047	0.9458±0.0043	0.9293±0.0050
	FHML	**0.6883±0.0047**	**0.6902±0.0051**	**0.9527±0.0042**	**0.9408±0.0045**
gneg	mKNN	0.6904±0.0019	0.6902±0.0018	0.9403±0.0015	0.9341±0.0019
	mSVM	0.6850±0.0018	0.6881±0.0020	0.9372±0.0016	0.9325±0.0015
	FHML	**0.7021±0.0014**	**0.6974±0.0013**	**0.9615±0.0011**	**0.9486±0.0016**
virus	mKNN	0.6371±0.0063	0.6225±0.0070	0.8633±0.0068	0.8379±0.0065
	mSVM	0.6323±0.0071	0.6209±0.0062	0.8676±0.0060	0.8351±0.0062
	FHML	**0.6487±0.0055**	**0.6315±0.0056**	**0.8804±0.0064**	**0.8642±0.0055**

the original study [80]. For our FHML method, the parameters λ_i's, the number of latent concepts r, the parameter δ of the sum-to-one constraint, and the convergence parameter ϵ are optimized by using 3-fold cross validation on the labeled set. The λ_i's are tuned from 10^{-5} to 10^{-3}. r is tuned from 50 to 500. We uniformly select twenty values for each parameter range and select the highest one to finetune. Here, $\lambda_1 = 0.00047$, $\lambda_2 = 0.00182$, $\lambda_3 = 0.00131$, $\lambda_4 = 0.00064$ and $r = 120$. For δ, we need to consider the balance between the sum-to-one constraint satisfaction and the convergence rate. We try the δ's from 0 to 100 with the $\mathrm{interval} = 5$. Finally, we find that when $\delta = 20$ the system has a relatively better prediction accuracy and an acceptable convergence rate. So, a relatively small value $\delta = 20$ is chosen in this work. For ϵ, the smaller ϵ leads to a more exact solution but slows down the convergence. We find that the ϵ smaller than 10^{-3} is helpless for the further significant improvement of the prediction accuracy but leads to a larger time cost. So $\epsilon = 10^{-3}$ is chosen in our work.

In statistical prediction, for evaluating the effectiveness of a predictor in practical application, the following three cross-validation methods are commonly used: independent dataset test, subsampling (e.g., K-fold cross validation) test, and jackknife test. Among the three methods, the jackknife test is considered as the most objective because it can always yield a unique result for a given benchmark dataset, as elucidated in [78] and [67]. Accordingly, the jackknife test has been increasingly and widely used to examine the performance of various prediction methods [82]–[88]. However, to reduce the computational time, we adopted the 10-fold cross-validation test in this study as done by many investigators with SVM as a prediction engine. As more detail, each 10-fold cross validation is repeated for ten times, where all the proteins are randomly divided into 10 mutually exclusive parts with approximately equal size and approximately equal class distribution. The averaged results are reported in this work.

D. Experimental Results

Table I illustrates the experimental results of the three compared methods on the six multi-location protein datasets in terms of the four multi-label performance measures, where the best result of each measure on each dataset is shown in bold face. From this result, we can conclude the following observations. 1) mKNN and mSVM perform similarly, and both worse than FHML on the six datasets in terms of the four measures. FHML achieves the best averaged performance 62.2% for *Precision*, 59.7% for *Recall*, 87.4% for *F-Measure* and 85.2% for *Accuracy* on the six datasets. 2) On the eukaryote, human, and gneg datasets, FHML performs better than those on the plant, gpos, and virus datasets. FHML achieves the averaged performance improvement 4.6% for *Precision*, 4.2% for *Recall*, 6.2% for *F-Measure* and 4.7% for *Accuracy* on the former three datasets, while 1.8% for *Precision*, 1.5% for *Recall*, 1.9% for *F-Measure* and 1.9% for *Accuracy* on the latter three datasets.

