PSL-Recommender: Protein Subcellular Localization Prediction using Recommender System

Ruhollah Jamali

Institute for Research in Fundamental Sciences(IPM)

Soheil Jahangiri-Tazehkand

University Health Network

Changiz Eslahchi (✉ ch-eslahchi@sbu.ac.ir)

Shahid Beheshti University

Research Article

Keywords: Protein Subcellular Localization Prediction, labor-intensive wet-lab methods, protein subcellular localization

Posted Date: September 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-878139/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

PSL-Recommender: Protein Subcellular Localization Prediction using Recommender System

Ruhollah Jamali¹, Soheil Jahangiri-Tazehkand², and Changiz Eslahchi^{3,*}

¹School of Biological Sciences, Institute for Research in Fundamental Sciences(IPM), Tehran, Iran

²Princess Margaret Cancer Centre, University Health Network, 101 College Street, Toronto, ON, M5G1L7, Canada

³Department of Computer Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

*corresponding author ch-eslahchi@sbu.ac.ir

ABSTRACT

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 labor-intensive wet-lab methods that are used for protein subcellular localization. Yet, there is still much room for improving the prediction accuracy of these methods.

In this article, we meant to develop a customized computational method rather than using common machine learning predictors, which are used in the majority of computational research on this topic. The neighbourhood regularized logistic matrix factorization technique was used to create PSL-Recommender (Protein subcellular location recommender), a GO-based predictor. We declared statistical inference as the driving force behind the PSL-Recommender here. Following that, it was benchmarked against twelve well-known methods using five different datasets, demonstrating outstanding performance. Finally, we discussed potential research avenues for developing a comprehensive prediction tool for protein subcellular location prediction.

The datasets and codes are available at: <https://github.com/RJamali/PSL-Recommender>

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 interactions¹. Recent advances in genomics and proteomics provide a massive amount of protein sequence data extending 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 manner². 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^{2-5,5-24}.

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 a better performance. Different types of sequence-derived features have been used for PSL prediction. For example, PSORT²⁵, WoLF PSORT⁴ and TargetP²⁶ employ sequence sorting signals⁹ while Cell-Ploc¹⁰ and LOCSVMPSI¹¹ use position-specific scoring matrix²⁷. Additionally, amino/pseudo-amino acid composition information^{12,28} is utilized by ngLOC¹³. There are also some methods that employ combinations of sequence-based features^{3,4}. 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^{2,7,8,17-20}. Moreover, text-based features derived by literature mining have also been employed beside other features for protein subcellular localization¹⁴⁻¹⁶. There is also recent imaging-based approach to predict PSL such as²³.

Parallel to the importance of features, selecting a suitable algorithm leads to higher accuracy in prediction. Many machine learning methods or statistical inferences are applied for the protein subcellular localization problem, such as support vector machine^{2,3}, K-nearest neighbors^{21,22}, Bayesian methods^{6,24}, and Deep Learning methods^{23,29}. One recent review³⁰ summarized the research progress of human PSL prediction tools.

In this paper, we have modeled the PSL prediction problem as a recommendation task that aims to suggest a list of subcellular locations to a new protein. In general, recommendation 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 previous knowledge about relations within and between items and users³¹. As of late framework strategies, recommendation systems have been utilized to predict associations in challenging bioinformatics problems³²⁻³⁵. Well-known PSL prediction methods assign equal importance

to all proteins information in both constructing model and prediction tasks^{2,5,6,36}, but utilization of prioritized information from similar proteins in model construction step is likely more meaningful. Additionally, due to a large number of protein features, dimension reduction methods which capture dependencies among proteins and subcellular locations could be useful to construct a PSL prediction model. To consider these concepts, our method, "PSL-Recommender" employs a probabilistic recommender system to predict the presence probability of a protein in a subcellular location. PSL-Recommender utilized both prioritized information to elucidate the importance of sharing similarity information over proteins and low-dimensional latent space projection of protein features during the PSL prediction process.

