Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

PSL-Recommender: Protein Subcellular Localization Prediction using Recommender System

Ruhollah Jamali [1,2], Changiz Eslahchi [1,*] and Soheil Jahangiri-Tazehkand [2,*]

[1]Department of Computer Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
[2]School of Biological Sciences, Institute for Research in Fundamental Sciences(IPM), Tehran, Iran.

*To whom correspondence should be addressed.

Abstract

Motivation: Identifying a protein's subcellular location is of great interest for understanding its function and behavior within the cell. In the last decade, many computational approaches have been proposed as a surrogate for expensive and inefficient wet lab methods that are used for protein subcellular localization. Yet, there is still much room for improving the prediction accuracy of these methods.
Results: In this paper we present PSL-Recommender, a method that employs neighborhood regularized logistic matrix factorization to build a recommender system for protein subcellular localization. By evaluating on four human and animals benchmark datasets. We have shown that PSL-Recommender significantly outperforms state-of-the-art methods, improving the previous best method by 13%, 8%, 31%, and 2% in $\mathrm{F1-mean}$, and also 12%, 8%, 28% and 2% in ACC.
Availability: PSL-Recommender is freely available online at https://github.com/RJamali/PSL-Recommender
Contact: ch-eslahchi@sbu.ac.ir

1 Introduction

Proteins are responsible for a wide range of functions within cells. The functionality of a protein is entangled with its subcellular location. Therefore, identifying Protein Subcellular Localization (PSL) is of great importance for both biologists and pharmacists, helping them inferring a protein's function and identifying drug-target interactions (Nair and Rost [2003]). Recent advances in genomics and proteomics provide massive amount of protein sequence data, which extends the gap between sequence and annotation data. Although PSLs can be identified by experimental methods, these methods are laborious and time-consuming explaining why only a narrow range of PSL information in Swiss-Prot database has been verified in this manner (Zhou *et al.* [2016]). This problem augments the demand for accurate computational prediction methods. Developments of computational and machine learning techniques have provided fast and effective methods for PSL prediction (Höglund *et al.* [2006], Horton *et al.* [2007], Briesemeister *et al.* [2010], Zhou *et al.* [2016], Cheng *et al.* [2017], Mehrabad *et al.* [2018]).
The desired PSL prediction can be reached typically by relying on sequence-derived features, taking into consideration that using annotation-derived features can lead up to better performance. Different types of sequence-derived features have been used for PSL prediction. For example, PSORT (Nakai and Horton [1999]), WoLF PSORT (Horton *et al.* [2007]) and TargetP (Emanuelsson *et al.* [2000]) employ sequence sorting signals (Bannai *et al.* [2002]) while Cell-Ploc (Chou and Shen [2008]) and LOCSVMPSI (Xie *et al.* [2005]) use position specific scoring matrix (Sinha [2006]). Additionally, amino/psudo-amino acid composition information (Nakashima and Nishikawa [1994], Chou and Cai [2003]) is utilized by ngLOC (King and Guda [2007]). There are also some methods that employ combinations of sequence based features (Höglund *et al.* [2006], Horton *et al.* [2007]). Alongside, there are different types of annotation derived features such as protein-protein interaction, Gene Ontology (GO) terms and functional domain and motifs which are used by different methods (Höglund *et al.* [2006], Chou and Shen [2007], Lee *et al.* [2008], Huang *et al.* [2008], Shin *et al.* [2009], Zhou *et al.* [2016], Cheng *et al.* [2017], Mehrabad *et al.* [2018]).
Parallel to the importance of features, selecting a suitable algorithm definitely leads to a higher accuracy in prediction. Many machine learning methods or statistical inferences are applied for the protein subcellular localization problem, such as support vector machine (Höglund *et al.* [2006], Zhou *et al.* [2016]), K-nearest neighbors (Xiao *et al.* [2011], He *et al.* [2012]), and Bayesian methods (Briesemeister *et al.* [2010], Simha and Shatkay [2014]). In this paper, we have modeled the problem of protein subcellular localization as a recommendation task that aims to suggest a list of subcellular locations to a new protein. In general, Recommender systems are methods and techniques that suggest users a preferred list of items (e.g. suggesting a movie to watch or suggesting an item to purchase) based on a previous knowledge about relations within

and between items and users (Francesco *et al.* [2011]). Our method, "PSL-Recommender", employs a probabilistic recommender system to predict the presence probability of a protein in a subcellular location.
PSL-Recommender employs logistic matrix factorization technique integrated with a neighborhood regularization method to capture the information from a set of previously known protein-subcellular location relations. Then, it utilizes this information to predict the presence probability of a new protein in a subcellular location using a logistic function. Logistic Matrix factorization was first introduced by (Johnson [2014]) for collaborative filtering. This technique has shown promising results for problems such as drug-target interaction prediction (Liu *et al.* [2016], Hao *et al.* [2017]) and lncRNA-protein interaction prediction (Liu *et al.* [2017]). However, to the best of our knowledge, it has not been used in PSL prediction problem.
By evaluating on different benchmark datasets, we have shown that PSL-Recommender significantly outperforms the current state-of-art methods.