E. Discussion

This experimental evaluation has shown the effectiveness of the proposed fuzzy hypergraph regularized hierarchical multi-label predictor. Generally, for a classification task, it becomes relatively more difficult when more class labels are considered. However, the superiority of FHML is more significant when the dataset contains more training samples and covers more subcellular locations. We notice that there are a larger number of proteins and subcellular locations in eukaryote, human, and plant datasets than the remaining three datasets. More informative relations embedded in feature space, and more inter-label and intra-label relations embedded in annotation label space can be expected in the former. Therefore, our proposed FHML method with exploiting correlations not only in feature space but also in label space is more effective, while the mKNN and mSVM

construct the flat prediction model directly from features to annotation labels without considering the correlations embedded in feature space and label space, and thus it performs worse than FHML, which confirms our intuition. This fact would suggest us that collecting more proteins to construct an abundant training dataset is necessary for a further development of protein subcellular localization predictors building.

IV. Conclusion and Future Work

In this work, we generate a hierarchical multi-label learning model with dual fuzzy hypergraph regularization. We explore the intrinsic relations both in feature space and label space. The experimental results have shown that our method outperforms the two state-of-the-art multi-location protein subcellular location prediction methods in terms of the four measures.

In this study, we only choose PseAAC and PSSM-ACT as our input features. In fact, in further work, other powerful protein feature extraction approaches, such as GO representation, are expected to improve the prediction performance of our proposed method. Furthermore, kinds of relations have been exploited with the help of the three-layer structure, while the relations within each individual layer are considered linearly. In fact, the more complex relation structure among subcellular locations would be explored in future work with the help of the biological evolutionary background. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [89], [90], and provide convenient tools for biology and drug researchers, we will make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgment

The authors gratefully thank the two anonymous reviewers for their helpful and constructive comments.

References

[1] F. Eisenhaber and P. Bork, "Wanted: Subcellular localization of proteins based on sequence," *Trends. Cell Biol.*, vol. 8, no. 4, pp. 169–170, Apr. 1998.

[2] K. Nishikawa, Y. Kubota, and T. Ooi, "Classification of proteins into groups based on amino acid composition and other characters," *J. Biochem.*, vol. 94, no. 3, pp. 981–1007, 1983.

[3] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," *Proteins: Struct. Funct., Genet.*, vol. 11, no. 2, pp. 95–110, 1991.

[4] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *J. Mol. Biol.*, vol. 238, no. 1, pp. 54–61, 1994.

[5] K.-C. Chou and D. W. Elrod, "Protein subcellular location prediction," *Protein Eng.*, vol. 12, no. 2, pp. 107–118, 1999.

[6] K.-C. Chou and Y. D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *J. Biol. Chem.*, vol. 277, no. 48, pp. 45765–45769, Nov. 2002.

[7] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Struct., Funct., Genet.*, vol. 43, no. 3, pp. 246–255, 2001.

[8] G.-P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins: Struct., Funct., Genet.*, vol. 50, no. 1, pp. 44–48, 2003.

[9] K. Nakai, "Protein sorting signals and prediction of subcellular localization," *Adv. Protein Chem.*, vol. 54, no. 1, pp. 277–344, 2000.

[10] K.-C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Anal. Biochem.*, vol. 370, no. 1, pp. 1–16, 2007.

[11] O. Emanuelsson, H. Nielsen, S. Brunak, and G. Heijne, "Predicting subcellular localization of proteins based on their n-Terminal amino acid sequence," *J. Mol. Biol.*, vol. 300, pp. 1005–1016, 2000.

[12] P. Horton, K. J. Park, T. Obayashi, and K. Nakai, "Protein subcellular localization prediction with WoLF PSORT," in *Proc. 4th Ann. Asia Pacific Bioinformat. Conf. (APBC' 06)*, 2006, pp. 39–48.

[13] H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano, "Extensive feature detection of N-terminal protein sorting signals," *Bioinformatics*, vol. 18, no. 2, pp. 298–305, 2002.

[14] J. Cedano, P. Aloy, J. A. Prez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *J. Mol. Biol.*, vol. 266, no. 3, pp. 594–600, 1997.

[15] Y. Huang and Y. Li, "Prediction of protein subcellular locations using fuzzy k-NN method," *Bioinformatics*, vol. 20, no. 1, pp. 21–28, 2004.

[16] K.-J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, vol. 19, no. 13, pp. 1656–1663, 2003.

[17] J. Wang, W.-K. Sung, A. Krishnan, and K.-B. Li, "Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines," *BMC Bioinformat.*, vol. 6, pp. 174–174, 2005.