PSL-Recommender employs logistic matrix factorization technique³⁷ integrated with a neighbourhood 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 has shown promising results for problems such as music recommendation³⁸, drug-target interaction prediction^{32,39}, and lncRNA-protein interaction prediction³³. However, to the best of our knowledge, it has not been used in PSL prediction problem.

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 terms⁴⁰ similarities, PSSM⁴¹ similarity and STRING⁴² 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 with respect to the similarity matrices. Construction pipeline of PSL-Recommender predictor has been demonstrated in Fig 1. The details of similarity measures and the recommender system are as follows.

Figure 1. Construction pipeline of PSL-Recommender

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-BLAST⁴³ 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.

STRING similarity

It has been shown that two interacting proteins have a higher chance to be in the same subcellular location^{8,18,44}. 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-based normalization.

Semantic similarity of GO terms

Gene Ontology terms are valuable sources of information for predicting subcellular localization^{2,45}. 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 similarities⁴⁶. Similarities were normalized using unity-based normalization.

PSL-Recommender

Let proteins and subcellular locations sets be denoted by X and Y , respectively and $|X| = m$ and $|Y| = n$. Moreover, let $S^p = [s_{i,k}^p]_{m \times m}$ 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:

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

In equation(1), $u_i \in \mathbb{R}^{1 \times d}$ and $v_j \in \mathbb{R}^{1 \times d}$ 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 equation(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 matrix 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:

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

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

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

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:

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

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:

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

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:

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

where $\|\cdot\|_F$ 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 by³². This process is known as neighbourhood 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 neighbourhood information as follows:

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

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

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

where $H^p = (B^p + \tilde{B}^p) - (A + A^T)$ and $\text{tr}(\cdot)$ 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 equation(8) into equation(6) we will have the following:

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

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 AdaGrad⁴⁷ algorithm to choose the gradient step size in each iteration adaptively. The

partial gradients of latent vectors and biases are given by:

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

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:

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

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.

Results

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^{2,5,6,29}. These datasets are summarized in Table 1. The Hum-mploc3.0², the BaCelLo IDS animals⁴⁸, the Höglund IDS³⁶, and the DeepLoc²⁹ datasets consist of two non-overlapping subsets for training and testing purposes while for DBMloc we have performed 5-fold cross validation.

	Hum-mPLoc 3.0		BaCelLo		Höglund		DBMloc	DeepLoc	
	Train	Test	Train	Test	Train	Test	All	Train	Test
Proteins count	3129	379	2597	576	5959	158	3056	6782	1628
Labels count	4229	541	2597	576	5959	158	6112	6782	1628
Locations count	12		4		6		6	10	

Table 1. Datasets summary

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 IDS 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 release 49 and 54 (February 2006 and July 2007 releases). Each protein in BaCelLo IDS animal dataset is assigned one of four subcellular locations (Cytoplasm, Mitochondrion, Nucleus, and Secreted).

In the Höglund IDS 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, and Lysosome). Accordingly, to train PSL-Recommender we only used 2682 proteins of training set that their subcellular location existed in the test set.

The protein data used in the DeepLoc dataset were extracted from the Swiss-Prot database release 2016_04. Both training and testing sets contain only single label proteins which are located in ten subcellular locations (Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell membrane, Endoplasmic reticulum, Plastid, Golgi apparatus, Lysosome/Vacuole, and Peroxisome). The training set consists of 6728 proteins and the testing set has 1682 proteins.

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 F1 – mean over subcellular locations for evaluation of multi-label classification performance methods which is introduced by⁴⁹ 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:

$$ACC_{x_i} = \frac{TP_{x_i}}{TP_{x_i} + FP_{x_i} + FN_{x_i}}, \quad (13)$$

where, TP_{x_i} , FP_{x_i} , and FN_{x_i} are number of true positive prediction, number of false positive predictions, and number of false negative predictions for protein x_i , respectively.