2 Materials and Method

2.1 Method

To recommend a subcellular position to a protein, PSL-Recommender employs two matrices; a matrix of currently known protein-subcellular location assignments(PSL interactions) and a similarity matrix between proteins. The proteins similarity matrix is the weighted average of similarity measures such as GO terms(Ashburner *et al.* [2000]) similarities, PSSM(Stormo *et al.* [1982]) similarity and STRING(Szklarczyk *et al.* [2014]) similarity.
The main idea is to model the localization probability of a protein in a location as a logistic function of two latent matrices. The latent matrices are acquired by matrix factorization of the protein-subcellular location matrix according to the similarity matrices. Construction pipeline of PSL-Recommender predictor has been demonstrated in Fig 2.1. The details of similarity measures and the recommender system are as follows.

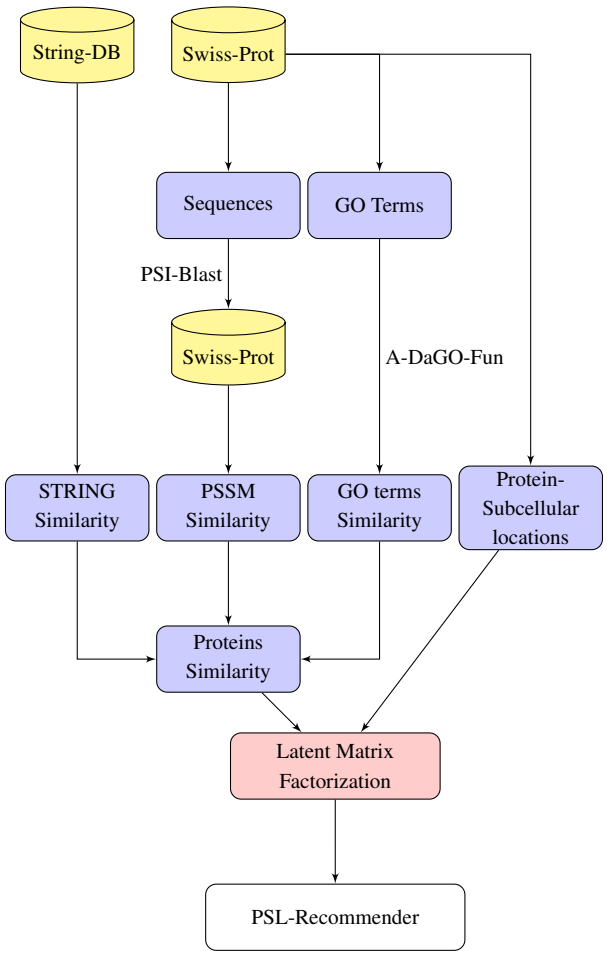

Fig. 1: Construction pipeline of PSL-Recommender

2.1.1 PSSM similarity

The PSSM similarity matrix, $S^{PSSM} = [s_{i,j}^{PSSM}]_{n\times n}$, contains the pairwise global alignment scores of proteins that are calculated using the position specific scoring matrices (PSSM). Accordingly, to compute the $s_{i,j}^{PSSM}$ of proteins i and j, first for each protein, PSI-BLAST(Altschul *et al.* [1997]) with e-value 0.001 is used to search the Swiss-Prot database to obtain each protein's PSSM. Then i and j are globally aligned twice, once using the PSSM of i and once using the PSSM of j. Finally, $s_{i,j}^{PSSM}$ is obtained by the mean of reciprocal alignment scores. The PSSM similarity matrix is normalized using unity based normalization.

2.1.2 STRING similarity

It has been shown that two interacting proteins have a higher chance to be in the same subcellular location (Scott *et al.* [2005], Lee *et al.* [2008], Mehrabad *et al.* [2018]). Accordingly, we extracted the interaction score of all pairs of proteins from STRING (Ver. 10.5) to construct the proteins interaction scoring matrix. If no interaction was available for a pair of proteins, we set their interaction score to zero. Since the STRING protein-protein interaction scores are in the range of [0, 999], we normalized the scores with unity-base normalization.

2.1.3 Semantic similarity of GO terms

Gene Ontology terms are valuable sources of information for predicting subcellular localization (Sayers *et al.* [2009], Zhou *et al.* [2016]). To exploit GO terms similarities, we first extracted GO molecular function, biological process and cellular component terms from Swiss-Prot database. Then we used A-DaGO-Fun to extract the BMA-based Resnik GO terms semantic similarities(Mazandu *et al.* [2015]). Similarities were normalized using unity-based normalization.

