Bayesian Collective Markov Random Fields for Subcellular Localization Prediction of Human Proteins

Lu Zhu*
Bielefeld University
Bio-/Medical Informatics Department
International Research
Training Group 1906 (DiDy)
Bielefeld, NRW 33611, Germany
lzhu@techfak.uni-bielefeld.de

Martin Ester
Simon Fraser University
School of computing science
Burnaby, BC V5A 1S6, Canada
ester@cs.sfu.ca

ABSTRACT

Advanced biotechnology makes it possible to access a multitude of heterogeneous proteomic, interactomic, genomic, and functional annotation data. One challenge in computational biology is to integrate these data to enable automated prediction of the Subcellular Localizations (SCL) of human proteins. For proteins that have multiple biological roles, their correct *in silico* assignment to different SCL can be considered as an imbalanced multi-label classification problem. In this study, we developed a Bayesian Collective Markov Random Fields (BCMRFs) model for multi-SCL prediction of human proteins. Given a set of unknown proteins and their corresponding protein-protein interaction (PPI) network, the SCLs of each protein can be inferred by the SCLs of its interacting partners. To do so, we integrate PPIs, the adjacency of SCLs and protein features, and perform transductive learning on the re-balanced dataset. Our experimental results show that the spatial adjacency of the SCLs improves multi-SCL prediction, especially for the SCLs with few annotated instances. Our approach outperforms the state-of-art PPI-based and feature-based multi-SCL prediction method for human proteins.

KEYWORDS

Human protein subcellular localization; markov random field; transductive learning; imbalanced multi-label classification.

1 INTRODUCTION

Detailed molecular knowledge of the human proteome has become an important asset in the understanding of human biology and disease. Rapid advances in biotechnology have made available a variety of high-throughput experimentally obtained proteomics and interactomics datasets [14, 22], and knowledge of SubCellular Localization (SCL) of proteins can provide important insights for understanding their functions in cells and the mechanism of disease [12]. Owing to the annotation efforts of model organism databases, high-quality subcellular localization information for human proteins can be obtained from various curated sources. However, manually annotating a protein, especially determining the subcellular localization using the enormous data from heterogeneous source, is always a challenging and low-throughput task.A variety of computational methods have been developed for predicting the SCL of proteins for various organisms [3, 8] in the past decade. Nevertheless, there are relatively few efficient specific prediction tools for human proteins in the face of rapidly increasing numbers of newly identified proteins.

Protein features, especially the sequence-based features, are always the essential part in various protein SCL predictors [4, 7, 10]. To carry out different functions, one protein can be located in different SubCellular Compartments (SCCs) simultaneously or at different times during different biological processes, e.g. protein trafficking. Sequence-based prediction methods have been successfully applied to genome-wide large-scale protein annotations and analysis. However it is hard to apply these methods to detect the translocation of proteins due to the fact that the primary sequences of the translocated protein are always about the same. The biological functions of proteins are carried out by interacting with other proteins. To interact, proteins (or any other molecules) must necessarily share a common SCC, or an interface between physically adjacent SCCs, transiently or conditionally. The SCL of a protein can therefore be inferred from the SCL of its interacting partners. Hence biological network information can complement feature-based approaches to SCL prediction.

Several methods have been developed which take advantage of Protein-Protein Interaction (PPI) networks to predict the SCL of proteins for different orgnisms using data integration from multiple data sources [13, 17, 21]. However, these approaches mainly focus on the co-localization (in the same SCL) of interacting proteins. The importance of the spatial adjacency among SCCs was underestimated. It was not investigated whether a protein SCL (e.g. plasma membrane) can be also inferred by its interacting partners in the adjacent SCLs (e.g. Extracellular and Cytoplasm). Secondly, for the proteins whose interacting partners are poorly annotated, the information of the adjacent SCLs can be used as the major prediction resource. In this study, we investigated whether the spatial adjacency among SCLs can improve PPI-based SCL prediction.

Conventional machine-learning approaches, such as supervised learning, predict protein SCLs by extracting information only from existing annotation. However, the number of unreviewed proteins

*Corresponding author

ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.
 ISBN 978-1-4503-4722-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3107411.3107412

increases at a remarkably faster rate than that of experimentally-annotated ones. It was shown that transductive leaning approaches are able to take advantage of the large number of available unknown data to improve the accuracy of classification [18, 23]. On the other hand, proteins are often annotated with multiple SCLs. The Multi-Labeled Datasets (MLDs) of protein SCLs are typically heavily imbalanced. The learning from an imbalanced multi-label classification is a well-known challenge in classification [6].

