Liu et al. BMC Genomics 2018, 19(Suppl 10):883
https://doi.org/10.1186/s12864-018-5276-7

BMC Genomics

RESEARCH Open Access

A partially function-to-topic model for protein function prediction

Lin Liu[1], Lin Tang[2*], Mingjing Tang[3] and Wei Zhou[4*]

From 29th International Conference on Genome Informatics
Yunnan, China. 3-5 December 2018

Abstract

Background: Proteins are a kind of macromolecules and the main component of a cell, and thus it is the most essential and versatile material of life. The research of protein functions is of great significance in decoding the secret of life. In recent years, researchers have introduced multi-label supervised topic model such as Labeled Latent Dirichlet Allocation (Labeled-LDA) into protein function prediction, which can obtain more accurate and explanatory prediction. However, the topic-label corresponding way of Labeled-LDA is associating each label (GO term) with a corresponding topic directly, which makes the latent topics to be completely degenerated, and ignores the differences between labels and latent topics.

Result: To achieve more accurate probabilistic modeling of function label, we propose a Partially Function-to-Topic Prediction (PFTP) model for introducing the local topics subset corresponding to each function label. Meanwhile, PFTP not only supports latent topics subset within a given function label but also a background topic corresponding to a 'fake' function label, which represents common semantic of protein function. Related definitions and the topic modeling process of PFTP are described in this paper. In a 5-fold cross validation experiment on yeast and human datasets, PFTP significantly outperforms five widely adopted methods for protein function prediction. Meanwhile, the impact of model parameters on prediction performance and the latent topics discovered by PFTP are also discussed in this paper.

Conclusion: All of the experimental results provide evidence that PFTP is effective and have potential value for predicting protein function. Based on its ability of discovering more-refined latent sub-structure of function label, we can anticipate that PFTP is a potential method to reveal a deeper biological explanation for protein functions.

Keywords: Multi-label classification, Topic model, Protein function, Probability distribution

Background

Proteins are the main component of a cell, which explain the basic activity of cellular life. The research of protein functions is of great significance in elucidating the phenomena of life [1]. Although there have been amount of protein sequences in biological database in recent years [2, 3], a small percentage of these proteins have experimental function annotations because of the high cost of biochemical experiment. In comparison with biochemical experiment, computational methods predict the functional annotations of proteins by using known information, such as sequence, structure, and functional behavior, which reduce time and effort, and have become important long-standing research works in post-genomic era [4].

The earlier computational approach for predicting protein function is to utilize the protein sequence or structure similarity to transfer functional information, such as BLAST. [5]With the rapid development of computational algorithms, an increasing types of algorithms have been introduced into the studies of predicting

* Correspondence: maitanweng2@163.com; zwei@ynu.edu.cn
[2]Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming 650500, Yunnan, China
[4]School of Software, Yunnan University, Kunming 650091, Yunnan, China
Full list of author information is available at the end of the article

protein function. At present, computational methods of protein function prediction can be classified as two types: classification-based approaches and graph-based approaches. In classification-based approaches, proteins are viewed as instances to be classified, and function annotations (such as Gene Ontology (GO) [6] terms) are regarded as labels. Each protein has a feature space composed by classification feature extracted from amino acid sequence, textual repositories, and so on. Based on these annotated proteins and their attribute features, we can train the classifier on training dataset and then predict function labels for unannotated proteins. For graph-based approaches, the network structure information of proteins is used to compute the distance between proteins, and then the closely related proteins are considered to have similar functional annotations [7, 8].

In classification-based approaches, since each protein is annotated with several functions, various multi-label classifiers can be adopted. Yu et.al [9] proposed a multiple kernels (ProMK) method to process multiple heterogeneous protein data sources for predicting protein functions; Fodeh et.al [10] used the binary-relevance for different classifiers to automatically assign molecular functions to genes; a new ant colony optimization algorithm is proposed in reference [11], which has applied to protein function dataset; Wang et.al [12] applied a new multi-label linear discriminant analysis approach to address protein function prediction problem; Liu et.al [4] introduced a multi-label supervised topic model called Labeled-LDA into protein function prediction, whose experimental results on yeast and human datasets demonstrated the effectiveness of Labeled-LDA on protein function prediction. This research is the first effort to apply a multi-label supervised topic model to protein function prediction. Besides, Pinoli et.al [13–15] applied two standard topic models, including latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA) [16, 17], to predict GO terms of proteins on the basic of available GO annotations.

In the topic modeling process of reference [4], each protein is viewed as a mixture of 'topics', where each 'topic' is also viewed as the mixture of amino acid blocks. In comparison with discriminative model, such as support vector machine (SVM), a multi-label supervised topic model can transform the word-level statistics of each document to its label-level distribution, and model all labels simultaneously rather than treating each label independently. Specially, topic model can provide the function label probability distribution over proteins as an output, and each function label is explained as a probability distribution over amino acid blocks. Nonetheless, in the study of Liu et.al [4], Labeled-LDA associates each label (GO term) with a corresponding topic directly, which makes the latent topics to be completely degenerated, and ignores the differences between labels and latent topics. Therefore, Labeled-LDA isn't able to discover the topic that represents common semantic of protein functions. For interpretable text mining, Ramage et.al [18] proposed a partially labeled LDA (PLDA), which associates each label with a topic subset partitioned from global topics set. PLDA overcame the shortfalls of Labeled-LDA, and improved the precision of text classification in experimental research.

Inspired by the application of multi-label topic model in protein function prediction and PLDA model, we introduce a Partially Function-to-Topic Prediction model (called PFTP). Firstly, we describe the related definitions by contrasting text data and protein function data. Then the topic modeling process of PFTP is described in detail, including the generative process and parameter estimation of PFTP. In a 5-fold cross validation experiment on predicting protein function, PFTP significantly outperforms five algorithms compared. All of the experimental results provide evidence that PFTP is effective and have potential value for predicting protein function.

Methods

Related definitions and notations

To better understand related objects of topic model, the corresponding relationship between protein function prediction and multi-label classification of text is first depicted in Fig. 1.

Several topic modeling concepts of protein function data and text data are displayed in Fig. 1, one on the left and the other on the right. First of all, the text dataset is composed of several documents numbered D1 to Dn, and the protein function dataset is composed of several protein sequences numbered P1 to Pn. Obviously, words are the main component of document, such as word 'table' and 'database'. But for protein sequence, we consider a protein sequence to be a text string, which is defined on a fixed 20 amino acids alphabet (G,A,V,L,I,F,P,Y,S,C,M,N,Q,T,D,E,K,R,H,W). Then amino acid blocks are the main component of protein sequence, such as 'MS' and 'TS'. Besides, a protein annotated by GO terms is equivalent to a document labeled by tags, so each GO term or tag can be viewed as a label, such as 'GO0003673' and 'language'. According to above statements, there are three types of equivalence relations between protein function data and text data: protein sequence and document, amino acid block and word, GO term and document tag. In general, the GO term (document tag), protein sequence (document) and amino acid block (word) are observable data for dataset.

As the input for topic model, the bag of words (BoW) is constructed by computing the word-document matrix, where matrix element is obtained by counting the times of word in each document. As an instance,