[18] S.-W. Zhang, Y.-L. Zhang, H.-F. Yang, C.-H. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: An approach by incorporating evolutionary information and von neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.

[19] R. Nair and B. Rost, "Sequence conserved for subcellular localization," *Protein Sci.*, vol. 11, pp. 2836–2847, 2002.

[20] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[21] J. K. Kim, G. P. S. Raghava, S. Y. Bang, and S. Choi, "Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine," *Pattern Recogn. Lett.*, vol. 27, no. 9, pp. 996–1001, 2006.

[22] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, and H. Shatkay, "SherLoc2: A high-accuracy hybrid method for predicting subcellular localization of proteins," *J. Proteome Res.*, vol. 8, no. 11, pp. 5363–5366, 2009.

[23] S.-M. Chi, "Prediction of protein subcellular localization by weighted gene ontology terms," *Biochem. Biophys. Res. Commun.*, vol. 399, no. 3, pp. 402–405, 2010.

[24] S. Wan, M.-W. Mak, and S.-Y. Kung, "GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *J. Theor. Biol.*, vol. 323, no. 0, pp. 40–48, 2013.

[25] T. Wang and J. Yang, "Predicting subcellular localization of gram-negative bacte rial proteins by linear dimensionality reduction method," *Protein Pept. Lett.*, vol. 17, pp. 32–37, 2010.

[26] Y.-S. Ding and T.-L. Zhang, "Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier," *Pattern Recogn. Lett.*, vol. 29, no. 13, pp. 1887–1892, Oct. 2008.

[27] B. Liao, J. B. Jiang, Q. G. Zeng, and W. Zhu, "Predicting apoptosis protein subcellular location with PseAAC by incorporating tripeptide composition," *Protein Pept. Lett.*, vol. 18, pp. 1086–1092, 2011.

[28] L. Zou, Z. Wang, and J. Huang, "Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weigh ted inputs," *J. Genet. Genomics*, vol. 34, pp. 1080–1087, 2007.

[29] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, no. 1, pp. 50–57, Jul. 2013.

[30] T. Lin, R. Murphy, and Z. Bar-Joseph, "Discriminative motif finding for predicting protein subcellular localization," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 8, pp. 441–451, 2011.

[31] T.-H. Chang, L.-C. Wu, T.-Y. Lee, S.-P. Chen, H.-D. Huang, and J.-T. Horng, "EuLoc: A web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC," *J. Comput. Aid Mol. Des.*, vol. 27, no. 1, pp. 91–103, 2013.