2.1.4 PSL-Recommender

Let proteins and subcellular locations sets be denoted by X and Y, respectively and $|X| = m$ and $|Y| = n$. Moreover, let $S^p = \left[s_{i,k}^p\right]_{m\times m}$ represent the similarity of proteins. The presence of proteins in subcellular locations is also denoted by a binary matrix $L = [l_{ij}]_{m\times n}$, where, $l_{ij} = 1$ if proteins i has been experimentally observed in subcellular location j and $l_{ij} = 0$ otherwise.
The localization probability of the protein i in subcellular location j can be modeled as a logistic function as follows:

$$p_{ij} = \frac{\exp\left(u_i v_j^T + \beta_i^p + \beta_j^l\right)}{1 + \exp\left(u_i v_j^T + \beta_i^p + \beta_j^l\right)}. \quad (1)$$

In Eq.(1), $u_i \in \mathbb{R}^{1\times \mathrm{d}}$ and $v_j \in \mathbb{R}^{1\times \mathrm{d}}$ are two latent vectors that reflect the properties of protein i and subcellular location j in a shared latent space of size $d < min\,(m, n)$. However, in our case matrix L is biased toward some proteins and subcellular locations, meaning that some proteins tend to localize in many locations and some subcellular locations include many proteins. Accordingly, for each protein and subcellular location we introduce a latent term to capture this bias. In Eq.(1), β_i^p

represent the bias factor for protein i and β_j^l represent the bias factor for subcellular location j.
Now the goal is to acquire the latent factors for a given L. Suppose $U \in \mathbb{R}^{m\times d}$, $V \in \mathbb{R}^{n\times d}$, $\beta^p \in \mathbb{R}^{m\times 1}$ and $\beta^l \in \mathbb{R}^{n\times 1}$ denote the latent matrices and bias vectors for proteins and subcellular locations. According to the Bayes' theorem and the independence of U and V we have:

$$p(U,V,\beta^p,\beta^l|L) \propto p(L|U,V,\beta^p,\beta^l) \times p(U) \times p(V) \tag{2}$$

On the other hand, by assuming that all entries of L are independent, we have:

$$p\left(L|U,V,\beta^p,\beta^l\right) = \prod_{i=1}^{m}\prod_{j=1}^{n} {p_{ij}}^{cl_{ij}}(1-p_{ij})^{(1-l_{ij})}, \tag{3}$$

where c is weighting factor on positive observations, since we have more confidence on positive observations than negative ones. Also, by placing a zero-mean spherical Gaussian prior on latent vectors of proteins and subcellular locations we have:

$$p\left(U|\sigma_p^2\right) = \prod_{i=1}^{m}\mathcal{N}\left(u_i|0,\sigma_p^2 I\right), p\left(V|\sigma_l^2\right) = \prod_{j=1}^{n}\mathcal{N}\left(v_j|0,\sigma_l^2 I\right), \tag{4}$$

where σ_p^2 and σ_l^2 are parameters controlling the variances of prior distributions and I denotes the identity matrix. According to the above equations, the log of the posterior is yielded as follows:

$$\begin{aligned}\log p\left(U,V,\beta|L,\sigma_p^2,\sigma_l^2\right) &= \sum_{i=1}^{m}\sum_{j=1}^{n} cl_{ij}\left(u_i v_j^T + \beta_i + \beta_j\right)\\ &- (1+cl_{ij}-l_{ij})\, log\left[1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)\right]\\ &- \frac{\lambda_p}{2}\sum_{i=1}^{m}\|u_i\|_2^2 - \frac{\lambda_l}{2}\sum_{j=1}^{n}\|v_j\|_2^2 + C,\end{aligned} \tag{5}$$

where $\lambda_p = \frac{1}{\sigma_p^2}$, $\lambda_l = \frac{1}{\sigma_l^2}$ and c is a constant term independent of the model parameters. Our goal is to learn U, V, β^p and β^l that maximize the log posterior above, which is equal to minimizing the following objective function:

$$\begin{aligned}\min_{U,V,\beta^p,\beta^l} &\sum_{i=1}^{m}\sum_{j=1}^{n}(1+cl_{ij}-l_{ij})\, log\left[1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)\right]\\ &-cl_{ij}\left(u_i v_j^T + \beta_i + \beta_j\right) + \frac{\lambda_p}{2}\sum_{i=1}^{m}\|u_i\|_F^2 - \frac{\lambda_l}{2}\sum_{j=1}^{n}\|v_j\|_F^2,\end{aligned} \tag{6}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix. By minimizing the above function U, V and β can effectively capture the information of protein localizations. However, we can further improve the model by incorporating protein similarities as suggested by (Liu *et al.* [2016]). This process is known as neighborhood regularization. This is done by regularizing the latent vectors of proteins such that the distance between a protein and its similar proteins is minimized in the latent space.
Accordingly, suppose that the set of k_1 most similar neighbors to protein x_i is denoted by $N_{k_1}(x_i) \subseteq X - x_i$. We constructed adjacency matrix $A = [a_{ij}]_{m\times m}$ that represents proteins neighborhood information as follows:

$$a_{ij} = \begin{cases} s_{ij}^p & if \quad x_j \in N(x_i)\\ 0 & otherwise\end{cases} \tag{7}$$

