bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

PSL-Recommender: Protein Subcellular Localization Prediction using Recommender System

Ruhollah Jamali[a], Changiz Eslahchi[b,*], Soheil Jahangiri-Tazehkand[a,c,*]

[a] *School of Biological Sciences, Institute for Research in Fundamental Sciences(IPM), Tehran, Iran.*
[b] *Department of Computer Science, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran*
[c] *Princess Margaret Cancer Centre, University Health Network, 101 College Street, Toronto, ON, M5G1L7, Canada*

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 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.

PSL-Recommender (Protein subcellular location recommender) is a method that employs neighborhood regularized logistic matrix factorization to build a recommender system for protein subcellular localization. The effectiveness of PSL-Recommender method is benchmarked on one human and three animals datasets. The results indicate that the PSL-Recommender significantly outperforms state-of-the-art methods, improving the previous best method up to 31% in F1 − mean, up to 28% in ACC, and up to 47% in AVG. The source of datasets and codes are available at: https://github.com/RJamali/PSL-Recommender

Keywords: Molecular biology, Recommender systems,, Proteins subcellular location

2010 MSC: 00-01, 99-00

*Corresponding authors

Email addresses: ch-eslahchi@sbu.ac.ir (Changiz Eslahchi), soheil.jahangiri@uhnresearch.ca (Soheil Jahangiri-Tazehkand)

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [1]. Recent advances in genomics and proteomics provide 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 manner [2]. 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, 5–23].

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 [24], WoLF PSORT [4] and TargetP [25] employ sequence sorting signals [9] while Cell-Ploc [10] and LOCSVMPSI [11] use position specific scoring matrix [26]. Additionally, amino/psudo-amino acid composition information [12, 27] is utilized by ngLOC [13]. 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 localization [14–16].

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 prob-

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lem, such as support vector machine [2, 3], K-nearest neighbors [21, 22], and Bayesian methods [6, 23].

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 a previous knowledge about relations within and between items and users [28]. As of late framework strategies, recommendation systems have been utilized to predict associations in challenging bioinformatics problems [29–32]. Well-known PSL prediction methods assign equal importance to all proteins information in both constructing model and prediction tasks [2, 5, 6, 33], but utilization prioritized information from similar proteins in model construction step is likely more meaningful. Additionally, due to 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. In order to considering 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 PSL prediction process.

PSL-Recommender employs logistic matrix factorization technique [34] 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 has shown promising results for problems such as music recommendation [35], drug-target interaction prediction [29, 36], and lncRNA-protein interaction prediction [30]. 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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significantly outperforms the results of 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 [37] similarities, PSSM [38] similarity and STRING [39] 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.

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 [40] 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 [8, 18, 41]. 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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