A Markov Random Field (MRF) is a graphical model of a joint probability distribution. Many problems in computer vision such as image segmentation, image restoration and systems biology such as identification of differentially expressed genes [24], protein function prediction [16] involve the solution of a probability distribution defined by a discrete MRF. In this paper, we propose a Bayesian Collective MRFs (BCMRFs) to predict the multi-SCLs of human proteins considering PPI network features, the proteins features, the spatial adjacency of SCCs and the imbalance of the dataset. The key contributions are summarized as follows:

(1) We introduce weighted MRFs based on the PPI network with label propagation to predict the SCL of the proteins in the network.
(2) We propose Collective MRFs, one MRF per SCL, which are trained collectively to exploit the spatial adjacency among SCLs.
(3) We show the transductive learning method is more efficient than supervised learning method.
(4) We discuss the drawbacks of the imbalance of SCL datasets for protein SCL prediction and show that it can be improved by balancing the minor class with the major class.
(5) We performed experiments evaluating the performance on a human protein SCL benchmark set. Our method outperforms the state-of-the-art methods including the PPI-based approache DC-kNN [17] and the feature-based approach Hum-mPLoc 3.0 [26].

The rest of the paper is organized as follows: In section 2 we introduce our MRFs and the corresponding learning procedure. section 3 details the experiment protocol and section 4 shows the experimental results. At last, we conclude our work along with the discussion of directions of future work.

2 THE BAYESIAN COLLECTIVE MRF MODEL

In this section, we firstly give basic definitions and notations of MRF, and the rational of using MRFs for protein SCL prediction. Then, we introduce our Bayesian Collective Markov Random Fields (BCMRFs) for predicting the multi-SCLs of human proteins. The BCMRFs are formed by interatively collecting and optimizing the labels from multiple binary Bayesian MRFs.

2.1 Markov Random Field (MRF) on protein SCL prediction

Markov Random Field is a graphical model of a joint probability distribution. It consists of an undirected graph $G = (X, E)$ in which the nodes X represent random variables $\mathbf{X} = X_1, X_2, \dot{X}_n$, where each variable $X_i \in \mathbf{X}$ takes a value from the label set $\mathcal{L} = l_1, l_2, \ldots, l_k$. A labeling $\mathbf{x}$ refers to any possible assignment of labels to the random variables and takes values from the set $\mathcal{L}^n$.

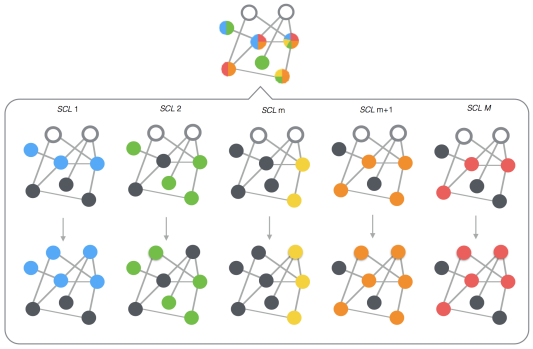

Figure 1: Multiple binary MRFs. The graph with multi-colored nodes on the top represents the general PPI network. Each node on the graph represents a protein associated with in total M SCL annotation terms. This network can be derived to M PPI networks for each single SCL term. The nodes are colored or in grey which represent 1 and 0 respectively if the SCL annotation of the protein is available for any of the M SCL terms. Otherwise, the node is not colored. And these nodes (proteins) are in need to be assigned with SCLs.

The posterior distribution $Pr(\mathbf{x}|\mathbf{D})$ over the labellings of the random field is a *Gibbs* distribution if it can be written in the form:

$$Pr(\mathbf{x}|\mathbf{D}) = \frac{1}{Z} \exp(-\sum_{c \in C} \phi_c(\mathbf{x}_c)) \quad (1)$$

where Z is a normalizing constant known as the partition function, and C is the set of all cliques. The term $\psi_c(\mathbf{x}_c)$ is known as the potential function of the clique c where $\mathbf{x}_c = x_i, i \in c$.

The corresponding Gibbs energy function E: $\mathcal{L}^n \longrightarrow \mathbb{R}$ maps any labelling $\mathbf{x} \in \mathcal{L}^n$ to a real number $E(\mathbf{x})$ called its energy. Energy function are the negative logarithm of the posterior probability distribution of the labeling. Maximizing the posterior probability equals to minimizing the energy function and leads to the MLE or MAP solution, which is defined as $\mathbf{x} = argmin_{x \in L} E(\mathbf{x})$.