To minimize the distance between proteins and their k most similar proteins we minimize the following objective function:

$$\frac{\alpha}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}\|u_i - u_j\|_F^2 = \frac{\alpha}{2} tr\left(U^T H^p U\right), \tag{8}$$

where $H^p = \left(B^p + \tilde{B}^p\right) - \left(A + A^T\right)$ and $tr(.)$ is the trace of matrix. In this equation, B^p and $\tilde{B}^p$ are two diagonal matrices, that their diagonal elements are $B_{ii}^p = \sum_{j=1}^{m} a_{ij}$ and $\tilde{B}_{jj}^p = \sum_{i=1}^{m} a_{ij}$, respectively.
Finally by plugging Eq.(8) into Eq.(6) we will have the following:

$$\begin{aligned}\min_{U,V,\beta^p,\beta^l} &\sum_{i=1}^{m}\sum_{j=1}^{n} cl_{ij}\left(u_i v_j^T + \beta_i + \beta_j\right)\\ &- (1+cl_{ij}-l_{ij})\, log\left[1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)\right]\\ &- \frac{1}{2} tr\left[U^t\left(\lambda_p I + \alpha H^p\right)U\right] - \frac{\lambda_l}{2}\sum_{j=1}^{n}\|v_j\|_F^2.\end{aligned} \tag{9}$$

A local minimum of above function can be found by employing the alternating gradient descent method. In each iteration of the gradient descent, first U and β_i are fixed to compute V and β_j and then V and β_j are fixed to compute U and β_i. To accelerate the convergence, we have employed the AdaGrad(Duchi *et al.* [2011]) algorithm to choose the gradient step size in each iteration adaptively. The partial gradients of latent vectors and biases are given by:

$$\begin{aligned}\frac{\partial F}{\partial u_i} &= \sum_{j=1}^{n} cl_{ij} v_j^T - \frac{v_j^T(1+cl_{ij}-l_{ij})\, exp\left(u_i v_j^T + \beta_i + \beta_j\right)}{1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)}\\ &- (\lambda_p u_i + \alpha H_{ij}^p u_i),\\ \frac{\partial F}{\partial \beta_i} &= \sum_{j=1}^{n} cl_{ij} - \frac{(1+cl_{ij}-l_{ij})\, exp\left(u_i v_j^T + \beta_i + \beta_j\right)}{1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)},\\ \frac{\partial F}{\partial v_j} &= \sum_{i=1}^{m} cl_{ij} u_i - \frac{u_i(1+cl_{ij}-l_{ij})\, exp\left(u_i v_j^T + \beta_i + \beta_j\right)}{1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)}\\ &- \lambda_l v_j,\\ \frac{\partial F}{\partial \beta_j} &= \sum_{i=1}^{m} cl_{ij} - \frac{(1+cl_{ij}-l_{ij})\, exp\left(u_i v_j^T + \beta_i + \beta_j\right)}{1+exp\left(u_i v_j^T + \beta_i + \beta_j\right)}.\end{aligned} \tag{10}$$

Once the latent matrices U, V, β_i and β_j are calculated, the presence probability of a protein i in a subcellular location can be estimated by the logistic function in formula 1. However for a new protein the latent factors u and b are not available. Hence, for a new protein the presence probability in subcellular location j is estimated as follows:

$$p_{ij} = \frac{\exp\left(\tilde{u}_i v_j^T + \beta_i^p + \beta_j^l\right)}{1+\exp\left(\tilde{u}_i v_j^T + \beta_i^p + \beta_j^l\right)}, \tag{11}$$

where $\tilde{u}_i$ is the weighted average of the latent vectors of k_2 nearest neighbors of i, as follows:

$$\tilde{u}_i = \frac{\sum_{k\in N(x_i)} s_{i,k}^p u_k}{\sum_{k\in N(x_i)} s_{i,k}^p}. \tag{12}$$

Eventually a threshold can be applied on probabilities to assign the subcellular locations to proteins.