[32] H.-B. Shen and K.-C. Chou, "Gpos-Ploc: An ensemble classifier for predicting subcellular localization of gram-positive bacterial proteins," *Protein. Eng. Des. Sel.*, vol. 20, no. 1, pp. 39–46, 2007.
[33] Q. Xu, D. Hu, H. Xue, W. Yu, and Q. Yang, "Semi-supervised protein subcellular localization," *BMC Bioinformat.*, vol. 10, no. Suppl 1, article S47, 2009.
[34] Q. Xu, S. J. Pan, H. H. Xue, and Q. Yang, "Multitask learning for protein subcellular location prediction," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 8, no. 3, pp. 748–759, May/Jun. 2011.
[35] S. Mei, W. Fei, and S. Zhou, "Gene ontology based transfer learning for protein subcellular localization," *BMC Bioinformat.*, vol. 12, no. Art. 44, 2011.
[36] Y. Yoon and G. G. Lee, "Subcellular localization prediction through boosting association rules," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 2, pp. 609–618, Mar./Apr. 2012.
[37] E. Pacharawongsakda and T. Theeramunkong, "Predict subcellular locations of singleplex and multiplex proteins by semi-supervised learning and dimension-reducing general mode of Chou's PseAAC," *IEEE Trans. NanoBiosci.*, vol. 12, no. 4, pp. 311–320, Dec. 2013.
[38] D. Yu, X. Wu, H. Shen, J. Yang, Z. Tang, Y. Qi, and J. Yang, "Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features," *IEEE Trans. NanoBiosci.*, vol. 11, no. 4, pp. 375–385, Dec. 2012.
[39] H. Lin, H. Ding, F.-B. Guo, and J. Huang, "Prediction of subcellular location of mycobacterial protein using feature selection techniques," *Mol. Divers.*, vol. 14, pp. 667–671, 2010.
[40] E. Glory and R. F. Murphy, "Automated subcellular location determination and highthroughput microscopy," *Dev. Cell*, vol. 12, pp. 7–16, 2007.
[41] C. Smith, Subcellular Targeting of Proteins and Drugs 2008 [Online]. Available: http://www.biocompare.com/Articles/TechnologySpotlight/976/ SubcellularTargeting-Of-Proteins-And-Drugs.html
[42] A. H. Millar, C. Carrie, B. Pogson, and J. Whelan, "Exploring the function-location nexus: Using multiple lines of evidence in defining the subcellular location of plant proteins," *Plant Cell*, vol. 21, pp. 1625–1631, 2009.
[43] J. X. Zhou, L. Brusch, and S. Huang, "Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model," *PLoS ONE*, vol. 6, no. 3, p. e14752, 2011.
[44] S. C. Basak, K. Balasubramanian, B. D. Gute, D. Mills, A. Gorczynska, and S. Roszak, "Prediction of cellular toxicity of halocarbons from computed chemodescriptors: A hierarchical QSAR approach," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 4, pp. 1103–1109, 2003.
[45] T. Hou, Y. Li, and W. Wang, "Prediction of peptides binding to the PKA RII alpha subunit using a hierarchical strategy," *Bioinformatics*, vol. 27, no. 13, pp. 1814–1821, Jul. 2011.
[46] B. Karacali, "Hierarchical motif vectors for prediction of functional sites in amino acid sequences using quasi-supervised learning," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 8, no. 5, pp. 441–451, 2012.
[47] D. Stojanova, M. Ceci, D. Malerba, and S. Dzeroski, "Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction," *BMC Bioinformat.*, vol. 14, p. 285, Sep. 2013.
[48] K.-C. Chou and H.-B. Shen, "Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *J. Proteome Res.*, vol. 6, no. 5, pp. 1728–1734, 2007.
[49] K.-C. Chou and H.-B. Shen, "A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mploc 2.0," *PLoS ONE*, vol. 5, no. 4, p. e9931, Apr. 2010.
[50] K.-C. Chou and H.-B. Shen, "Plant-mPLoc: A top-down strategy to augment the power for predicting plant protein subcellular localization," *PLoS ONE*, vol. 5, no. 6, p. e11335, Jun. 2010.
[51] H.-B. Shen and K.-C. Chou, "Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites," *Biochem. Biophys. Res. Commun.*, vol. 355, no. 4, pp. 1006–1011, Apr. 2007.
[52] H.-B. Shen and K.-C. Chou, "A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0," *Anal. Biochem.*, vol. 394, no. 2, pp. 269–274, Nov. 2009.
[53] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: A multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, p. e18258, Mar. 2011.
[54] Z.-C. Wu, X. Xiao, and K.-C. Chou, "iLoc-Plant: A multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites," *Mol. BioSyst.*, vol. 7, pp. 3287–3297, 2011.
[55] X. Xiao, Z.-C. Wu, and K.-C. Chou, "iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *J. Theor. Biol.*, vol. 284, no. 1, pp. 42–51, 2011.
[56] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Mol. Biosyst.*, vol. 8, pp. 629–641, 2012.
[57] Z.-C. Wu, X. Xiao, and K.-C. Chou, "iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins," *Protein Pept. Lett.*, vol. 19, no. 1, pp. 4–14, 2012.
[58] X. Xiao, Z.-C. Wu, and K.-C. Chou, "A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites," *PLoS ONE*, vol. 6, no. 6, p. e20592, Jun. 2011.