The F1 – mean is the average of $F1_{y_j}$ of all subcellular locations, where F1 of subcellular location y_j is the harmonic mean of $Precision_{y_j}$ and $Recall_{y_j}$, defined as follows:

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

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

Alongside, SherLoc2⁵ applied two other evaluation criterias, named ACC2 (ratio of correctly predicted proteins) and AVG (average fraction of called instances) which are defined as follow:

$$\begin{aligned} ACC2 &= \frac{tp + tn}{tp + tn + fp + fn}, \\ AVG &= \frac{1}{d} \sum_{i=1}^d \frac{tp_i}{tp_i + fn_i}, \end{aligned} \quad (15)$$

where, d denotes the number of subcellular locations and tp , tn , fp , and fn indicate the number of true positive, true negative, false positive, and false negative instances, respectively.

Furthermore, another accuracy (ACC3) and Gorodkin measure⁵⁰ used to compare the performance of PSL-Recommender and DeepLoc, which are employed in DeepLoc paper.

Learning Hyperparameters

For all datasets, to prevent overfitting in tuning hyperparameters, they were learned from a comprehensive dataset, HumB, which is not considered in testing stage. It means that these hyperparameters are considered for all datasets. Since HumB -among the four mentioned datasets- contains both the single label and multi label PSL data, this dataset has been used for tuning task. Following 5-fold cross validation procedure is applied on HumB and hyperparameters were chosen empirically by maximizing the F1 – mean: HumB is divided into 5 equal subsets and PSL-Recommender is trained on union of 4 subsets and one other subset was hold for test the F1 – mean. This process is repeated 5 times, such that each time one of the 5 subsets is used as validation set and other 4 subsets are put together to form a training set.

For each set of hyperparameters, whole 5-fold process is repeated for 20 times and average of F1 – mean has been calculated. Due to the large search space, a grid-search procedure is applied for selecting the hyperparameters.

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 from $\{2^{-5}, 2^{-4}, \dots, 2^1\}$. Impact factor of nearest neighbors in equation 8, α , was picked from $\{2^{-5}, 2^{-4}, \dots, 2^2\}$. Finally, The learning rate of the gradient descent criteria, θ , was selected from $\{2^{-5}, 2^{-4}, \dots, 2^0\}$.

Table 2 represents the learned hyperparameters using HumB dataset. For all datasets, these learned hyperparameters are considered to construct the models.

	r	c	k_1	k_2	λ_p	λ_l	α	θ
Value	10	11	4	10	0.25	0.5	2	1

Table 2. Learned hyperparameters based on HumB dataset. (r is latent space dimension, c is weighting factor for positive observations, k_1 is the number of nearest neighbors for constructing $N_{k_1}(x_i)$, k_2 is the number of nearest neighbors for constructing $N_{k_2}(x_i)$, λ_p variance controlling parameter of proteins, λ_l variance controlling parameter of subcellular locations, α is the impact factor of nearest neighbors, and θ is the learning rate of the gradient descent criteria)

Evaluation

PSL-Recommender can be employed to predict the protein’s subcellular 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.

It is important to compare the performance of multiple different prediction algorithms fairly. While various state-of-the-art protein subcellular location predictors developed more than five years ago, having fair comparison could be a complicated task. Since they are trained with older data sets, it is understandable that new methods demonstrate better results, which could cause an illusion about a method’s performance.

First, we have employed the Hum-mPLOC 3.0² human protein dataset to compare the performance of PSL-Recommender to six methods that were introduced for protein localization in humans. The methods include YLoc+⁶, iLoc-Hum⁵¹, WegoLoc⁵², mLASSO-Hum⁵³ and Hum-mPloc 3.0. The F1 – score for each location and the ACC and F1 – mean of all methods on Hum-mPloc 3.0 dataset is depicted in Table 3. In which, PSL-Recommender significantly outperforms the F1 – mean and ACC of all other methods improving the best method by 12% in both F1 – mean and 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 17%, 21% and 18% improvement respectively over the second best method. It is only in Endosome that PSL-Recommender shows unsatisfactory results (41% 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% F1 – score. Moreover, for Extracellular, WegoLoc slightly (3%) outperforms PSL-Recommender.