Energy functions can be decomposed into sum over unary(ϕ_i) and pairwise(ϕ_{ij}) potentials as:

$$E(\mathbf{x}) = \sum_{i \in v} \phi_i(x_i) + \sum_{(i,j) \in \varepsilon} \phi_{ij}(x_i, x_j) \quad (2)$$

where v is the set of all random variables and ε is the set of all pairs of interacting variables. However, the potential functions could be with more interacting variables[15].

As we described in Section 1, the SCLs of a protein can be inferred by the SCLs of its physically interacting proteins. A physical PPI network $G = (P, I)$ with N proteins, $N = |P|$, that are assigned in M different SCLs in total fullfills the definition and properties of a MRF. It's reasonable to apply MFRs on PPI network to predict the SCL(s) of a set of proteins in the network. Moreover, a PPI network in which each protein is labeled by single or multi SCLs can be considered as multi-label MRFs. Inspired by the statistical power of those MRF models, we applied MRFs to PPI network for solving protein SCL prediction problem.

Using the binary reference approach [9], for the SCL noted as $l_m, 1 \le m \le M$, the network is encoded in an N-dimensional vector $\mathbf{x} = \{x_1, \ldots, x_N\}$, where $x_i = 1$ if the protein $p_i, 1 \le i \le |P|$ is assigned with l_m, else $x_i = 0$. The multi-label classification problem is thus reduced to multiple binary classification problems (Figure 1). For each SCL, we build corresponding binary MRFs to predict SCL labeling of unknown proteins by maximizing the posterior probability distribution of the SCL labeling of proteins. The following elements are used in our MRFs model: (1) prior probability of any protein being located in l_m, (2) the number of interacting neighbors being located in l_m, (3) the number of interacting neighbors being located in the adjacent SCLs of l_m, and (4) the sequence-based features of protein.

Meanwhile, the quality of PPI data and the connectivity of PPI network are crucial for inferring the SCL of a protein by its interacting neighbors. However the confidence of PPIs varies from one to another depending on the method, and the size of experiment etc.[5]. To balance of having a high quality of PPI network and reduce the risk of losing valuable information by removing too many edges, we use the confidence scores of the PPIs to weight our MRFs. The detailed method is described in the following sections.

2.2 The weighted markov random field model

By definition, the posterior distribution $Pr(\mathbf{x})$ over the SCL labelings of the MRF is a *Gibbs* distribution which can be written in the form:

$$Pr(\mathbf{x}) = \frac{1}{Z}\exp(-E(\mathbf{x})) \tag{3}$$

where Z is a normalizing constant known as the partition function. $E(\mathbf{x})$ is the energy function of the MRFs which is defined as follows:

$$\begin{aligned} E(\mathbf{x}) = -(&\sum_{i\in v}\phi_i^S(x_i) + \sum_{i\in v}\phi_i^F(x_i,F_i) + \sum_{i,j\in\varepsilon}\omega_{i,j}\phi_{ij}^P(x_i,x_j) \\ &+ \sum_{i,j\in\varepsilon}\omega_{i,j}\phi_{ij}^A(x_i,x_j,A_{ij})) \end{aligned} \tag{4}$$

with the unary potential

$$\phi_i^S = \begin{cases} 0 & x_i = 0 \\ \alpha & x_i = 1 \end{cases} \tag{5}$$

where α is the probability of a protein located in l_m. $\phi_i^F(x_i, F_i)$ is feature-based potential. F_i is a vector that includes the features for protein i. Conditional probability of a protein p_i being located in l_m given its features $Pr(x_i = 1|F_i)$.

$$\phi_i^F(x_i,F_i) = \begin{cases} 0 & x_i = 0 \\ \eta Pr(x_i = 1|F_i) & x_i = 1 \end{cases} \tag{6}$$

with

$$Pr(x_i = 1|F_i) = Pr(x_i = 1)\prod_{f=1}^{F} Pr(F_i^f|x_i = 1) \tag{7}$$

We includes thirty features which are generated from previous widely used sequence-based protein SCL predictor YLoc[4] into our model. These features include various types from simple amino acid composition to annotation information.Certain features are general such as protein size, number of small residues etc., while others specifically describing one certain SCL only. η is an unknown parameter associate to the ensemble of the 30 features F_i for protein i. The class priors and the feature probability distributions are estimated using the entropy-based supervised discretization of the training data. The final probabilities are obtained by normalizing the posterior such that the sum of all posterior is one. η together with other unknown parameters are estimated during parameters learning process.