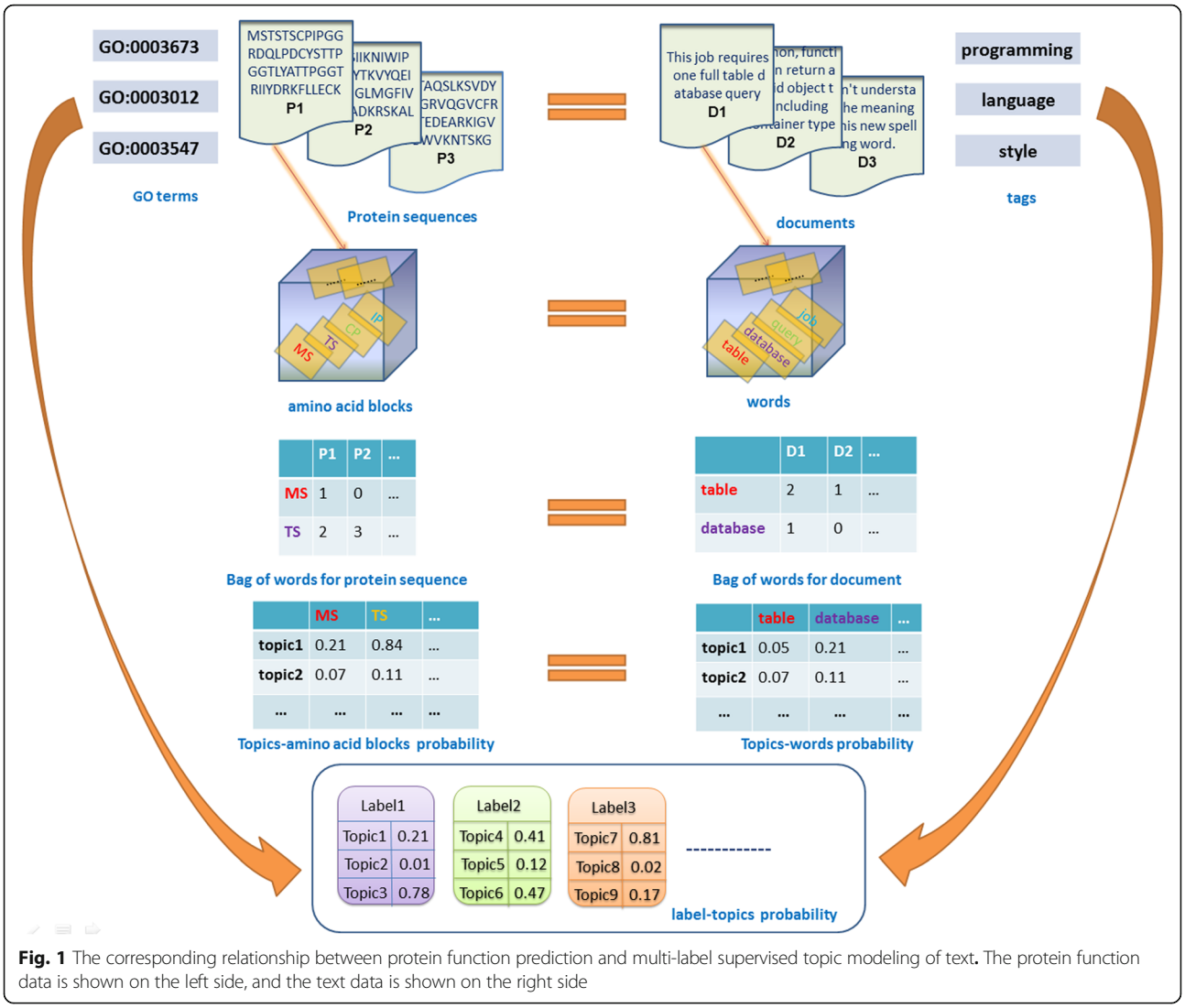

Fig. 1 The corresponding relationship between protein function prediction and multi-label supervised topic modeling of text. The protein function data is shown on the left side, and the text data is shown on the right side

the word 'table' appears two times in document D1. Likewise for protein function data, an amino acid block - protein sequence matrix is computed for the construction of protein BoW. As an example, the amino acid block 'MS' appears one times in protein P1. Besides, the fixed amino acid blocks set or words set is also called 'vocabulary'.

For topic model, a 'topic' is viewed as a probability distribution over a fixed vocabulary. Taking the text data as an example, the probabilities of word 'table' over 'topic 1' are 0.05. For the protein data, the probabilities of amino acid block 'MS' over 'topics 1' are 0.21. Obviously, topics are latent and needed to be inferred by topic modeling. Finally, in order to establish the connection between labels and topics, the latent topics discovered by our PFTP are divided into several non-overlapping subsets, each of which associates a label. As can be seen in Fig. 1, we split whole topic set into several groups: 'label1' connects with 'topic1' to 'topic3'; 'lable2' connects with 'topic 4' to 'topic 5', and so on. It is worth noting that our PFTP define a special type of topics as background topics. The background topics are divided from whole latent topics set, and don't associate any observable label, which express the common sematic of documents. For instance, the background topic on text dataset may be some topics with a high probability on several universal words, such as 'text', 'other' and so on. To formalize the above description, the related notations are given below.

Suppose there are D proteins in the protein set which compose the protein space $\mathbf{D} = \{1, ..., D\}$, and the vocabulary of amino acid blocks is in a space of $\mathbf{W} = \{1, ..., W\}$, then W is the size of vocabulary. The topic space including Ktopics is represented by $\mathbf{K} = \{1, ..., K\}$, which is shared by whole protein set. Therefore, $\mathbf{K}$ is

also called global topic space. The protein function label space is expressed as $\mathbf{L} = \{1, \ldots, L\}$.

In PFTP model, the global topic space $\mathbf{K}$ is divided into L groups without overlap, and each group corresponds to a subspace of topic $\mathbf{K}_l$. Besides, there is a 'background subspace of topics' $\mathbf{K}_B$.

$$\begin{aligned} &\mathbf{K} = (\cup_{l\in\mathbf{L}_d}\mathbf{K}_l)\cup\mathbf{K}_B, \quad \mathbf{K}_l, \mathbf{K}_B\subset\mathbf{K}, \\ &\mathbf{K}_l\neq\varnothing \quad (l\in\mathbf{L}), \quad \mathbf{K}_B\neq\varnothing, \\ &\forall\mathbf{K}_i, \mathbf{K}_j\subset\mathbf{K}, \quad i, j\in\mathbf{L}, \quad i\neq j \quad \Rightarrow \\ &\mathbf{K}_i\cap\mathbf{K}_j = \varnothing, \quad \mathbf{K}_i\cap\mathbf{K}_B = \varnothing \end{aligned}$$

Then, each of labels is assigned a subspace of topic $\mathbf{K}_l$, the background topic subspace $\mathbf{K}_B$ associates a background label l_B.In this case, the label space is expanded to $L+1$ dimensions and expressed as $\mathbf{L}'$. Similar to topic modeling of text in Labeled-LDA, each of topics can be represented as a multinomial distribution of parameter $\boldsymbol{\theta}_k = \{\theta_{kw}\}_{w=1}^{W}$ (the equivalent of the topic-word matrix in Fig. 1) on the vocabulary $\mathbf{W}$, and $\boldsymbol{\theta}_k$ obeys a Dirichlet prior distribution of hyper parameter $\boldsymbol{\lambda} = \{\lambda_w\}_{w\in\mathbf{W}}$. But what is different about our PFTP is that each of labels l is represented as a multinomial distribution of parameter $\boldsymbol{\pi}_l = \{\pi_{lk}\}_{k\in\mathbf{K}_l}$ (the equivalent of the label-topics probability in Fig. 1) on its topic subspace, where π_{lk} is the probabilities of topic k among topic subspace $\mathbf{K}_l$ corresponding to label l. Suppose $\boldsymbol{\pi}_l$ obeys a Dirichlet prior distribution of hyper parameter $\boldsymbol{\alpha}$.

$$\boldsymbol{\pi}_l \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \{\alpha_k\}_{k\in\mathbf{K}}, \quad |\,\boldsymbol{\alpha}\,| = |\mathbf{K}_l| = K_l \tag{1}$$

We utilize a binary vector $\boldsymbol{\Lambda}_d$ to map global label space $\mathbf{L}'$ to $\mathbf{L}_d$:

$$\begin{aligned} &\mathbf{L}_d = \{l\Lambda_{dl}\}_{l=1}^{L+1} = \mathbf{L}\boldsymbol{\Lambda}_d \\ &\boldsymbol{\Lambda}_d = \{\Lambda_{dl}\}_{l\in\mathbf{L}'} = \{\Lambda_{d1}, \Lambda_{d2}, \ldots, \Lambda_{dL}, 1\}, \\ &\Lambda_{dl} = \begin{cases} 1, & l\in\mathbf{L}_d \\ 0, & l\notin\mathbf{L}_d \end{cases} \end{aligned} \tag{2}$$

$\Lambda_{d,\,L+1} = 1$ illustrates that latent background label l_B is assigned to each protein d. Then, the probabilities of $L_d = |\mathbf{L}_d|$ labels of protein d is represented by a weight of protein-label $\boldsymbol{\psi}_d = \{\psi_{dl}\}_{l\in\mathbf{L}_d} = \{\psi_{dl}\Lambda_{dl}\}_{l\in\mathbf{L}'}$, and $\boldsymbol{\psi}_d$ obeys a Dirichlet prior distribution of hyper parameter $\boldsymbol{\beta}_d$ constrained by $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}_d$:

$$\begin{aligned} \boldsymbol{\beta}_d &= \{\beta_l\}_{l\in\mathbf{L}_d} = \{\beta_l\Lambda_{dl}\}_{l\in\mathbf{L}'} = \boldsymbol{\beta}\boldsymbol{\Lambda}_d, \quad \boldsymbol{\beta} \\ &= \{\beta_l\}_{l\in\mathbf{L}'} \end{aligned} \tag{3}$$

In this paper, the shared parameters of whole protein sets is called global parameter in this paper, and the parameter facing one protein is called local parameter.

The topic modeling process of PFTP

Based on above expression, the process of PFTP topic modeling is divided into three steps: BoW construction, the description of model (the generative process or graphic model) and parameter estimation (model training and predicting).These steps are depicted in Fig. 2.

As shown in Fig. 2, PFTP model takes BoW as input. As we construct BoW of protein in exactly the same way as reference [4], this step will not repeat in this paper. There are two ways to describe our topic model, including the generative process and the graphic model. After identifying the model structure, the joint distribution of whole model is obtained. Based on this joint distribution, we can learn and infer unknown parameters of our model, which are also the output of PFTP. In fact, unknown parameters represent several matrixes. For instance, $\boldsymbol{\theta}_k = \{\theta_{kw}\}_{w=1}^{W}$ represents the topic-word matrix in Fig. 2, and $\boldsymbol{\pi}_l = \{\pi_{lk}\}_{k\in\mathbf{K}_l}$ represents the label-topics matrix in Fig. 2.

The second and third steps are discussed in next sections. It is worth noting that the third step includes two sub-steps for realizing function prediction: model training and predicting. Both of these two sub-steps need adopt learning and inference algorithm to estimate parameters of model, and are described with more detail as follows.