However, our proposed method is trained and evaluated on the same data set as Hum-mPLOC 3.0 while other methods are trained with older data sets. Therefore, it is understandable that our method has better results on this dataset with respect to the other methods. At the same time, PSL-Recommender outperformed Hum-mPLOC 3.0 which is employed SVM⁵⁴. While both methods are using almost the same data sets, this superiority demonstrates the advantage of using a customized machine learning solution rather than using a popular machine learning method for this problem.

To evaluate the performance of PSL-Recommender on other species, we have employed previously introduced datasets that include proteins from animals and eukaryotes. Then we compared the results with five state-of-the-art methods including YLoc⁶, MultiLoc2³⁶, BaCelLo⁵⁵, PMLPR⁸, and Hum-mPloc 3.0². The results are depicted in Table 4.

Location	Yloc+			iLoc-Human			WegoLoc			mLASSO-Hum			Hum-mPloc3.0			PSL-Recommend		
	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	<u>0.63</u>	0.94	0.69	0.80
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	<u>0.74</u>	0.81	0.78	0.79
Cytoskeleton	-	-	-	0	0	0	0.32	0.34	0.33	0.9	0.22	0.35	0.8	0.68	<u>0.74</u>	0.97	0.70	0.82
ER	0.71	0.12	0.21	0	0	0	0.73	0.2	0.31	0.74	0.49	<u>0.59</u>	0.83	0.37	0.51	0.91	0.72	0.80
Endosome	-	-	-	0	0	0	0.25	0.07	0.11	0.38	0.2	0.26	0.58	0.47	0.52	0.63	0.31	<u>0.41</u>
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	<u>0.68</u>
Golgi apparatus	0.1	0.05	0.07	0.6	0.3	0.4	0.6	0.15	0.24	0.72	0.65	<u>0.68</u>	0.69	0.45	0.55	0.86	0.59	0.70
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	<u>0.67</u>	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	<u>0.85</u>	0.78	0.75	0.76	0.93	0.86	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	<u>0.76</u>	0.75	0.71	0.73	0.83	0.91	0.87
Peroxisome	0.07	0.5	0.13	1	0.5	<u>0.67</u>	0.5	1	<u>0.67</u>	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	<u>0.57</u>	0.65	0.44	0.52	0.77	0.73	0.75
ACC	0.45			0.41			0.50			<u>0.65</u>			0.63			0.77		
F1-mean	0.34			0.32			0.44			0.56			<u>0.65</u>			0.77		

Table 3. Comparison of PSL-Recommend on Human proteins dataset(Hum-mPloc 3.0) 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	<u>0.86/0.84</u>	<u>0.64/0.59</u>	<u>0.87/0.84</u>
PSL-Recommend	0.93/0.92	0.91/0.88	0.88/0.85

Table 4. Comparison of PSL-Recommend ACC/F1 – mean on other species proteins datasets with state-of-the-art methods.

As seen in Table 4, PSL-Recommend demonstrated excellent performance by reaching almost 90 in both F1 – mean and ACC. This could be for development of GO databases in recent years which boost our performance. However, in the next section, we demonstrated that PSL-Recommend could perform great only by using the PSSM features over these datasets, outperforming some states-of-the-art methods.

In order to compare PSL-Recommend performance with some other prominent works like SherLoc2⁵, WoLF PSORT⁴, and Euk-mPloc¹⁷, we have investigated AVG and ACC2 of PSL-Recommend results over two data set BaCelLo IDS animals and Höglund IDS animals and compared them with reported results in SherLoc2 paper which is represented in Table 5. As demonstrated in Table 5 PSL-Recommend has highlighted the best performance with respect to AVG and ACC2. PSL-Recommend shown great improvement by 16% in AVG and 25% in ACC2 over BaCelLo IDS dataset and also outperforming results over Höglund IDS dataset with 47% and 40% improvement over AVG and ACC2 respectively.