2.2 Datasets and evaluation criteria

Evaluating the protein subcellular prediction methods is a challenging task. In one hand, the standalone version of state-of-the-art methods are not available and on the other hand, the protein databases are updated quickly. Hence, to achieve a fair evaluation and comparison we have employed the same datasets and evaluation criteria as used in previous studies (Zhou *et al.* [2016], Briesemeister *et al.* [2010]). These datasets are summarized in Table 1. The Hum-mploc3.0, the BaCelLo animals, and the Höglund datasets consist of two non-overlapping subsets for training and testing purposes while for DBMloc we have performed 5-fold cross validation.
The training set of Hum-mploc 3.0, HumB, is constructed from Swiss-Prot database release 2012_01 (January 2012) and consists of 3122 proteins of which 1023 proteins are labeled with more than one subcellular locations and the rest are single location proteins. Alongside HumB, HumT is used as the testing set to evaluate the method's performance. HumT is also constructed from Swiss-Prot database release 2015_05 (May 2015 release) and consists of 379 proteins of which 120 proteins are labeled with more than one subcellular locations and the rest are single location proteins. Each protein in Hum-mploc 3.0 is assigned to at least one of 12 subcellular locations (Centrosome, Cytoplasm, Cytoskeleton, Endoplasmic reticulum, Endosome, Extracellular, Golgi apparatus, Lysosome, Mitochondrion, Nucleus, Peroxisome, and Plasma membrane).
The training set of BaCelLo animals dataset is extracted from Swiss-Prot release 48 (September 2005 release) containing 2597 single label proteins, while the testing set consists of 576 single label proteins extracted from Swiss-Prot between relese 49 and 54 (February 2006 and July 2007 releases). Each protein in BaCelLo animal dataset is assigned one of four subcellular locations (Cytoplasm, Mitochondrion, Nucleus, and Secreted). In the Höglund dataset, the training set contains 5959 single label proteins extracted from Swiss-Prot release 42 and includes nine subcellular locations (Nucleus, Cytoplasm, Mitochondrion, Endoplasmic reticulum, Golgi apparatus, Peroxisome, Plasma membrane, Extracellular space, Lysosome, and Vacuole) while the testing set contains 158 single label proteins extracted from Swiss-Prot release 55.3 including six subcellular locations (Endoplasmic reticulum, Golgi apparatus, Peroxisome, Plasma membrane, Extracellular space, Lysosome). Accordingly, to train PSL-Recommender we only used 2682 proteins of training set that their subcellular location existed in the test set.
Unlike the previous datasets, the DBMLoc dataset does not have a separate training and testing dataset. This dataset contains 3054 double locational proteins with paired subcellular locations: (cytoplasm and nucleus), (extracellular and plasma membrane), (cytoplasm and plasma membrane), (cytoplasm and mitochondrion), (nucleus and mitochondrion), (endoplasmic reticulum and extracellular) and (extracellular and nucleus). We have performed 5-fold cross validation technique to produce training and testing sets on this dataset.

We assessed PSL-Recommender performance against other methods by using customized ACC and $\mathrm{F1-mean}$ over subcellular locations for evaluation of multi-label classification performance methods which is introduced by (Tsoumakas and Katakis [2007]) and used by other state-of-the-art methods for this problem. ACC is the average of ACC_{x_i} of all proteins in the test set, calculated for each protein as follows:

$$\mathrm{ACC_{x_i}} = \frac{TP_{x_i}}{TP_{x_i} + FP_{x_i} + FN_{x_i}}, \tag{13}$$

where, TP, FP and FN stand for true positive, false positive and false negative, respectively.
The $\mathrm{F1-mean}$ is the average of $\mathrm{F1_{y_j}}$ of all subcellular locations, where $\mathrm{F1}$ of subcellular location y_j is the harmonic mean of $\mathrm{Precision_{y_j}}$ and $\mathrm{Recall_{y_j}}$, defined as follows:

$$\begin{aligned}
\mathrm{Precision_{y_j}} &= \frac{\sum_{\{x_i \in R_j\}} \frac{TP_{x_i}}{TP_{x_i}+FP_{x_i}}}{|R_j|}, \\
\mathrm{Recall_{y_j}} &= \frac{\sum_{\{x_i \in T_j\}} \frac{TP_{x_i}}{TP_{x_i}+FN_{x_i}}}{|T_j|}, \\
\mathrm{F1_{y_j}} &= \frac{2 \times \mathrm{Precision_{y_j}} \times \mathrm{Recall_{y_j}}}{\mathrm{Precision_{y_j}} + \mathrm{Recall_{y_j}}},
\end{aligned} \tag{14}$$

where, R_j and T_j are sets of predicted proteins for location y_j and true proteins for location y_j, respectively.

2.2.1 Learning parameters

Minimizing formula 6 in section 2.1 depends on 13 parameters. The parameters were picked empirically for each dataset by maximizing the $\mathrm{F1-mean}$. Due to the large search space, we used a greedy grid search method for selecting the parameters.
The weight of similarity measures used to build the protein similarity matrix was picked from 1 to 10 by step of 1. The dimension of latent space, r, was selected between 1 and the number of subcellular locations by step of 1. The weighting factor for positive observations, c, was chosen between 5 and 80 by step of 1. The number of nearest neighbors for constructing $N_{k_1}(x_i)$ in equation 7, k_1, was selected from 1 to 60 by step of 1. Similarly, The number of nearest neighbors for constructing $N_{k_2}(x_i)$, in equation 12, k_2, was selected from 1 to 60 by step of 1. The variance controlling parameters, λ_p and λ_l, were chosen between 0 and 1 by step of 0.1. Impact factor of nearest neighbors in equation 8, α, was picked from 0.1 to 1 by step of 0.1. Finally, The learning rate of the gradient descent criteria, θ, was selected from 0.1 to 1 by step of 0.1.
All learned parameters for each dataset is available in the code.