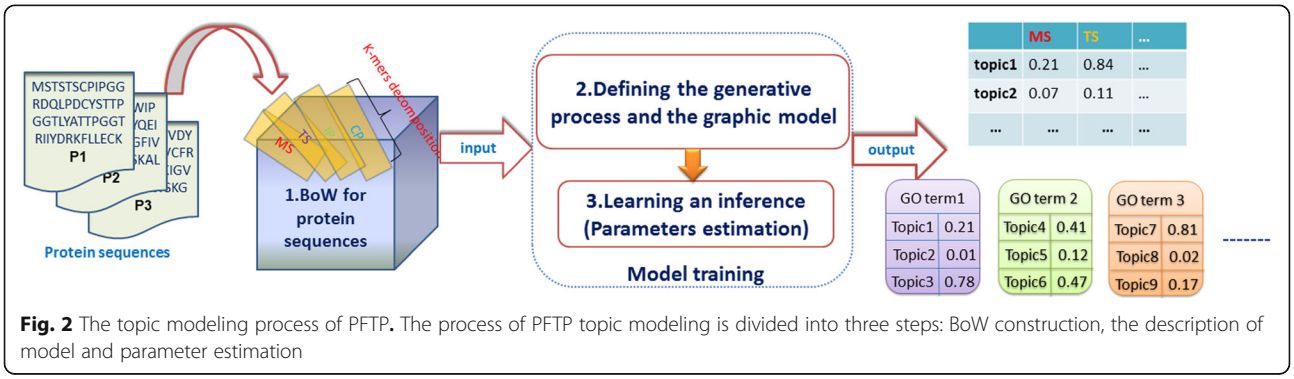

Fig. 2 The topic modeling process of PFTP**.** The process of PFTP topic modeling is divided into three steps: BoW construction, the description of model and parameter estimation

The process of model training

PFTP takes a training protein set with known function as an input of training model. The unknown parameter includes$\boldsymbol{\pi}_l$, $\boldsymbol{\theta}_k$ and $\boldsymbol{\psi}_d$. The local hidden variables include the label number and topic number of each word sample. The unknown parameter and local hidden variables can be estimated by inferring algorithm in model training.

The process of model predicting

For unannotated proteins, based on the estimated parameters and local hidden variables, unknown local parameter $\boldsymbol{\psi}_d$ and hidden variables are updating by constraining the global parameter $\boldsymbol{\pi}_l$ and $\boldsymbol{\theta}_k$. Then, the label probabilities over protein are obtained.

The description of PFTP model

According to the above definitions, the whole word sample x is composed by protein set, where $x_d = \{\mathbf{x}_{dn}\}_{n=1}^{N_d}$. It illustrates that there are N_d word samples in protein d, $\mathbf{x}_{dn}$ represents one word sample. At this point, word sample $\mathbf{x}_{dn}$ not only associates a word number $\mathbf{w}_{dn}$($\mathbf{w}_{dn}\in\mathbf{W}$), but also is assigned a label number $\mathbf{l}_{dn}$($\mathbf{l}_{dn}\in\mathbf{L}$) and a topic number$\mathbf{z}_{dn}$($\mathbf{z}_{dn}\in\mathbf{K}$).

The generative process of word sample can be described as follows. The corresponding graphical model is shown in Fig. 3.

1. For each global label $l\in\mathbf{L}' = \{1, \ldots, L, L+1\}$

Sample multinomial parameter vector $\boldsymbol{\pi}_l$ from K_l dimensions Dirichlet distribution:

$$\boldsymbol{\pi}_l = \{\pi_{lk}\}_{k\in\mathbf{K}} \sim \mathrm{Dir}(\boldsymbol{\alpha}),\ \ \boldsymbol{\alpha} = \{\alpha_k\}_{k\in\mathbf{K}} \tag{4}$$

2. For each global topic $k\in\mathbf{K} = \{1, \ldots, K\}$

Sample multinomial parameter vector $\boldsymbol{\theta}_k$ from W dimensions Dirichlet distribution:

$$\boldsymbol{\theta}_k = \{\theta_{kw}\}_{w\in\mathbf{W}} \sim \mathrm{Dir}(\boldsymbol{\lambda}),\ \ \boldsymbol{\lambda} = \{\lambda_w\}_{w\in\mathbf{W}} \tag{5}$$

3. For each local protein $d\in\mathbf{D} = \{1, \ldots, D\}$
   (a) Sample label weight vector of protein d from L_d dimensions Dirichlet distribution:

$$\begin{aligned}\boldsymbol{\psi}_d &= \{\psi_{dl}\}_{l\in\mathbf{L}_d} \sim \mathrm{Dir}(\boldsymbol{\beta}_d),\ \ \boldsymbol{\beta}_d = \{\beta_l\}_{l\in\mathbf{L}_d} \\ &= \{\beta_l \Lambda_{dl}\}_{l\in\mathbf{L}'} = \boldsymbol{\beta}\boldsymbol{\Lambda}_d\end{aligned} \tag{6}$$

where:

$$\begin{aligned}\boldsymbol{\beta} &= \{\beta_l\}_{l\in\mathbf{L}'}, \boldsymbol{\Lambda}_d = \{\Lambda_{dl}\}_{l\in\mathbf{L}'}, \Lambda_{dl} \\ &= \begin{cases} 1, & l\in\mathbf{L}_d \\ 0, & l\notin\mathbf{L}_d \end{cases}, \Lambda_{d,L+1} \equiv 1\end{aligned} \tag{7}$$

   (b) For each word sample $\mathbf{x}_{dn}$,
      i. Sample label number $\mathbf{l}_{dn}$ of $\mathbf{x}_{dn}$ from L_d dimensions multinomial distribution of parameter $\boldsymbol{\psi}_d$:

$$\mathbf{l}_{dn} \sim \boldsymbol{\psi}_d \ \text{ or } \ L_d = \{\mathbf{l}_{dn}\}_{n=1}^{N_d} \sim \mathrm{Mul}(\boldsymbol{\psi}_d, N_d) \tag{8}$$

      ii. Sample topic number $\mathbf{z}_{dn}$ of $\mathbf{x}_{dn}$ from K dimensions multinomial distribution of parameter$\boldsymbol{\pi}_{\mathbf{l}_{dn}}$:

$$\mathbf{z}_{dn} \sim \boldsymbol{\pi}_{\mathbf{l}_{dn}} \ \text{ or } \ \mathrm{Z}_d = \{\mathbf{z}_{dn}\}_{n=1}^{N_d} \sim \mathrm{Mul}(\boldsymbol{\pi}_{d\,\mathbf{l}_{dn}}, N_d) \tag{9}$$

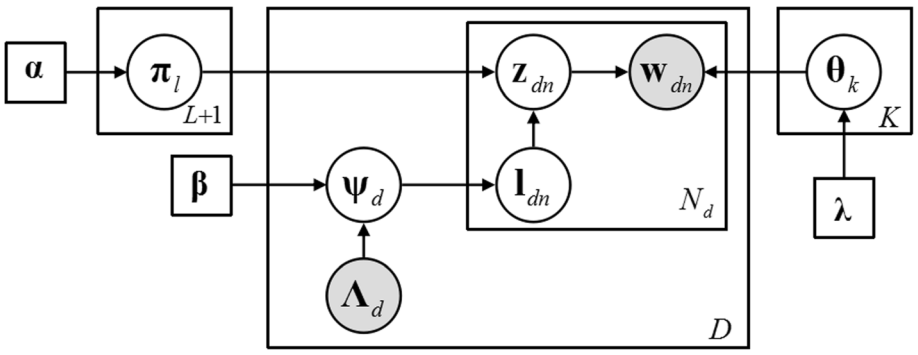

Fig. 3 The graphic model of PTPF. Box indicates repeated contents, and the number in the bottom right corner is the times of repetition

iii. Sample word number $\mathbf{w}_{dn}$ of $\mathbf{x}_{dn}$ from W dimensions multinomial distribution of parameter $\boldsymbol{\theta}_{\mathbf{z}_{dn}}$:

$$\mathbf{w}_{dn}\sim\boldsymbol{\theta}_{\mathbf{z}_{dn}} \text{ or } W_d=\{\mathbf{w}_{dn}\}_{n=1}^{N_d}\sim\mathrm{Mul}(\boldsymbol{\theta}_{Z_d},N_{d\mathbf{z}_{dn}}) \quad (10)$$

Parameter estimation

In PFTP model, the unknown parameters to be estimated are the global label multinomial parameters $\boldsymbol{\pi}=\{\boldsymbol{\pi}_l\}_{l\in\mathbf{L}'}=\{\pi_{lk}\}_{l\in\mathbf{L}',k\in\mathbf{K}_l}$, the global topic multinomial parameters $\boldsymbol{\theta}=\{\boldsymbol{\theta}_k\}_{k\in\mathbf{K}}=\{\theta_{kw}\}_{k\in\mathbf{K},w\in\mathbf{W}}$ and the local document label weight $\boldsymbol{\psi}_d=\{\psi_{dl}\}_{l\in\mathbf{L}_d}$; the local hidden variables are document label $L_d=\{\mathbf{l}_{dn}\}_{n=1}^{N_d}$ and topic $Z_d=\{\mathbf{z}_{dn}\}_{n=1}^{N_d}$; the known information are the observed label vector $\boldsymbol{\Lambda}_d$, word samples $W_d=\{\mathbf{w}_{dn}\}_{n=1}^{N_d}$ and their joint distribution. As shown in Eq. (11):

$$\begin{aligned}&p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\psi},L,Z,W|\boldsymbol{\Lambda},\boldsymbol{\alpha},\boldsymbol{\lambda},\boldsymbol{\beta})\\&=p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{\theta}|\boldsymbol{\lambda})\prod_{d\in\mathbf{D}}p(\boldsymbol{\psi}_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta}_d)\\&\quad p(L_d|\boldsymbol{\psi}_d)p(Z_d|L_d,\boldsymbol{\pi})p(W_d|Z_d,\boldsymbol{\theta})\\&=\prod_{l\in\mathbf{L}'}p(\boldsymbol{\pi}_l|\boldsymbol{\alpha})\prod_{k\in\mathbf{K}}p(\boldsymbol{\theta}_k|\boldsymbol{\lambda})\prod_{d\in\mathbf{D}}p(\boldsymbol{\psi}_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta}_d)\\&\quad\prod_{n=1}^{N_d}p(\mathbf{l}_{dn}|\boldsymbol{\psi}_d)p(\mathbf{z}_{dn}|\boldsymbol{\pi}_{\mathbf{l}_{dn}})p(\mathbf{w}_{dn}|\boldsymbol{\theta}_{\mathbf{z}_{dn}})\end{aligned} \quad (11)$$