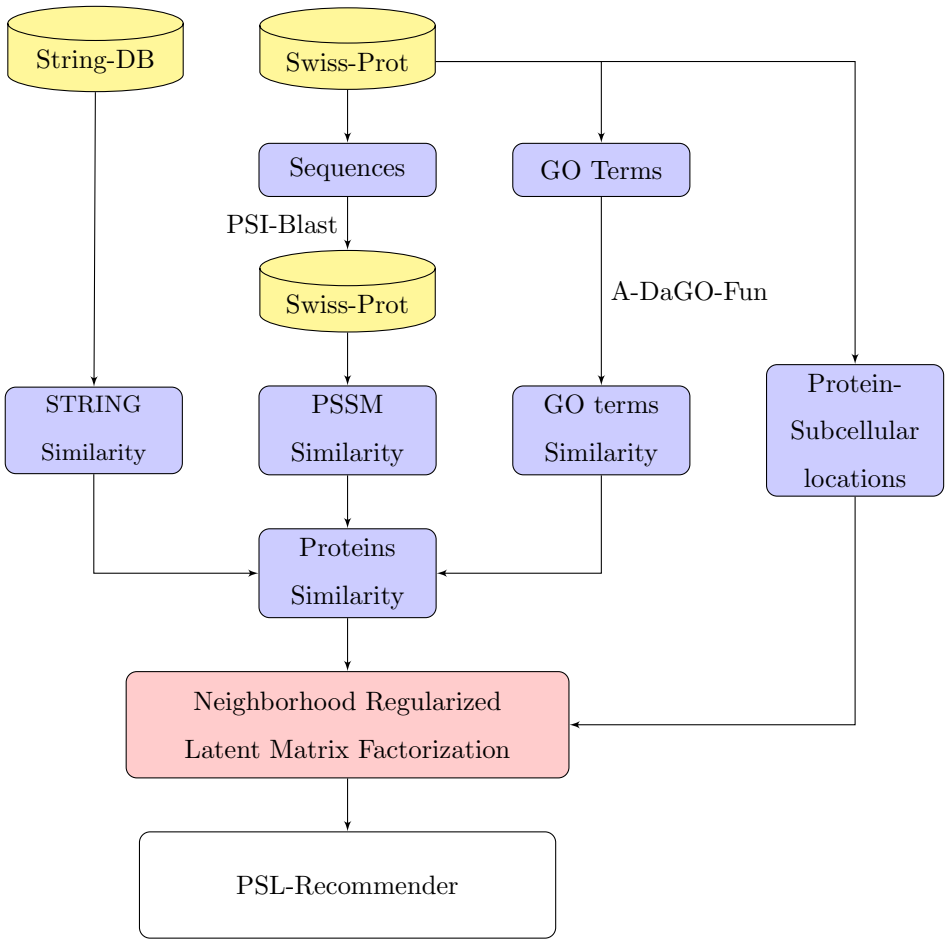

Figure 1: Construction pipeline of PSL-Recommender

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 [2, 42]. 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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Resnik GO terms semantic similarities [43]. 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 $S^p = \left[s^p_{i,k}\right]_{m\times m}$ 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:

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

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 $U \in \mathbb{R}^{\mathrm{m}\times \mathrm{d}}$, $V \in \mathbb{R}^{\mathrm{n}\times \mathrm{d}}$, $\beta^p \in \mathbb{R}^{\mathrm{m}\times 1}$ and $\beta^l \in \mathbb{R}^{\mathrm{n}\times 1}$ 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:

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [29]. 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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [44] algorithm to choose the gradient step size in each iteration adaptively. The

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

partial gradients of latent vectors and biases are given by:

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

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 [2, 5, 6]. These datasets are summarized in Table **??**. The Hum-mploc3.0, the BaCelLo IDS animals[45], and

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the Höglund IDS[33] 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 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 relese 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)

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

Table 1: Datasets summary

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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 [46] and used by other state-of-the-art methods for this problem. ACC is the average of $\mathrm{ACC_{x_i}}$ of all proteins in the test set, calculated for each protein as follows:

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Alongside, SherLoc2 [5] applied two other evaluation criterias, named ACC2 (ratio of correctly predicted proteins) and AVG (average fraction of called instances) which are defined as follow:

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

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

2.2.1. 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 $\text{F1} - \text{mean}$: HumB is devided into 5 equal subsets and PSL-Recommender is trained on union of 4 subsets and one other subset was hold for test the $\text{F1} - \text{mean}$. 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 $\text{F1} - \text{mean}$ 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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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 parameterof 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)

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 [2] human protein dataset to compare the performance of PSL-Recommender to six methods that were in-

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

troduced for protein localization in human. The methods include YLoc+ [6], iLoc-Hum [47], WegoLoc [48], mLASSO-Hum [49] 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.

As seen in Table 3, 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.

To show the performance of PSL-Recommender on other species we have em-

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	0.69	**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	0.74	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	0.74	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	0.59	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	0.41
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	0.86	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	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	0.93	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	0.76	0.75	0.71	0.73	0.83	0.91	**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	0.77	0.73	**0.75**
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			**0.77**	

Table 3: Comparison of PSL-Recommender on Human proteins dataset(Hum-mPloc 3.0) with other methods.