[59] W. Z. Lin, J. A. Fang, X. Xiao, and K.-C. Chou, "iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins," *Mol. Biosyst.*, vol. 9, no. 4, pp. 634–644, Apr. 2013.
[60] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *J. Theor. Biol.*, vol. 310, pp. 80–87, Oct. 2012.
[61] G.-Z. Li, X. Wang, X. Hu, J.-M. Liu, and R.-W. Zhao, "Multilabel learning for protein subcellular location prediction," *IEEE Trans. NanoBiosci.*, vol. 11, no. 3, Sep. 2012.
[62] K.-C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Mol. BioSyst.*, no. 9, pp. 1092–1100, 2013.
[63] R. Nair and B. Rost, "Mimicking cellular sorting improves prediction of subcellular localization," *J. Mol. Biol.*, vol. 348, no. 1, pp. 85–100, 2005.
[64] A. Pierleoni, P. L. Martelli, P. Fariselli, and R. Casadio, "BaCelLo: A balanced subcellular localization prediction," *Bioinformatics*, vol. 22, pp. e408–e416, 2006.
[65] A. Bulashevska and R. Eils, "Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains," *Bioinformatics*, vol. 7, p. 298, 2006.
[66] W. Y. Yang, B. L. Lu, and J. T. Kwok, "Incorporating cellular sorting structure for better prediction of protein subcellular locations," *J. Exp. Theor. Artif. Intell.*, vol. 23, no. 1, pp. 79–95, 2011.
[67] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.
[68] W. Chen, P. M. Feng, H. Lin, and K.-C. Chou, "iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Res.*, vol. 41, no. 6, p. e68, Apr. 2013.
[69] J.-L. Min, X. Xiao, and K.-C. Chou, "iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking," *BioMed Res. Int.*, p. 701317, 2013.
[70] Y. Xu, X.-J. Shao, L.-Y. Wu, N.-Y. Deng, and K.-C. Chou, "iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, p. e171, 2013.
[71] X. Xiao, J. L. Min, P. Wang, and K.-C. Chou, "iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints," *J. Theor. Biol.*, vol. 337, pp. 71–79, Nov. 2013.
[72] Y. N. Fan, X. Xiao, J. L. Min, and K.-C. Chou, "iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking," *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 4915–4937, Mar. 2014.
[73] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen, and K.-C. Chou, "iNuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, Mar. 2014.
[74] B. Liu, D. Zhang, R. Xu, J. Xu, X. Wang, Q. Chen, Q. Dong, and K.-C. Chou, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, Feb. 2014.
[75] W. R. Qiu, X. Xiao, and K.-C. Chou, "iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components," *Int. J. Mol. Sci.*, vol. 15, no. 2, pp. 1746–1766, Jan. 2014.
[76] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[77] X. Lu, H. Wu, and Y. Yuan, "Double constrained NMF for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sci.*, 2013.
[78] K.-C. Chou and H.-B. Shen, "Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Nat. Sci.*, vol. 2, pp. 1090–1103, 2010.
[79] S. Nowak, H. Lukashevich, P. Dunker, and S. Ruger, "Performance measures for multilabel evaluation: A case study in the area of image classification," in *Proc. Int. Conf. Multimedia Inf. Retrieval*, 2010, pp. 35–44.
[80] L. Zhu, J. Yang, and H. Shen, "Multilabel learing for prediction of human protein subcellular localizations," *Protein J.*, vol. 28, pp. 384–390, 2009.
[81] [Online]. Available: http://www.csbio.sjtu.edu.cn/bioinf/ML-PLoc
[82] H. Mohabatkar, "Prediction of cyclin proteins using Chou's pseudo amino acid composition," *Protein Pept. Lett.*, vol. 17, no. 10, pp. 1207–1214, Oct. 2010.
[83] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Comput. Biol. Chem.*, vol. 34, no. 5-6, pp. 320–327, Dec. 2010.
[84] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *J. Theor. Biol.*, vol. 264, no. 2, pp. 203–209, 2010.
[85] G. L. Fan and Q. Z. Li, "Predict mycobacterial proteins subcellular locations by incorporating pseudo-average chemical shift into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 304, pp. 88–95, Jul. 2012.
[86] G. L. Fan and Q. Z. Li, "Discriminating bioluminescent proteins by incorporating average chemical shift and evolutionary information into the general form of Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 334, pp. 45–51, Oct. 2013.
[87] J. Guo, N. Rao, G. Liu, Y. Yang, and G. Wang, "Predicting protein folding rates using the concept of Chou's pseudo amino acid composition," *J. Comput. Chem.*, vol. 32, no. 8, pp. 1612–1617, Jun. 2011.
[88] Z. Hajisharifi, M. Piryaiee, B. M. Mohammad, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *J. Theor. Biol.*, vol. 341, pp. 34–40, Jan. 2014.
[89] K.-C. Chou and H.-B. Shen, "Review: Recent advances in developing web-servers for predicting protein attributes," *Nat. Sci.*, vol. 1, no. 2, pp. 63–92, Sep. 2009.
[90] S.-X. Lin, "Theoretical and experimental biology in one – A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers," *J. Biomed. Sci. Engin.*, vol. 6, no. 4, pp. 435–442, Apr. 2013.