Based on the joint distribution, several parameter estimations can be obtained, including $p(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\psi},L,Z|\,W,\boldsymbol{\Lambda},\boldsymbol{\alpha},\boldsymbol{\lambda},\boldsymbol{\beta})$, the posterior distribution of unknown model parameters and hidden variables. In this paper, we use the Collapsed Gibbs sampling (CGS) to train a PFTP model. By marginalizing the model parameters $(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\psi})$ from the joint distribution (11), the collapsed joint distribution of (L,Z,W) is obtained. The collapsed inference is as follows.

In the joint distribution Eq. (11), function label weight $\boldsymbol{\psi}_d$ only appears in $p(\boldsymbol{\psi}_d|\,\boldsymbol{\Lambda}_d,\boldsymbol{\beta}_d)$ and $p(L_d|\,\boldsymbol{\psi}_d)$:

$$\begin{aligned}p(\boldsymbol{\psi}_d,L_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta}_d)&=p(\boldsymbol{\psi}_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta}_d)p(L_d|\boldsymbol{\psi}_d)\\&=\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\beta_l\Lambda_{dl}\right)}\prod_{l\in\mathbf{L}'}\left(\psi_{dl}\Lambda_{dl}\right)^{\beta_l\Lambda_{dl}-1}\\&\quad\cdot\frac{\left(\sum_{l\in\mathbf{L}'}N_{dl}\Lambda_{dl}\right)!}{\prod_{l\in\mathbf{L}'}(N_{dl}\Lambda_{dl})!}\prod_{l\in\mathbf{L}'}\left(\psi_{dl}\Lambda_{dl}\right)^{N_{dl}\Lambda_{dl}}\\&=\mathrm{C}_1\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\beta_l\Lambda_{dl}\right)}\prod_{l\in\mathbf{L}'}\left(\psi_{dl}\Lambda_{dl}\right)^{\Lambda_{dl}\left(\beta_l+N_{dl}\right)-1}\end{aligned} \quad (12)$$

N_{dl} is the number of samples assigned to observed label $l\in\mathbf{L}_d$ of protein d; C_1 is the constant of multinomial distribution coefficient:

$$\mathrm{C}_1=\frac{\left(\sum_{l\in\mathbf{L}'}N_{dl}\Lambda_{dl}\right)!}{\prod_{l\in\mathbf{L}'}(N_{dl}\Lambda_{dl})!}=\frac{\left(\sum_{l\in\mathbf{L}'}N_{dl}\right)!}{\prod_{l\in\mathbf{L}_d}N_{dl}!} \quad (13)$$

Suppose$\hat{\beta}_{dl}=\Lambda_{dl}(\beta_l+N_{dl})$, $\hat{\psi}_{dl}=\psi_{dl}\Lambda_{dl}$. This parameter is eliminated by doing the integral of $\boldsymbol{\psi}_d$ in Eq. (11), the marginal distribution of local hidden variable L_d is shown in below:

$$\begin{aligned}p(L_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta})&=\int_{\Psi_d}p(\boldsymbol{\psi}_d,L_d|\boldsymbol{\Lambda}_d,\boldsymbol{\beta})\mathrm{d}\boldsymbol{\psi}_d\\&=\int_{\Psi_d}\mathrm{C}_1\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\beta_l\Lambda_{dl}\right)}\prod_{l\in\mathbf{L}'}\left(\psi_{dl}\Lambda_{dl}\right)^{\Lambda_{dl}\left(\beta_l+N_{dl}\right)-1}\mathrm{d}\boldsymbol{\psi}_d\\&=\mathrm{C}_1\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\beta_l\Lambda_{dl}\right)}\left(\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\hat{\beta}_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\hat{\beta}_{dl}\right)}\right)^{-1}\\&\quad\cdot\int_{\hat{\Psi}_d}\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\hat{\beta}_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\hat{\beta}_{dl}\right)}\prod_{l\in\mathbf{L}'}\hat{\psi}_{dl}^{\hat{\beta}_{dl}-1}\mathrm{d}\hat{\boldsymbol{\psi}}_d\\&\propto\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}\right)}{\Gamma\left(\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}+N_d\right)}\prod_{l\in\mathbf{L}'}\frac{\Gamma\left(\beta_l\Lambda_{dl}+N_{dl}\Lambda_{dl}\right)}{\Gamma\left(\beta_l\Lambda_{dl}\right)}\end{aligned} \quad (14)$$

$N_d=\sum_{l\in\mathbf{L}'}N_{dl}\Lambda_{dl}=\sum_{l\in\mathbf{L}_d}N_{dl}$ is the number of observed samples of protein d. The integral of Eq. (14) satisfies probabilistic completeness:

$$\begin{aligned}&\int_{\hat{\Psi}_d}\frac{\Gamma\left(\sum_{l\in\mathbf{L}'}\hat{\beta}_{dl}\right)}{\prod_{l\in\mathbf{L}'}\Gamma\left(\hat{\beta}_{dl}\right)}\prod_{l\in\mathbf{L}'}\hat{\psi}_{dl}^{\hat{\beta}_{dl}-1}\mathrm{d}\hat{\boldsymbol{\psi}}_d\\&=\int_{\hat{\Psi}_d}p\left(\hat{\boldsymbol{\psi}}_d|\hat{\boldsymbol{\beta}}_d\right)\mathrm{d}\hat{\boldsymbol{\psi}}_d=1\end{aligned} \quad (15)$$

Therefore, deducing from Eq. (14), the predictive probability distribution for the label-assignment $\mathbf{l}_{dn}=l$of sample $\mathbf{x}_{dn}$ is:

$$p\left(\mathbf{l}_{dn}=l|L_d^{(\backslash dn)},\boldsymbol{\Lambda}_d,\boldsymbol{\beta}\right)\propto\frac{\left(\beta_l+N_{dl}^{(\backslash dn)}\right)\Lambda_{dl}}{\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl}+N_d^{(dn)}} \quad (16)$$

$N_{dl}^{(\backslash dn)}$ is the number of samples that were assigned to label l and word w in addition to the current sample x_{dn}.

By the same way, in the joint distribution Eq. (11), global label parameter only appears in $p(\boldsymbol{\pi}|\,\boldsymbol{\alpha})$ and $p(Z_d|\,L_d,\boldsymbol{\pi})$.

$$\begin{aligned} p(\boldsymbol{\pi}, Z|L, \boldsymbol{\alpha}) &= p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(Z|L, \boldsymbol{\pi}) \\ &= p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{d\in\mathbf{D}} p(Z_d|L_d, \boldsymbol{\pi}) \\ &= \prod_{l\in\mathbf{L}'} p(\boldsymbol{\pi}_l|\boldsymbol{\alpha}) \prod_{d\in\mathbf{D}} \prod_{n=1}^{N_d} p(\mathbf{z}_{dn}|\mathbf{l}_{dn} = l, \boldsymbol{\pi}_l) \\ &= \prod_{l\in\mathbf{L}'} \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\alpha_k)} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{\alpha_k - 1} \frac{\left(\sum_{k\in\mathbf{K}} N_{lk}\right)!}{\prod_{k\in\mathbf{K}} N_{lk}!} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{N_{lk}} \\ &= \prod_{l\in\mathbf{L}'} \mathrm{C}_2 \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\alpha_k)} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{\alpha_k + N_{lk} - 1} \end{aligned} \tag{17}$$

N_{lk} represents the number of samples assigned to topic k of global label l; C_2 is the constant of multinomial distribution coefficient:

$$\mathrm{C}_2 = \frac{\left(\sum_{k\in\mathbf{K}} N_{lk}\right)!}{\prod_{k\in\mathbf{K}} N_{lk}!} \tag{18}$$