Method	BaCelLo	Höglund
PSL-Recommend	0.92/0.96	0.86/0.97
SherLoc2	<u>0.76/0.71</u>	<u>0.39/0.54</u>
MultiLoc2	0.75/0.68	0.38/0.57
WoLF PSORT	0.69/ <u>0.71</u>	0.24/0.56
Euk-mPloc	0.48/0.58	0.18/0.22

Table 5. Comparisons of PSL-Recommend, SherLoc2, MultiLoc2, WoLF PSORT, and Euk-mPloc performance with respect to AVG/ACC2.

Deeploc²⁹ is almost the best single label PSL predictor, which employed only sequence derived features to predict subcellular location. On the other hand, PSL-Recommend considered PSSM, three types of GO terms semantic similarity, and STRING features. In order to have a fair comparison between these two methods, we excluded cellular component GO terms and STRING features from PSL-Recommend in both training and testing processes. The threshold is removed from the model and subcellular location with the highest probability is suggested for each protein. The result was satisfying, PSL-Recommend

achieved 0.7818 and 0.7213 in ACC3 and Gorodkin, respectively, which is slightly better than DeepLoc performance with 0.7511 and 0.6988 in ACC3 and Gorodkin, respectively.

It also worth mentioning that, for PSL prediction problem, to the best of our knowledge, PMLPR⁸ is the only recommender system based method that employs the well-known network-based inference(NBI)⁵⁶ approach. As seen in Table 4, PSL- Recommender outperforms PMLPR by 50% and 18% in F1 – mean, and also 27% and 16% in ACC on Höglund and DBMloc datasets, respectively.

Impact of each similarity matrix

The proteins similarity matrix is used for neighbourhood 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 6. shows the result of each combination on all datasets. As can be seen in Table 6., those combinations excluding 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 Höglund 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 Höglund datasets contain proteins from multiple species we were unable to use String interaction scores in these datasets.

Features	BaCelLo	Höglund	DBMLoc	Hum-mPloc 3.0
PSSM	0.69/0.53	0.63/0.26	0.81/0.77	0.33/0.17
STRING	-	-	-	0.44/0.40
GO	0.93/0.91	0.90/0.87	0.86/0.84	0.76/0.75
PSSM+STRING	-	-	-	0.46/0.37
GO +PSSM	0.93/0.92	0.91/0.88	0.88/0.85	0.77/0.76
GO+STRING	-	-	-	0.78/0.77
All	0.93/0.92	0.91/0.88	0.88/0.85	0.77/0.77

Table 6. PSL-Recommender ACC/F1 – mean comparisons by using different features.

Stability of the PSL-Recommender

Choosing appropriate hyperparameters plays a vital role in the performance of a model. As mentioned in learning hyperparameters section, the models for all of the datasets constructed by same set of hyperparameters based on HumB dataset (Table 2). The results of using these hyperparameters represented on Table 3, Table 4, and Table 5 for all datasets.

In order to investigate the stability of the models, for each datasets, the hyperparametr are selected according to their training set by applying 5-fold cross validation with similar procedure which is explained in learning hyperparameters section. By considering different hyperparameters, F1 – mean reached to 0.92, 0.90, and 0.89 and ACC get to 0.94, 0.92, and 0.89 for BaCelLo IDS, Höglund IDS, and DBMloc, respectively.

For each datasets, by applying selected hyperparameters with respect to their training set, the F1 – mean and ACC can be increased only by 2 percent. It can be concluded that, despite large number of hyperparameters, PSL-Recommender is a stable method for PSL prediction.

Discussion

In the absence of efficient experimental methods, computational tools could 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 has outstanding performance. While using newer GO datasets could cause an illusion about our method's performance, we demonstrated that PSL-Recommender could perform great on single-label datasets only using sequence-driven features.

However, we believe that the performance of PSL-Recommender can be improved further by employing a better approach for searching the parameter space.

Code availability

Datasets and the codes of the algorithms are available in our github repository. <https://github.com/RJamali/PSL-Recommender>.