Jing Chen received the B.S. degree in computational mathematics from Shandong University, Ji'nan, China, the M.S. and Ph.D. degrees in computer science and technology from Chongqing University, Chongqing, China. He is currently a postdoctoral fellow with the Faculty of Science and Technology, University of Macau, Macau. His research interests include machine learning, pattern recognition, and biomedical information processing and analysis.

Yuan Yan Tang (F'04) received the B.S. degree in electrical and computer engineering from Chongqing University, Chongqing, China, the M.S. degree in electrical engineering from the Beijing University of Post and Telecommunications, Beijing, China, and the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada. He is a chair professor in the Faculty of Science and Technology at the University of Macau (UM). Before joining UM, he served as a chair professor in the Department of Computer Science at Hong Kong Baptist University and dean of the College of Computer Science at Chongqing University, China. He is a chair of the Technical Committee on Pattern Recognition of the IEEE Systems, Man, and Cybernetics Society (IEEE SMC) for his great contributions to wavelet analysis, pattern recognition, and document analysis. Recently, he was elected as one of the executive directors of the Chinese Association of Automation Council. With all his distinguished achievement, he is also the founder and editor-in-chief of the *International Journal of Wavelets, Multi-resolution, and Information Processing* (IJWMIP), and an associate editor of the *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI), and the *International Journal on Frontiers of Computer Science* (IJFCS). He has been presented with numerous awards such as the First Class of Natural Science Award of Technology Development Centre, Ministry of Education of the People's Republic of China in November 2005 and the Outstanding Contribution Award by the IEEE Systems, Man, and Cybernetics Society in 2007. He has published more than 300 papers, books, and book chapters. He is a Fellow of the Pattern Recognition Society (IAPR).

C. L. Philip Chen (F'07) received the M.S. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985 and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. From 1989 to 2002, he was with the Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA, as an Assistant, an Associate, and a Full Professor before he joined the University of Texas, San Antonio, TX, USA, where he was a Professor and the Chair of the Department of Electrical and Computer Engineering and the Associate Dean for Research and Graduate Studies of the College of Engineering. He is currently a Chair Professor of the Department of Computer and Information Science and Dean of the Faculty of Science and Technology, University of Macau, Macau, China. Dr. Chen is a Fellow of the AAAS. He is currently the President of the IEEE Systems, Man, and Cybernetics Society. In addition, he is an Accreditation Board of Engineering and Technology Education Program Evaluator for Computer Engineering, Electrical Engineering, and Software Engineering programs.

Bin Fang (SM'10) received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, the M.S. degree in electrical engineering from Sichuan University, Chengdu, China, and the Ph.D. degree in electrical engineering from the University of Hong Kong, Hong Kong. He is currently a Professor with the Department of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, medical image processing, biometrics applications, and document analysis. He has published more than 100 technical papers and is an Associate Editor of the *International Journal of Pattern Recognition and Artificial Intelligence*.

Yuewei Lin received the B.S. degree in optical information science and technology from Sichuan University, Chengdu, China, and the M.E. degree in optical engineering from Chongqing University, Chongqing, China. He is currently working toward the Ph.D. degree in the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA. His current research interests include computer vision, machine learning, and image/video processing.

Zhaowei Shang received the B.S. degree in computer science from the Northwest Normal University, Lanzhou, China, in 1991, the M.S. degree from the Northwest Polytechnical University, Xi'an, China, in 1999, and the Ph.D. degree in computer engineering from Xi'an Jiaotong University, Xi'an, in 2005. He is currently a professor with the Department of Computer Science, Chongqing University, Chongqing, China. His research interests include computer vision, pattern recognition, image processing, and wavelet analysis.