Suppose $\hat{\alpha}_k = \alpha_k + N_{lk}$. This parameter is eliminated by doing the integral of $\boldsymbol{\pi}$ in Eq. (17), the marginal distribution of local hidden variable Z is shown in below:

$$\begin{aligned} p(Z|L, \boldsymbol{\alpha}) &= \prod_{l\in\mathbf{L}'} \int_{\Pi_l} p(\boldsymbol{\pi}_l|\boldsymbol{\alpha}) \prod_{d\in\mathbf{D}} \prod_{n=1}^{N_d} p(\mathbf{z}_{dn}|\mathbf{l}_{dn} = l, \boldsymbol{\pi}_l) \mathrm{d}\boldsymbol{\pi}_l \\ &= \prod_{l\in\mathbf{L}'} \int_{\Pi_l} \mathrm{C}_2 \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\alpha_k)} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{\alpha_k + N_{lk} - 1} \mathrm{d}\boldsymbol{\pi}_l \\ &= \prod_{l\in\mathbf{L}'} \mathrm{C}_2 \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\alpha_k)} \left(\frac{\Gamma\left(\sum_{k\in\mathbf{K}} \hat{\alpha}_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\hat{\alpha}_{lk})}\right)^{-1} \\ &\cdot \int_{\Pi_l} \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \hat{\alpha}_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\hat{\alpha}_{lk})} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{\hat{\alpha}_{lk} - 1} \mathrm{d}\boldsymbol{\pi}_l \\ &\propto \prod_{l\in\mathbf{L}'} \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k + N_l\right)} \prod_{k\in\mathbf{K}} \frac{\Gamma(\alpha_k + N_{lk})}{\Gamma(\alpha_k)} \end{aligned} \tag{19}$$

$N_l = \sum_{k\in\mathbf{K}} N_{lk}$ is the number of observed samples assigned to global l in protein set. The integral of Eq. (19) satisfies probabilistic completeness:

$$\begin{aligned} &\int_{\Pi_l} \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \hat{\alpha}_k\right)}{\prod_{k\in\mathbf{K}} \Gamma(\hat{\alpha}_{lk})} \prod_{k\in\mathbf{K}} {\pi_{lk}}^{\hat{\alpha}_{lk} - 1} \mathrm{d}\boldsymbol{\pi}_l = \int_{\Pi_l} p(\boldsymbol{\pi}_l|\hat{\boldsymbol{\alpha}}_l) \\ &= 1 \end{aligned} \tag{20}$$

Therefore, deducing from Eq. (19), the predictive probability distribution for the topic-assignment k of sample $\mathbf{x}_{dn}$ in label l is:

$$p\left(\mathbf{z}_{dn} = k|\mathbf{l}_{dn} = l, L^{(\backslash dn)}, Z^{(\backslash dn)}, \boldsymbol{\alpha}\right) \propto \frac{\alpha_k + N_{lk}^{(\backslash dn)}}{\sum_{k\in\mathbf{K}} \alpha_k + N_l^{(\backslash dn)}} \tag{21}$$

$N_{lk}^{(\backslash dn)}$ represents the number of samples that were assigned to the topic k of global label l in addition to the current sample x_{dn}, $N_l^{(\backslash dn)} = \sum_{k\in\mathbf{K}} N_{lk}^{(\backslash dn)}$.

The integral of $\boldsymbol{\theta}$ is same as LDA in Eq. (11):

$$\begin{aligned} p(W|Z, \boldsymbol{\lambda}) &= \prod_{k\in\mathbf{K}} \int_{\Theta_k} p(\boldsymbol{\theta}_k|\boldsymbol{\lambda}) \prod_{d\in\mathbf{D}} \prod_{n=1}^{N_d} p\left\{\mathbf{w}_{dn}|\mathbf{z}_{dn} \right. \\ &= \left. k, \boldsymbol{\theta}_k\right) \mathrm{d}\boldsymbol{\theta}_k \propto \prod_{k\in\mathbf{K}} \int_{\Theta_k} \frac{\Gamma\left(\sum_{w\in\mathbf{W}} \lambda_w\right)}{\prod_{w\in\mathbf{W}} \Gamma(\lambda_w)} \\ &\prod_{w\in\mathbf{W}} \theta_{kw}^{\lambda_w + N_{kw} - 1} \mathrm{d}\boldsymbol{\theta}_k \propto \prod_{k\in\mathbf{K}} \frac{\Gamma\left(\sum_{w\in\mathbf{W}} \lambda_w\right)}{\Gamma\left(\sum_{w\in\mathbf{W}} \lambda_w + N_k\right)} \\ &\prod_{w\in\mathbf{W}} \frac{\Gamma(\lambda_w + N_{kw})}{\Gamma(\lambda_w)} \end{aligned} \tag{22}$$

Then the predictive probability distribution over the word-assignment $w$of topic k for observed sample $\mathbf{x}_{dn}$ is:

$$p\left(\mathbf{w}_{dn} = w|\mathbf{z}_{dn} = k, Z^{(\backslash dn)}, W^{(\backslash dn)}, \boldsymbol{\lambda}\right) \propto \frac{\lambda_w + N_{kw}^{(\backslash dn)}}{\sum_{w\in\mathbf{W}} \lambda_w + N_k^{(\backslash dn)}} \tag{23}$$

$N_{kw}^{(\backslash dn)}$ is the number of samples that were assigned to the word w of topic k in addition to the current sample x_{dn}, $N_k^{(\backslash dn)} = \sum_{w\in\mathbf{W}} N_{kw}^{(\backslash dn)}$.

Given the above, the collapsed joint distribution of (L, Z, W) is obtained by doing the integral of $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\psi})$ in Eqs. (14), (19) and (22).

$$\begin{aligned} p(L, Z, W|\boldsymbol{\Lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) &= p(L|\boldsymbol{\Lambda}, \boldsymbol{\beta}) p(Z|L, \boldsymbol{\alpha}) p(W|Z, \boldsymbol{\lambda}) \\ &\propto \prod_{d\in\mathbf{D}} \frac{\Gamma\left(\sum_{l\in\mathbf{L}'} \beta_l \Lambda_{dl}\right)}{\Gamma\left(\sum_{l\in\mathbf{L}'} \beta_l \Lambda_{dl} + N_d\right)} \\ &\prod_{l\in\mathbf{L}'} \frac{\Gamma\left(\beta_l \Lambda_{dl} + N_{dl} \Lambda_{dl}\right)}{\Gamma\left(\beta_l \Lambda_{dl}\right)} \\ &\cdot \prod_{l\in\mathbf{L}'} \frac{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k\right)}{\Gamma\left(\sum_{k\in\mathbf{K}} \alpha_k + N_l\right)} \\ &\prod_{k\in\mathbf{K}} \frac{\Gamma(\alpha_k + N_{lk})}{\Gamma(\alpha_k)} \\ &\cdot \prod_{k\in\mathbf{K}} \frac{\Gamma\left(\sum_{w\in\mathbf{W}} \lambda_w\right)}{\Gamma\left(\sum_{w\in\mathbf{W}} \lambda_w + N_k\right)} \\ &\prod_{w\in\mathbf{W}} \frac{\Gamma(\lambda_w + N_{kw})}{\Gamma(\lambda_w)} \end{aligned} \tag{24}$$

To simplify computation, the Dirichlet prior distributions are symmetric Dirichlet distributions:

$$\begin{aligned}\boldsymbol{\beta} &= \{\beta_l\}_{l\in\mathbf{L}'} = \{\beta, ..., \beta_{|\mathbf{L}'|=L+1}\}\\ \boldsymbol{\alpha} &= \{\alpha_k\}_{k\in\mathbf{K}} = \{\alpha, ..., \alpha_{|\mathbf{K}|=K}\}\\ \boldsymbol{\lambda} &= \{\lambda_w\}_{w\in\mathbf{W}} = \{\lambda, ..., \lambda_{|\mathbf{W}|=W}\}\end{aligned} \tag{25}$$

$\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl} = \sum_{l\in\mathbf{L}_d}\beta_l = \beta L_d$, $\sum_{k\in\mathbf{K}}\alpha_k = \alpha K$ and $\sum_{w\in\mathbf{W}}\lambda_w = \lambda W$ can be substituted to Eq. (24):

$$\begin{aligned}p(L, Z, W|\boldsymbol{\Lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) &= p(L|\boldsymbol{\Lambda}, \boldsymbol{\beta})p(Z|L, \boldsymbol{\alpha})p(W|Z, \boldsymbol{\lambda})\\ &\propto \prod_{d\in\mathbf{D}}\frac{\Gamma(\beta L_d)}{\Gamma(\beta L_d + N_d)}\prod_{l\in\mathbf{L}_d}\frac{\Gamma(\beta + N_{dl})}{\Gamma(\beta)}\\ &\cdot\prod_{l\in\mathbf{L}'}\frac{\Gamma(\alpha K)}{\Gamma(\alpha K + N_l)}\prod_{k\in\mathbf{K}}\frac{\Gamma(\alpha + N_{lk})}{\Gamma(\alpha)}\\ &\cdot\prod_{k\in\mathbf{K}}\frac{\Gamma(\lambda W)}{\Gamma(\lambda W + N_k)}\prod_{w\in\mathbf{W}}\frac{\Gamma(\lambda + N_{kw})}{\Gamma(\lambda)}\end{aligned} \tag{26}$$

Then, the prediction probability distribution of hidden variable $\mathbf{z}_{dn}$ and $\mathbf{l}_{dn}$can be computed from that collapsed joint distribution as a transition probability of state space in the Markov chain. Through Gibbs Sampling iteration, Markov chain converges to the target stationary distribution after the burn-in time. Finally, collecting sufficient statistic samples from the converged Markov chain state space and averaging among the samples, we can get a posteriori estimates of corresponding parameters.