References

1. Nair, R. & Rost, B. Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins: Struct. Funct. Bioinforma.* **53**, 917–930 (2003).
2. Zhou, H., Yang, Y. & Shen, H.-B. 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**, 843–853 (2016).
3. Höglund, A., Dönnies, P., Blum, T., Adolph, H.-W. & Kohlbacher, O. Multiloc: prediction of protein subcellular localization using n-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* **22**, 1158–1165 (2006).
4. Horton, P. *et al.* Wolf psort: protein localization predictor. *Nucleic acids research* **35**, W585–W587 (2007).
5. Briesemeister, S. *et al.* Sherloc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. proteome research* **8**, 5363–5366 (2009).
6. Briesemeister, S., Rahnenführer, J. & Kohlbacher, O. Going from where to why—interpretable prediction of protein subcellular localization. *Bioinformatics* **26**, 1232–1238 (2010).
7. Cheng, X., Xiao, X. & Chou, K.-C. ploc-mhum: predict subcellular localization of multi-location human proteins via general pseAAC to winnow out the crucial go information. *Bioinformatics* **34**, 1448–1456 (2017).
8. Mehrabad, E. M., Hassanzadeh, R. & Eslahchi, C. Pmlpr: A novel method for predicting subcellular localization based on recommender systems. *Sci. reports* **8**, 12006 (2018).
9. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics* **18**, 298–305 (2002).
10. Chou, K.-C. & Shen, H.-B. Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms. *Nat. protocols* **3**, 153 (2008).
11. Xie, D., Li, A., Wang, M., Fan, Z. & Feng, H. Locsvmpsi: a web server for subcellular localization of eukaryotic proteins using svm and profile of psi-blast. *Nucleic acids research* **33**, W105–W110 (2005).
12. Chou, K.-C. & Cai, Y.-D. Prediction and classification of protein subcellular location—sequence-order effect and pseudo amino acid composition. *J. cellular biochemistry* **90**, 1250–1260 (2003).
13. King, B. R. & Guda, C. ngloc: an n-gram-based bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome biology* **8**, R68 (2007).
14. Nair, R. & Rost, B. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* **18**, S78–S86 (2002).
15. Shatkay, H. *et al.* Sherloc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* **23**, 1410–1417 (2007).
16. Brady, S. & Shatkay, H. Epiloc: a (working) text-based system for predicting protein subcellular location. In *Pacific Symposium on Biocomputing*, 604–615 (2008).
17. Chou, K.-C. & Shen, H.-B. Euk-mploc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* **6**, 1728–1734 (2007).
18. Lee, K. *et al.* Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research* **36**, e136–e136 (2008).
19. Huang, W.-L., Tung, C.-W., Ho, S.-W., Hwang, S.-F. & Ho, S.-Y. Proloc-go: utilizing informative gene ontology terms for sequence-based prediction of protein subcellular localization. *BMC bioinformatics* **9**, 80 (2008).
20. Shin, C. J., Wong, S., Davis, M. J. & Ragan, M. A. Protein-protein interaction as a predictor of subcellular location. *BMC systems biology* **3**, 28 (2009).
21. Xiao, X., Wu, Z.-C. & Chou, K.-C. iloc-virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. *J. theoretical biology* **284**, 42–51 (2011).
22. He, J., Gu, H. & Liu, W. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PloS one* **7**, e37155 (2012).