Table 1. Dataset summary

	Hum-mPLoc 3.0		BaCelLo		Höglund		DBMloc
	Train	Test	Train	Test	Train	Test	All
Proteins count	3126	379	2597	576	5959	158	3054
Labels count	4229	541	2597	576	5959	158	6108
Locations count	12		4		6		6

3 Results and discussion

PSL-Recommender can be employed to predict the subcellular protein localization in different species. Accordingly, we evaluated the performance of PSL-Recommender on different datasets and compared it to other state-of-the-arts methods. We further investigated the role of each protein similarity measures that are employed by the PSL-Recommender.

3.1 Comparison with the State-of-art method

We have first employed the Hum-mPLoc 3.0 (Zhou *et al.* [2016]) human protein dataset to compare the performance of PSL-Recommender to six methods that were introduced for protein localization in human. The methods include YLoc+ (Briesemeister *et al.* [2010]), iLoc-Hum (Chou *et al.* [2012]), WegoLoc (Chi and Nam [2012]), mLASSO-Hum (Wan *et al.* [2015]) and Hum-mPloc 3.0. The $\mathrm{F1-score}$ for each location and the ACC and $\mathrm{F1-mean}$ of all methods on Hum-mploc 3.0 dataset is depicted in Table 2.
As seen in Table 2, PSL-Recommender significantly outperforms the

Table 2. Comparison of PSL-Recommender on Human proteins dataset(Hum-mPloc 3.0) with other methods

Location	Yloc+			iLoc-Human			WegoLoc			mLASSO-Hum			Hum-mPLoc3.0			PSL-Recommender		
	pre	re	F1	pre	re	F1	pre	re	F1	pre	re	F1	pre	re	F1	pre	re	F1
Centrosome	-	-	-	0	0	0	0.75	0.14	0.23	0.59	0.59	0.59	0.75	0.55	0.63	0.94	0.75	**0.83**
Cytoplasm	0.55	0.85	0.67	0.5	0.54	0.52	0.69	0.53	0.60	0.93	0.51	0.66	0.76	0.73	0.74	0.79	0.81	**0.80**
Cytoskeleton	-	-	-	0	0	0	0.32	0.34	0.33	0.9	0.22	0.35	0.8	0.68	0.74	0.93	0.77	**0.84**
ER	0.71	0.12	0.21	0	0	0	0.73	0.2	0.31	0.74	0.49	0.59	0.83	0.37	0.51	0.9	0.7	**0.79**
Endosome	-	-	-	0	0	0	0.25	0.07	0.11	0.38	0.2	0.26	0.58	0.47	**0.52**	0.57	0.37	0.45
Extracellular	0.39	0.85	0.54	0.62	0.62	0.62	0.67	0.77	**0.71**	0.16	0.69	0.26	0.5	0.46	0.48	0.66	0.71	0.68
Golgi apparatus	0.1	0.05	0.07	0.6	0.3	0.4	0.6	0.15	0.24	0.72	0.65	0.68	0.69	0.45	0.55	0.88	0.61	**0.72**
Lysosome	0	0	0	0.5	0.13	0.2	0.2	0.13	0.15	0.55	0.75	0.63	0.71	0.63	0.67	1	0.55	**0.71**
Mitochondrion	0.65	0.43	0.52	0.95	0.33	0.49	0.79	0.73	0.76	0.83	0.88	0.85	0.78	0.75	0.76	0.92	0.88	**0.90**
Nucleus	0.41	0.57	0.48	0.54	0.7	0.61	0.65	0.64	0.64	0.85	0.7	0.76	0.75	0.71	0.73	0.81	0.92	**0.87**
Peroxisome	0.07	0.5	0.13	1	0.5	0.67	0.5	1	0.67	0.29	1	0.44	1	1	**1**	1	1	**1**
Plasma membrane	0.41	0.44	0.42	0.42	0.33	0.37	0.44	0.53	0.48	0.58	0.56	0.57	0.65	0.44	0.52	0.78	0.74	**0.76**
ACC		0.45			0.41			0.50			0.65			0.63			**0.77**	
F1-mean		0.34			0.32			0.44			0.56			0.65			**0.78**	

$\mathrm{F1-mean}$ and ACC of all other methods improving the best method by 13% in $\mathrm{F1-mean}$ and 12% in ACC. Also, in 10 out of 12 subcellular locations, PSL-Recommender has the best performance amongst all methods while in the other two locations it has the second best performance. The most significant improvements have been observed in Centrosome, ER (Endoplasmic Reticulum) and Plasma Membrane showing 20%, 20% and 19% improvement respectively over the second best method.