Deducing from Eqs. (16), (21) and (23), the predictive probability distribution for the word-assignment $w$of topic k in label l for sample $\mathbf{x}_{dn}$ is:

$$\begin{aligned}&p\left(\mathbf{l}_{dn} = l, \mathbf{z}_{dn} = k, \mathbf{x}_{dn} = w|L^{(\backslash dn)}, Z^{(\backslash dn)}, W^{(\backslash dn)}, \boldsymbol{\Lambda}_d, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}\right)\\ &\propto p\left(\mathbf{l}_{dn} = l|L_d^{(\backslash dn)}, \boldsymbol{\Lambda}_d, \boldsymbol{\beta}\right)\cdot p\left(\mathbf{z}_{dn} = k|\mathbf{l}_{dn} = l, L^{(\backslash dn)}, Z^{(\backslash dn)}, \boldsymbol{\alpha}\right)\\ &\cdot\ p\left(\mathbf{w}_{dn} = w|\mathbf{z}_{dn} = k, Z^{(\backslash dn)}, W^{(\backslash dn)}, \boldsymbol{\lambda}\right)\\ &\propto\frac{\left(\beta_l + N_{dl}^{(\backslash dn)}\right)\Lambda_{dl}}{\sum_{l\in\mathbf{L}'}\beta_l\Lambda_{dl} + N_d^{(dn)}}\cdot\frac{\alpha_k + N_{lk}^{(\backslash dn)}}{\sum_{k\in\mathbf{K}}\alpha_k + N_l^{(\backslash dn)}}\\ &\cdot\frac{\lambda_w + N_{kw}^{(\backslash dn)}}{\sum_{w\in\mathbf{W}}\lambda_w + N_k^{(\backslash dn)}}\end{aligned} \tag{27}$$

Results

Dataset

To investigate the performance of the proposed method, we utilize two types of datasets. The first one is S.cerevisiae dataset (S.C) proposed in [19], and the second one is human dataset constructed by ourselves.

In S.C dataset, there are several sub datasets that constructed from different characteristics of yeast genome. Meanwhile, each sub dataset use two kinds of function annotation standard, FunCat and GO. We mainly use the sub dataset that depends on the amino acid sequence of protein and GO. What's more, to compare the performance of PFTP between difference label numbers, we construct a dataset named S.C-CC from S.C, which only includes GO terms belonging to cellular component. Then, there are two datasets constructed from S.C.

The human dataset is constructed from the Universal Protein Resource (UniProt) databank [2] and constructed by the similar way of reference [4]. Meanwhile, we construct two Human datasets for different word length, where the max word length of Human1 dataset is two alphabet, and which of Human2 dataset is three alphabet.

Due to the large number of GO terms in protein function dataset, we adopted a label space dimension reduction (LSDR) method to overcome the classification difficulty of classifiers. Boolean Matrix Decomposition (BMD) has been studied for LSDR recently, which can recovery the label space after classification conveniently. Therefore, a BMD method proposed in reference [20] has conducted in S.C and Human dataset. The statistics of above two datasets is displayed in Table 1. 'L' represents the number of GO terms after BMD; 'D' denotes the number of proteins in each dataset; 'W' denotes the size of vocabulary.

Parameter settings

PFTP model involves three parameters: α, λ and K. α and λ are the parameters of two Dirichlet distribution, where the larger the value of λ, the more balanced the probabilistic of word in a topic. According to the experience, we set $\alpha = 50/K$,$\lambda = 200/W$. The settings and impact of K value are explained later.

In the Gibbs sampling process of model training, we set the number of Markov chain as 1, the maximum number of iterations as 2000 times, where the number of iteration of burn-in time is set to 1000. We record the state space at intervals of 50 times on converged Markov chain, and 20 times of record is conducted. In the process of model predicting, we set the number of iterations as 1000 times. After 500

Table 1 The statistic of four datasets

Dataset	D	W	L
Human 1	4962	5297	1477
Human 2		400	
S.C	1692	400	1538
S.C-CC			319

times of iterations for burn-in time, we record the state space at intervals of 50 times.

Evaluation criterias

In all of our experiments, we use three representative multi-label learning evaluation criteria, including Hamming loss(HL), Average precision(AP) and One Error. Besides, we also use three kinds of area under Precision-Recall curve proposed in reference [19], including $\overline{AUPRC}$, $AU(\overline{PRC})$ and $\overline{AUPRCw}$. Meanwhile, the 5-fold cross validation is adopted to assess the performance of PFTP and contrast methods. The average results of 5 independent rounds are reported in following sections.

The impact of topic number on experimental results

K denotes the number of global topics. The analysis about impact of K on model performance is discussed in this section. According to the description of Section 2, as PFTP allocates one or more latent topics to each GO term, then the value of K should range from $L$to infinity in theory. Specifically, if we allocate only one topic to each GO term ($K = L$), then the model reduces to Labeled-LDA. Obviously, setting $K < L$makes our PFTP have no ability to discover the sub-structure of function. In our experiment, each function is assigned exactly the same number of topics for the simplicity of computation. For example, we set $K = 3L$, then each GO term corresponds to a topic set with three topics. In view of above reason, the lower bounded of K value is set to $2L$. On the other hand, although theory insists that the larger K value equals to the more refined sub-structure of label, incorporating more latent topics per function will increase the computational load. In reference [18], the impact of K value on the effectiveness of PLDA model has been discussed in several texts collections. Along with the growth of topic size, the performance of PLDA model approaches a fixed value which was obtained by a non-parametric model. In other words, the infinitely larger size of topics doesn't equal to an infinitely greater performance, but an unbearable running time. Therefore, we set the upper bound of K value as $5L$based on our empirical experience and the acceptable level of time overhead. In sum, the Kvalue should be set to an integer between $2L$ and$5L$. Then, the performance of PFTP under different Kvalue is shown in Fig. 4.

As shown in Fig. 4, all of the evaluation criteria value is relatively stable when $K$is set to$2L$~$4L$. Nonetheless, when Kvalue is greater than $4L$, the values of AP,$\overline{AUPRC}$,$AU(\overline{PRC})$ and $\overline{AUPRCw}$ decrease with the increase of K, the value of Hamming loss and One Error slowly increase with the increase of K. These results suggest that the optimum value range of K is $2L$ to$4L$. This was due to that the lower K value makes the fewer topics allocated to each label, and the higher K value makes the small difference of word distribution between topics. What's more, the problem of huge labels is particularly obvious in protein function dataset, even if a BMD method has applied to reduce the label dimension. Therefore, we set K as $3L$ in our experiment.

Evaluation against widely adopted method

Firstly, we compare PFTP with Labeled-LDA [4] and multi-label K-nearest neighbor (MLKNN) [21] on four datasets. MLKNN is a representative multi-label classifier and can be applied by an open source tool called Mulan [22]. Figure 5 shows the HL, AP, One Error, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRCw}$ values of these three models in SC, SC-CC, Human1 and Human2 dataset, respectively. For AP, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRCw}$, the larger the value, the better the performance. Conversely, for HL and One-Error, the smaller the value, the better the performance. The red asterisk of Fig. 4 represents the best result on each dataset.

As shown in Fig. 5, we can observe that PTPF shown more advantages in contrast to Labeled-LDA and MLKNN in four datasets. Concrete analysis is as follows:

For Human1 dataset, PFTP obtain a better performance in all evaluation criteria. On HL, PTPF achieves 9.7 and 2% improvements over Labeled-LDA and MLKNN. On One-Error, PTPF achieves 80 and 99% improvements over Labeled-LDA and MLKNN. On AP, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRCw}$, PFTP achieves 2.5, 0.2, 47 and 18% improvements over Labeled-LDA, and achieves 48, 40, 43 and 41% improvements over MLKNN. Obviously, the improvements on $\overline{AUPRC}$ and $\overline{AUPRCw}$ is more significant than $AU(\overline{PRC})$.

For Human2 dataset, PFTP obtain a better performance in four evaluation criteria except $AU(\overline{PRC})$ and $\overline{AUPRC}$. On HL, PTPF achieves 30 and 7.9% improvements over Labeled-LDA and MLKNN. On One-Error, PTPF achieves 66 and 99% improvements over Labeled-LDA and MLKNN. On AP and $\overline{AUPRCw}$, PFTP achieves 3.3 and 0.2% improvements over Labeled-LDA, and achieves 40 and 29% improvements over MLKNN. Nevertheless, on $AU(\overline{PRC})$ and $\overline{AUPRC}$, MLKNN and Labeled-LDA get better results respectively.

For S.C dataset, PFTP obtain a better performance in four evaluation criteria except HL and One-Error.