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ployed previously introduced datasets that include proteins from animals and eukaryotes. We then compared the results to five state-of-the-art methods including [2, 6, 8, 33, 50]. The results are depicted in Table 4.

	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	**0.93/0.92**	**0.91/0.88**	**0.88/0.85**

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

As seen in Table 4, PSL-Recommender outperforms all methods in all datasets by both F1 − mean and ACC. In Höglund IDS animals dataset, PSL-Recommender significantly outperforms the second best method by 27% and 29% in F1 − mean and ACC respectively. In BaCelLo IDS animals dataset, the improvement over the second best method is 7% in F1 − mean and 8% in ACC, while in DBMloc dataset, PSL-Recommender slightly improves the second best method by 1% in both F1 − mean and ACC.

In order to compare PSL-Recommender performance with some other prominent works like SherLoc2 [5], WoLF PSORT [4], and Euk-mPloc [17], we have investigated AVG and ACC2 of PSL-Recommender 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.

Table 6 is also demonstrate great performance of PSL-Recommender with re-

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spect to AVG and ACC2. PSL-Recommender 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.

It also worth mentioning that, for PSL prediction problem, to the best of our knowledge, PMLPR [8] is the only recommender system based method that employs the well-known network-based inference(NBI) [51] 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, repectively.

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

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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.

3.3. Stability of the PSL-Recommender

Choosing appropriate hyperparameters plays a vital role in the performance of a model. As mentioned in section 2.2.1, 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 hyperparametrs are selected according to their training set by applying 5-fold cross validation with similar procedure which is explained in section 2.2.1. By considering different hyperparameters, F1 − mean reached to 0.92, 0.90, and 0.89 and

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] H. Zhou, Y. Yang, H.-B. Shen, 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) (2016) 843–853.

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

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

[5] S. Briesemeister, T. Blum, S. Brady, Y. Lam, O. Kohlbacher, H. Shatkay, Sherloc2: a high-accuracy hybrid method for predicting subcellular localization of proteins, Journal of proteome research 8 (11) (2009) 5363–5366.

[6] S. Briesemeister, J. Rahnenführer, O. Kohlbacher, Going from where to whyinterpretable prediction of protein subcellular localization, Bioinformatics 26 (9) (2010) 1232–1238.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

[14] R. Nair, B. Rost, Inferring sub-cellular localization through automated lexical analysis, Bioinformatics 18 (suppl_1) (2002) S78–S86.

[15] H. Shatkay, A. Höglund, S. Brady, T. Blum, P. Dönnes, O. Kohlbacher, Sherloc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data, Bioinformatics 23 (11) (2007) 1410–1417.

[16] S. Brady, H. Shatkay, Epiloc: a (working) text-based system for predicting protein subcellular location, in: Pacific Symposium on Biocomputing, World Scientific, 2008, pp. 604–615.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

[21] X. Xiao, Z.-C. Wu, K.-C. Chou, 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) (2011) 42–51.

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

[31] X. Zheng, H. Ding, H. Mamitsuka, S. Zhu, 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, ACM, 2013, pp. 1025–1033.

[32] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, J. Wang, Computational drug repositioning using low-rank matrix approximation and randomized algorithms, Bioinformatics 34 (11) (2018) 1904–1912.

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

[34] A. Mnih, R. R. Salakhutdinov, Probabilistic matrix factorization, in: Advances in neural information processing systems, 2008, pp. 1257–1264.

[35] C. C. Johnson, Logistic matrix factorization for implicit feedback data, Advances in Neural Information Processing Systems 27.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[38] G. D. Stormo, T. D. Schneider, L. Gold, A. Ehrenfeucht, Use of the perceptronalgorithm to distinguish translational initiation sites in e. coli, Nucleic acids research 10 (9) (1982) 2997–3011.

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

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

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

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

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

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

[45] R. Casadio, P. L. Martelli, A. Pierleoni, The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation, Briefings in Functional Genomics and Proteomics 7 (1) (2008) 63–73.

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

bioRxiv preprint doi: https://doi.org/10.1101/462812; this version posted April 21, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

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

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