It is only in Endosome that PSL-Recommender shows unsatisfactory results (45% $\mathrm{F1-score}$). This is while other methods also fail to provide good results for this location such that the best method (Hum-mPLOC 3.0) only achieves 52% $\mathrm{F1-score}$. Moreover, for Extracellular, WegoLoc slightly (3%) outperforms PSL-Recommender.

To show the performance of PSL-Recommender on other species we have employed previously introduced datasets that include proteins from animals, plants and eukaryotes. We then compared the results to five state-of-the-art methods including (Briesemeister *et al.* [2010], Blum *et al.* [2009], Pierleoni *et al.* [2006], Mehrabad *et al.* [2018], Zhou *et al.* [2016]). The results are depicted in Table 3.

As seen in Table 3, PSL-Recommender outperforms all methods in all datasets by both $\mathrm{F1-mean}$ and ACC. In Höglund dataset, PSL-Recommender significantly outperforms the second best method by 31% and 28% in $\mathrm{F1-mean}$ and ACC respectively. In BaCelLo dataset, the improvement over the second best method is 8% in both ACC and $\mathrm{F1-mean}$ while in DBMloc dataset, PSL-Recommender slightly improves the second best method by 2% in both $\mathrm{F1-mean}$ and ACC.

It also worth mentioning that, for PSL prediction problem, to the best of our knowledge, PMLPR (Mehrabad *et al.* [2018]) is the only recommender system based method that employs the well-known network-based inference(NBI) (Zhou *et al.* [2007]) approach. As seen in Table 3, PSL-Recommender outperforms PMLPR by 52% and 19% in $\mathrm{F1-mean}$, and also 28% and 17% in ACC on Höglund and DBMloc datasets.

Table 3. Comparison of PSL-Recommender $\mathrm{ACC/F1-mean}$ on other species proteins datasets with other methods

	BaCelLo	Höglund	DBMloc
YLoc-LowRes	0.79/0.75	-	-
YLoc-HighRes	0.74/0.69	0.56/0.34	-
YLoc+	0.58/0.67	0.53/0.37	0.64/0.68
MultiLoc2-LowRes	0.73/0.76	-	-
MultiLoc2-HighRes	0.68/0.71	0.57/0.41	-
BaCelLo	0.64/0.66	-	-
PMLPR	-	0.64/0.38	0.72/0.67
Hum-mPloc 3.0	0.86/0.84	0.64/0.59	0.87/0.84
PSL-Recommender	**0.94/0.92**	**0.92/0.90**	**0.89/0.86**

Table 4. Comparisons of $ACC/\mathrm{F1-mean}$ by using different features

Features	BaCelLo	Höglund	DBMLoc	Hum-mPloc 3.0
PSSM	0.72/0.58	0.63/0.32	0.88/0.85	0.33/0.32
STRING	-	-	-	0.44/0.47
GO	0.94/0.93	0.92/90	0.88/0.85	0.79/0.79
PSSM+STRING	-	-	-	0.46/49
GO +PSSM	0.94/0.92	0.92/0.90	0.89/0.86	0.79/0.79
GO+STRING	-	-	-	0.78/0.79
All	0.94/0.95	0.92/0.90	0.89/0.86	0.77/0.78

3.2 Impact of each similarity matrix

The proteins similarity matrix is used for neighborhood regularization and also the prediction step. To acquire this matrix PSL-Recommender combines three sources of protein similarity measures (PSSM similarity, String-DB interactions similarity and GO terms semantic similarity) using weighted averaging. The weights are acquired through the learning process.

To investigate the impact of different similarity measures, we repeated previous experiments using different combination of similarity measures. Table 4. shows the result of each combination on all datasets. As can be seen in Table 4., those combinations including the GO terms semantic similarities do not provide reliable predictions showing that GO terms semantic similarities play an important role in protein subcellular localization.

It should be noted that GO terms are not available for all proteins. In the absence of GO terms semantic similarities, PLS-Recommender is still able to provide acceptable results for DBMLoc and BacelLo datasets but its performance significantly drops for Hoglund and Hum-mPloc 3.0.

Moreover, the usage of String protein-protein interaction scores is only limited to datasets that contain proteins from single species. Since DBMLoc, BacelLo and Hoglund datasets contain proteins from multiple species we were unable to use String interaction scores in these datasets.

4 Conculusion

In the absence of efficient experimental methods, computational tools play an important role for predicting protein subcellular localizations. Yet, there is still much room for improving the prediction accuracy of these methods. In this paper, we introduced PSL-Recommender, a recommender system that employs logistic matrix factorization for efficient prediction of protein subcellular localization. By evaluating on human and animals datasets it was shown that PSL-Recommender significantly outperforms other state-of-the-art methods. However, we believe that the performance of PSL-Recommender can be improved further by employing a better approach for searching the parameter space. The standalone version of PSL-Recommender and all the datasets are available online at: https://github.com/RJamali/PSL-Recommender

Acknowledgements

This work is supported by Iran National Science Foundation.