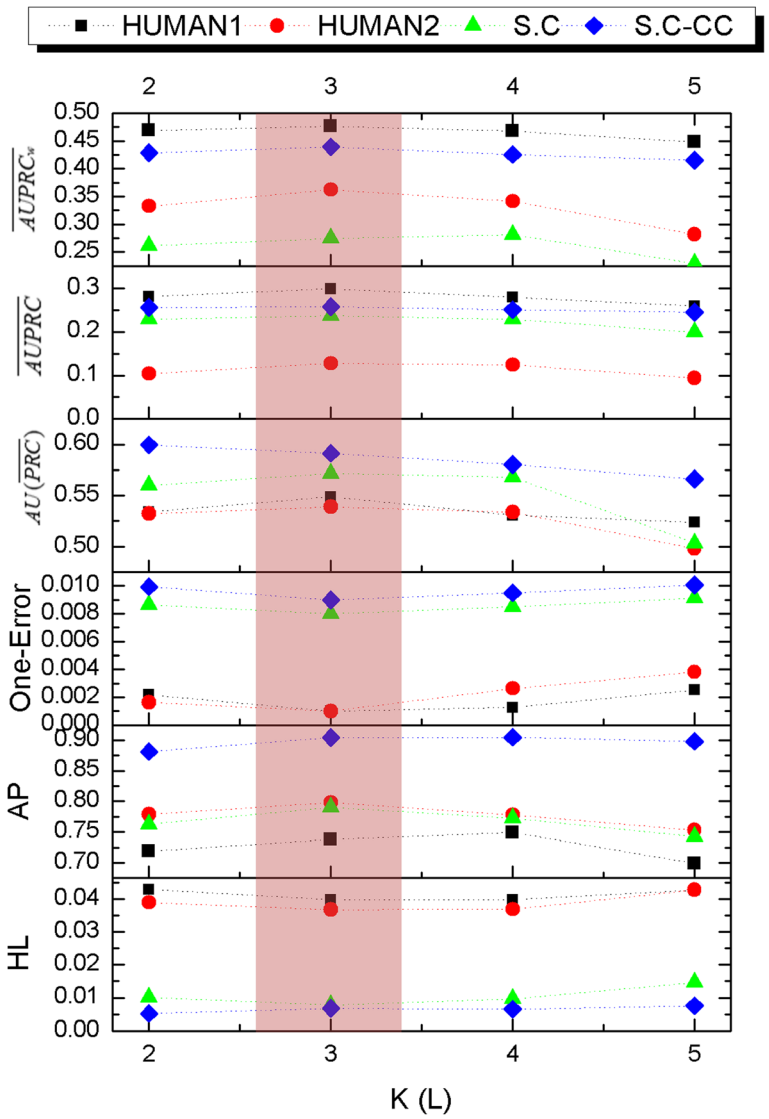

Fig. 4 The performance comparison of different K setting. For AP, $AU(\overline{PRC})$, $\overline{AUPRC}$ and $\overline{AUPRCw}$, the larger the value, the better the performance; for HL and One-Error, the smaller the value, the better the performance; The red background represents the best value range

On AP, $\overline{AUPRC}$ and $\overline{AUPRCw}$, PTPF achieves 2.8%, 22 and 16% improvements over Labeled-LDA, and achieves 48, 17 and 32% improvements over MLKNN; on $AU(\overline{PRC})$, the results of Labeled-LDA and PFTP are almost the same. Nevertheless, on HL, MLKNN gets better results than PFTP; on One-Error, almost identical results were obtained by these three methods.

For S.C-CC dataset, PFTP obtain a better performance on AP, $\overline{AUPRC}$ and $\overline{AUPRCw}$. On AP, PTPF achieves 2.6 and 27% improvements over Labeled-LDA and MLKNN. On $\overline{AUPRC}$, PTPF achieves 14 and 32% improvements over Labeled-LDA and MLKNN. On $\overline{AUPRCw}$, PTPF achieves 7.8 and 41% improvements over Labeled-LDA and MLKNN.

Besides, we compare PFTP with three hierarchal multi-label classification (HMC) algorithm based on decision tree, namely HMC/SC (single-label classification)/HSC (hierarchical single-label classification) [19]. These three algorithms have been studied on protein function prediction dataset and proved to be a kind of multi-label classifiers with great performance. Since the results of CLUS-HMC/SC/HSC in reference [19] are only on S.C dataset, the comparison results with our PFTP are also on S.C dataset, and are plotted in Fig. 6.

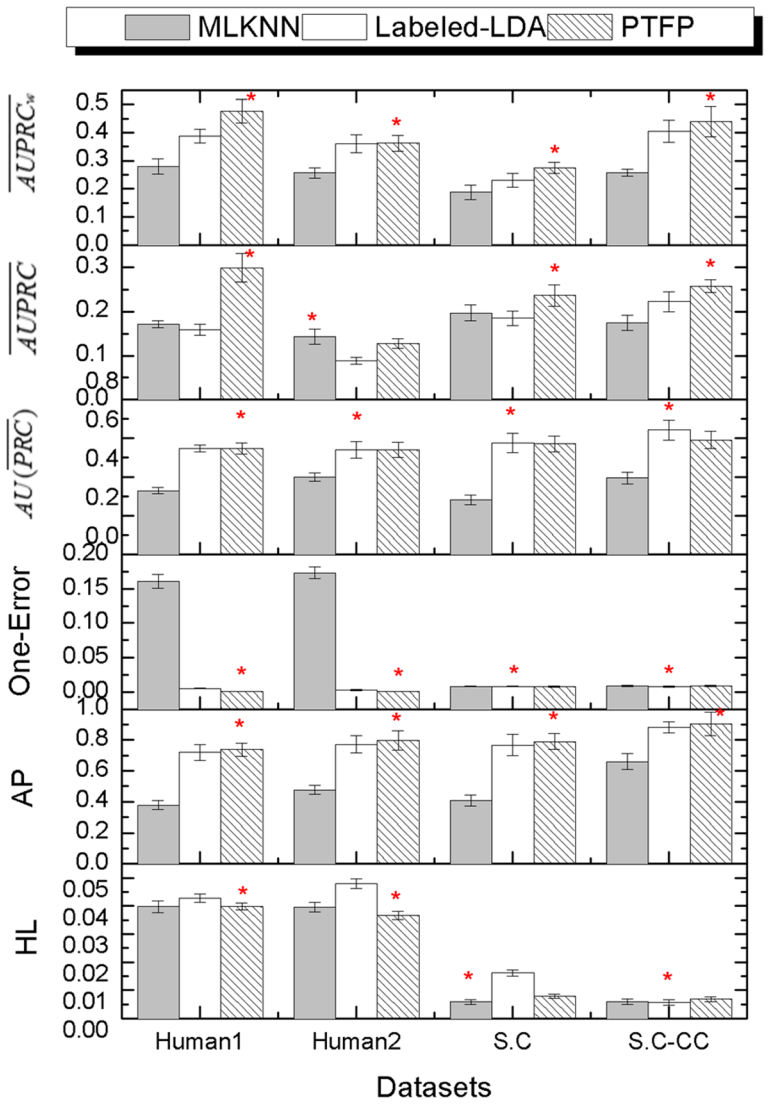

Fig. 5 The comparison results with PTFP and Labeled-LDA. For AP, $\overline{AUPRC}$, $AU(\overline{PRC})$ and $\overline{AUPRCw}$, the larger the value, the better the performance; for HL and One-Error, the smaller the value, the better the performance; the red asterisk on bar represents the best result on each dataset

On $\overline{AUPRC}$, our method exhibits dominant advantage against all of the three comparison methods. The performance improvements are 85, 85 and 84% against CLUS-SC, CLUS-HSC and CLUS-HMC, respectively. On $AU(\overline{PRC})$, PTPF achieves 65, 51 and 32% improvements over CLUS-SC, CLUS-HSC and CLUS-HMC. Nonetheless, on $\overline{AUPRCw}$, CLUS-HMC gets better results than PFTP.

The topics discovered by PFTP

The greatest strength of our protein function topic modeling is that, it can not only provide the function label probability distribution over proteins as an output, but also each function label can be explained as a probability distribution over topic subset, where each topic is represented as the probability distribution over amino acid blocks. To better understand this topic modeling process, we take GO term 'GO0016020' as an example, whose corresponding topics are shown in Table 2.

As shown in Table 2, the 2-mers BoW is used in this example. For Labeled-LDA, the one-to-one correspondence between label and word is the key design consideration. Therefore, 'GO0016020' only corresponds with a topic numbered 288, and also corresponds with a probability distribution over word. The top 20 words are listed from large to small order.

For PFTP model, each GO term is a partition of global topics set. Such as for S.C-CC dataset, the number of function label is 319, while the number of global topics is three times that of the labels, that's a

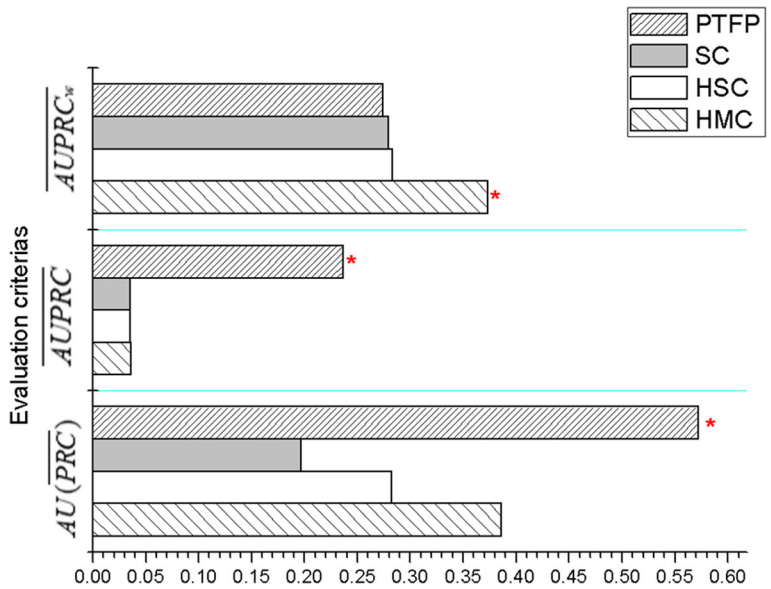

Fig. 6 The comparison results with PTFP and HMC/SC/HSC. For three evaluation criteria, the larger the value, the better the performance, and the red asterisk on bar represents the best result on each dataset

total of 958(including a background topic). Therefore, each GO term corresponds with four topics (including three local topics and one background topic). The topic number 863,864,865 and 1 are the four topics corresponded by 'GO0016020', where the number 1 is a background topic. Likewise, the top 20 words of these four topics are listed from large to small order.