23. Su, R., He, L., Liu, T., Liu, X. & Wei, L. Protein subcellular localization based on deep image features and criterion learning strategy. *Briefings Bioinforma.* **22**, bbaa313 (2021).
24. Simha, R. & Shatkay, H. Protein (multi-) location prediction: using location inter-dependencies in a probabilistic framework. *Algorithms for Mol. Biol.* **9**, 8 (2014).
25. Nakai, K. & Horton, P. Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends biochemical sciences* **24**, 34–35 (1999).
26. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *J. molecular biology* **300**, 1005–1016 (2000).
27. Sinha, S. On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* **22**, e454–e463 (2006).
28. Nakashima, H. & Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. molecular biology* **238**, 54–61 (1994).
29. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).
30. Shen, Y., Ding, Y., Tang, J., Zou, Q. & Guo, F. Critical evaluation of web-based prediction tools for human protein subcellular localization. *Briefings bioinformatics* **21**, 1628–1640 (2020).
31. Francesco, R., Lior, R. & Bracha, S. Introduction to recommender systems handbook, recommendersystems handbook (2011).
32. Liu, Y., Wu, M., Miao, C., Zhao, P. & Li, X.-L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS computational biology* **12**, e1004760 (2016).
33. Liu, H. *et al.* Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* **8**, 103975 (2017).
34. Zheng, X., Ding, H., Mamitsuka, H. & Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1025–1033 (ACM, 2013).
35. Luo, H. *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* **34**, 1904–1912 (2018).
36. Blum, T., Briesemeister, S. & Kohlbacher, O. Multiloc2: integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC bioinformatics* **10**, 274 (2009).
37. Mnih, A. & Salakhutdinov, R. R. Probabilistic matrix factorization. In *Advances in neural information processing systems*, 1257–1264 (2008).
38. Johnson, C. C. Logistic matrix factorization for implicit feedback data. *Adv. Neural Inf. Process. Syst.* **27** (2014).
39. Hao, M., Bryant, S. H. & Wang, Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci. reports* **7**, 40376 (2017).
40. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nat. genetics* **25**, 25 (2000).
41. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the ‘perceptron’ algorithm to distinguish translational initiation sites in e. coli. *Nucleic acids research* **10**, 2997–3011 (1982).
42. Szklarczyk, D. *et al.* String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* **43**, D447–D452 (2014).
43. Altschul, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389–3402 (1997).
44. Scott, M. S., Calafell, S. J., Thomas, D. Y. & Hallett, M. T. Refining protein subcellular localization. *PLoS computational biology* **1**, e66 (2005).
45. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic acids research* **38**, D5–D16 (2009).
46. Mazandu, G. K., Chimusa, E. R., Mbiyavanga, M. & Mulder, N. J. A-dago-fun: an adaptable gene ontology semantic similarity-based functional analysis tool. *Bioinformatics* **32**, 477–479 (2015).

47. Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011).
48. Casadio, R., Martelli, P. L. & Pierleoni, A. The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Briefings Funct. Genomics Proteomics* **7**, 63–73 (2008).
49. Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min. (IJDWM)* **3**, 1–13 (2007).
50. Gorodkin, J. Comparing two k-category assignments by a k-category correlation coefficient. *Comput. biology chemistry* **28**, 367–374 (2004).
51. Chou, K.-C., Wu, Z.-C. & Xiao, X. iloc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **8**, 629–641 (2012).
52. Chi, S.-M. & Nam, D. Wegoloc: accurate prediction of protein subcellular localization using weighted gene ontology terms. *Bioinformatics* **28**, 1028–1030 (2012).
53. Wan, S., Mak, M.-W. & Kung, S.-Y. mlasso-hum: A lasso-based interpretable human-protein subcellular localization predictor. *J. theoretical biology* **382**, 223–234 (2015).
54. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. their applications* **13**, 18–28 (1998).
55. Pierleoni, A., Martelli, P. L., Fariselli, P. & Casadio, R. Bacello: a balanced subcellular localization predictor. *Bioinformatics* **22**, e408–e416 (2006).
56. Zhou, T., Ren, J., Medo, M. & Zhang, Y.-C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **76**, 046115 (2007).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.[?].

For data citations of datasets uploaded to e.g. *figshare*, please use the `howpublished` option in the bib entry to specify the platform and the link, as in the `Hao:gidmaps:2014` example in the sample bibliography file.

Acknowledgements

This work is supported by Iran National Science Foundation.

Author contributions statement

All authors conceived the experiment(s), R.J. conducted the experiment(s), R.J. authors analysed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.E. .