References

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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics*, **18**(2), 298–305.

Blum, T., Briesemeister, S., and Kohlbacher, O. (2009). Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC bioinformatics*, **10**(1), 274.

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Going from where to whyâŁ"interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**(9), 1232–1238.

Cheng, X., Xiao, X., and Chou, K.-C. (2017). ploc-mhum: predict subcellular localization of multi-location human proteins via general pseaac to winnow out the crucial go information. *Bioinformatics*, **34**(9), 1448–1456.

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

Chou, K.-C. and Cai, Y.-D. (2003). Prediction and classification of protein subcellular locationâŁ"sequence-order effect and pseudo amino acid composition. *Journal of cellular biochemistry*, **90**(6), 1250–1260.

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

Chou, K.-C. and Shen, H.-B. (2008). Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nature protocols*, **3**(2), 153.

Chou, K.-C., Wu, Z.-C., and Xiao, X. (2012). iloc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Molecular Biosystems*, **8**(2), 629–641.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, **12**(Jul), 2121–2159.

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

Francesco, R., Lior, R., and Bracha, S. (2011). Introduction to recommender systems handbook, recommendersystems handbook.

Hao, M., Bryant, S. H., and Wang, Y. (2017). Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Scientific reports*, **7**, 40376.

He, J., Gu, H., and Liu, W. (2012). Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PloS one*, **7**(6), e37155.

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

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. (2007). Wolf psort: protein localization predictor. *Nucleic acids research*, **35**(suppl_2), W585–W587.

Huang, W.-L., Tung, C.-W., Ho, S.-W., Hwang, S.-F., and Ho, S.-Y. (2008). Proloc-go: utilizing informative gene ontology terms for sequence-based prediction of protein subcellular localization. *BMC bioinformatics*, **9**(1), 80.

Johnson, C. C. (2014). Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, **27**.

King, B. R. and Guda, C. (2007). ngloc: an n-gram-based bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome biology*, **8**(5), R68.

Lee, K., Chuang, H.-Y., Beyer, A., Sung, M.-K., Huh, W.-K., Lee, B., and Ideker, T. (2008). Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research*, **36**(20), e136–e136.

Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., and Zhao, Q. (2017). Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget*, **8**(61), 103975.

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology*, **12**(2), e1004760.

Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M., and Mulder, N. J. (2015). A-dago-fun: an adaptable gene ontology semantic similarity-based functional analysis tool. *Bioinformatics*, **32**(3), 477–479.

Mehrabad, E. M., Hassanzadeh, R., and Eslahchi, C. (2018). Pmlpr: A novel method for predicting subcellular localization based on recommender systems. *Scientific reports*, **8**(1), 12006.

Nair, R. and Rost, B. (2003). Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins: Structure, Function, and Bioinformatics*, **53**(4), 917–930.

Nakai, K. and Horton, P. (1999). Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in biochemical sciences*, **24**(1), 34–35.

Nakashima, H. and Nishikawa, K. (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *Journal of molecular biology*, **238**(1), 54–61.

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

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., *et al.* (2009). Database resources of the national center for biotechnology information. *Nucleic acids research*, **38**(suppl_1), D5–D16.

Scott, M. S., Calafell, S. J., Thomas, D. Y., and Hallett, M. T. (2005). Refining protein subcellular localization. *PLoS computational biology*, **1**(6), e66.

Shin, C. J., Wong, S., Davis, M. J., and Ragan, M. A. (2009). Protein-protein interaction as a predictor of subcellular location. *BMC systems biology*, **3**(1), 28.

Simha, R. and Shatkay, H. (2014). Protein (multi-) location prediction: using location inter-dependencies in a probabilistic framework. *Algorithms for Molecular Biology*, **9**(1), 8.

Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*, **22**(14), e454–e463.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the âŁ˜perceptronâŁ™algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research*, **10**(9), 2997–3011.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., *et al.* (2014). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447–D452.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, **3**(3), 1–13.

Wan, S., Mak, M.-W., and Kung, S.-Y. (2015). mlasso-hum: A lasso-based interpretable human-protein subcellular localization predictor. *Journal of theoretical biology*, **382**, 223–234.

Xiao, X., Wu, Z.-C., and Chou, K.-C. (2011). iloc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *Journal of theoretical biology*, **284**(1), 42–51.

Xie, D., Li, A., Wang, M., Fan, Z., and Feng, H. (2005). Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using svm and profile of psi-blast. *Nucleic acids research*, **33**(suppl_2), W105–W110.

Zhou, H., Yang, Y., and Shen, H.-B. (2016). Hum-mploc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics*, **33**(6), 843–853.

Zhou, T., Ren, J., Medo, M., and Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, **76**(4), 046115.