Discussions

The results in Figs. 5 and 6 indicate that PFTP has the significant advantage against several widely adopted multi-label classifiers.

Compared with traditional multi-label classifiers(non--topic model), our method can further improve the accuracy of protein function prediction by introducing topics subset into supervised topic model, which can discover the topic that represents common semantic of documents and reflect the differences between labels and latent topics. Especially for CLUS-HMC/SC/HSC, our method exhibit the dominant advantage on $\overline{AUPRC}$. We attribute this success of our method to its utilization of BMD method on dataset. As the computation of $\overline{AUPRC}$ doesn't bias toward the accuracy of function label annotating more proteins, and focus on the average of whole accuracy. The GO term annotating fewer proteins will be deleted after BMD processing, and recovered after predicting, but the prediction accuracy don't reduce. In other words, the combination of PFTP and BMD can improve the average accuracy of protein function prediction.

Compared with Labeled-LDA, PFTP is able to discovery more-refined latent sub-structure of function label than Labeled-LDA. By introducing topic subset for each label in PTPF, the relationship between functions and variety words, labels and topics were disclosed. Therefore, we can anticipate that PFTP is a potential method to reveal a deeper biological explanation for protein functions.

Meanwhile, the performance comparison of different dataset is also shown in Fig. 4. For S.C-CC dataset, six evaluation criteria values vary relatively smoothly. It may be due to the fewer labels of S.C-CC dataset, then changing the K value doesn't lead to great impact on prediction effect. In the comparison of S.C and S.C-CC

Table 2 The topics discovered by two models

Method	Topic number	words
Labeled-LDA	288	GM IH LH VH LK IG GC IC AK VM FG AM LW IK VG VW FC IG FH GK
PFTP	863	LM SM FG FC VG SG FT VM IT IM AK LG LW LK SC FK ST AG VK GM
	864	GK IC VH GV SM TH IH VM AW GM AV GE VK AG IK LV GC GL TK LK
	865	LT GC AH IK IH LH SK SW LC YM VH TG IG LG AX FW FK SF YX AM
	1	LC AC AM VW VC GM AH AV AW VH GW AK AT GC TC GH LH LW EC TH

dataset, we find that the value of AP, $\overline{AU(PRC)}$, $\overline{AUPRC}$ and $\overline{AUPRCw}$ on S.C is lower than S.C-CC, and the value of One-Error and HL is almost equal between S.C and S.C-CC. This is due to the same word space and different label number between these two dataset. The fewer labels of S.C-CC can make a higher classifying performance. In the comparison of Human1 and Human2 dataset, we find that the value of $\overline{AUPRC}$ and $\overline{AUPRCw}$ on Human1 is higher than Human2; the value of AP on Human1 is lower than Human2; the value of One-Error, HL and $\overline{AU(PRC)}$ is almost equal on Human1 and Human2. These results show that, the classification performance of PFTP on Human1 and Human2 is almost the same, which reveal that the larger word space might not obtain a better classifying performance.

Conclusions

In this paper, we introduced an improved multi-label supervised topic model for predicting protein function. In our previous study, a multi-label supervised topic model Labeled-LDA has been applied to protein function prediction, which associates each label (GO term) with a corresponding topic directly. This way makes the latent topics to be completely degenerated, and ignores the differences between labels and latent topics. To address the faultiness, we proposed a Partially Function-to-Topic Prediction model for introducing the local topic subset corresponding to each function label. PFTP not only supports latent topics subsets within given function labels but also a background topic corresponding to a 'fake' function label. In a 5-fold cross validation experiment on predicting protein function, PFTP significantly outperforms compared methods. Due to the more-refined way of function label modeling, PFTP shows the effectiveness and potential value in predicting protein function through experimental studies. Meanwhile, there are several problems in topic modeling of protein function prediction to be improved, such as the introduction of protein extra features and hierarchical function label structure. However, multi-label topic model is a potential method in many applications of bioinformatics.

Abbreviations

BMD: Boolean Matrix Decomposition; BoW: Bag of Words; CGS: Collapsed Gibbs sampling; GO: Gene Ontology; HL: Hamming loss, AP: Average precision; HMC: Hierarchal Multi-label Classification; HSC: Hierarchical Single-label Classification; LDA: Latent Dirichlet Allocation; LSDR: Label Space Dimension Reduction; MLKNN: Multi-label K-nearest neighbor; PFTP: Partially Function-to-Topic Prediction; PLDA: Partially Labeled LDA; PLSA: Probabilistic Latent Semantic Analysis; S.C: S.cerevisiae; SC: Single-label Classification; SVM: Support Vector Machine; UniProt: Universal Protein Resource

Acknowledgements

We would like to thank the researchers in State Key Laboratory of Conservation and Utilization of Bio-resources, Yunnan University, Kunming, China. Their very helpful comments and suggestions have led to an improved version of paper.

Funding

This research was supported by the National Natural Science Foundation of China (no. 61862067, no. 61363021), and the Doctor Science Foundation of Yunnan normal university (no. 01000205020503090, no. 2016zb009). Publication costs are funded by the Doctor Science Foundation of Yunnan normal university (no. 2016zb009).

Availability of data and materials

The data and source code is available upon request.

About this supplement

This article has been published as part of *BMC Genomics Volume 19 Supplement 10, 2018: Proceedings of the 29th International Conference on Genome Informatics (GIW 2018): genomics*. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-10.

Authors' contributions

LT and WZ conceived the study, and revised the manuscript. LL analyzed materials and literatures, and drafted the manuscript. LT and MT participated in the literatures analyses. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]School of Information, Yunnan Normal University, Kunming 650500, Yunnan, China. [2]Key Laboratory of Educational Informatization for Nationalities Ministry of Education, Yunnan Normal University, Kunming 650500, Yunnan, China. [3]President's Office, Yunnan Normal University, Kunming 650500, Yunnan, China. [4]School of Software, Yunnan University, Kunming 650091, Yunnan, China.

Published: 31 December 2018

References

1. Weaver RF. Molecular biology (WCB Cell & Molecular Biology). 5th ed. New York: cGraw-hill Education; 2011.
2. Consortium UP. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2016;45(D1):D158–69.
3. Berman HM, Battistuz T, Bhat TN. The protein data Bank. Berlin: Atomic evidence: Springer International Publishing; 2016. p. 218–22.
4. Liu L, Tang L, He L, Wei Z, Shaowen Y. Pedicting protein function via multi-label supervised topic model on gene ontology. Biotechnol. Biotechnol. Equip. 2017;31(1):1–9.
5. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids. 1997;25:3389–402.
6. Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32(Suppl 1):D258–61.
7. Cao R, Cheng J. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. Methods. 2016;93:84–91.
8. Erdin S, Venner E, Lisewski AM, Lichtarge O. Function prediction from networks of local evolutionary similarity in protein structure. BMC bioinformatics. 2013;14(3):S6.
9. Yu G, Rangwala H, Domeniconi C, Zhang G, Zhang Z. Predicting protein function using multiple kernels. IEEE/ACM Trans Comput Biol Bioinf. 2015; 12(1):219–33.
10. Fodeh S, Tiwari A, Yu H. Exploiting PubMed for protein molecular function prediction via NMF based multi-label classification. In: Proceeding of

international conference on data mining workshops. 2017 IEEE conference on; 2017. p. 446–51.
11. However. Orderly roulette selection based ant Colony algorithm for hierarchical multilabel protein function prediction. Math Probl Eng. 2017; 2017(2):1–15.
12. Wang H, Yan L, Huang H, Ding C. From protein sequence to protein function via multi-label linear discriminant analysis. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(3):503–13.
13. Pinoli P, Chicco D, Masseroli M. Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations. In: Proceeding of the 13th international conference on bioinformatics and bioengineering (BIBE). 2013 IEEE conference on; 2013. p. 1–4.
14. Masseroli M, Chicco D, Pinoli P. Probabilistic latent semantic analysis for prediction of gene ontology annotations. In: Proceeding of international joint conference on neural networks (IJCNN). 2012 IEEE conference on; 2012. p. 1–8.
15. Pinoli P, Chicco D, Masseroli M. Latent Dirichlet allocation based on Gibbs sampling for gene function prediction. In: Proceeding of international conference on computational intelligence in bioinformatics and computational biology. 2014 IEEE conference on; 2014. p. 1–8.
16. Dumais ST. Latent semantic analysis. Ann Rev Inf Sci Technol. 2004; 38(1):188–230.
17. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
18. Ramage D, Manning CD, Dumais S. Partially labeled topic models for interpretable text mining. In: International conference on knowledge discovery and data mining, 2011 ACM conference on; 2011. p. 457–65.
19. Vens C, Struyf J, Schietgat L, Džeroski S, Blockeel H. Decision trees for hierarchical multi-label classification. Mach Learn. 2008;73(2):185–214.
20. Sun Y, Ye S, Sun Y, Kameda T. Improved algorithms for exact and approximate Boolean matrix decomposition. In: International conference on data science and advanced analytics, 2015 IEEE conference on; 2015. p. 1–10.
21. Zhang M, Zhou Z. ML-KNN : a lazy learning approach to multi-label learning. Pattern Recogn. 2007;40(7):2038–48.
22. Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. In: Maimonn O, Rokach L, editors. Data mining and knowledge discovery handbook. New York: Springer US; 2009. p. 667–85.