ϕ^P is the pairwise potential of two proteins locating in l_m.

$$\phi_{ij}^P(x_i,x_j) = \begin{cases} 0 & (i,j) \notin \varepsilon \\ 0 & (i,j) \in \varepsilon \;\&\; x_i = x_j = 0 \\ \beta^{11} & (i,j) \in \varepsilon \;\&\; x_i = x_j = 1 \\ \beta^{10} & (i,j) \in \varepsilon \;\&\; x_i = 1 - x_j \end{cases} \tag{8}$$

where $\omega_{i,j}$ is a constant parameter , the confidential score of the PPI between P_i and P_j. $\phi_{ij}^A(x_i, x_j, A_{ij})$ is the potential which depends on if the protein p_i interacts with the proteins locating in the adjacent SCLs of l_m,

$$\phi_{ij}^A(x_i,x_j,A_{ij}) = \begin{cases} 0 & i,j \notin \varepsilon \\ \sum_{h=1}^{H}\mu_h A_{ij}^h & (i,j) \in \varepsilon \;\&\; x_i = 1 \end{cases} \tag{9}$$

where H is the total number of adjacent SCLs of SCL l_m. Given a set of H adjacent SCLs of SCL l_m, for each protein p_i which has N_{ne} of neighbors, we construct an $N_{ne} \times H$ binary matrix A, where the element A_{ij}^h is equal to 1 if protein p_i has an interacting neighbor p_j located in the adjacent SCL l_h and 0 otherwise. μ_h is an unknown parameter for the adjacent SCL l_h. The parameters $\alpha, \eta, \beta^{11}, \beta^{10}$, and μ are estimated during optimization.

2.3 Gibbs sampler and likelihood estimation

Energy functions are the negative logarithm of the posterior probability distribution of the SCL labeling. Maximizing the posterior probability equals to minimizing the energy function, which is defined as $\mathbf{x} = argmin_{x\in L}E(\mathbf{x})$. In this study we apply the approximation method Maximum pseudo-likelihood estimation(MPLE) to solve the maximization problem [1, 16]. Since the SCL datasets are usually highly imblanced, the posterior $Pr_\theta(x_i|\mathbf{x}_{-i})$ will tend to be overwhelmed by the majority classes (in this case negative examples in individual binary classifier). In order to deal with this problem, an imbalance coefficientis used to re-balance the influence on the joint likelihood by enhancing the minority classes [11]. Thus the re-balanced pseudo-likelihood function (PLF) can be written as

$$PLF(\mathbf{x}) = \prod_{i=1}^{N}(Pr(x_i|\mathbf{x}_{-i}))^{c_i^m} \tag{10}$$

where c_i is the imbalance coefficient

$$c_i^m = \begin{cases} \frac{n^-}{n^+} & x_i = 1 \\ 1 & x_i = 0 \end{cases} \tag{11}$$

where n^+ and n^- denote the numbers of positive samples and negative samples for the SCL l_m, respectively.

2.4 Collective MRFs

In our MRFs, each variable x_i in vector $\mathbf{x} = \{x_1, \ldots, x_N\}$, represents whether a protein being located to l_m or not. For protein p_i, it is possible that its neighbors located in adjacent SCLs are also unknown. To respect the property of MRF, we need to initialize the

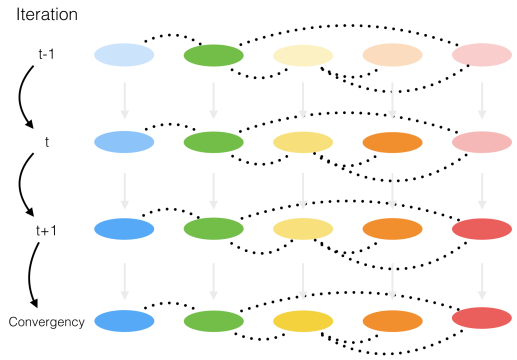

Figure 2: Collective MRFs. Each colored eclipse represents a MRF for one SCL term. The different shades of color represent the PLF value. The deeper the color is, the higher the PLF value calculated from this MRF is. The dotted line between eclipses represents the spatial relationship of SCLs.

labels of unknown proteins. Instead of using random labelings, we initialize in a more efficient way by the labeling results from the MRFs model without considering the adjacent SCLs. The results from the previous MRFs $MRF - l_m^t$ are collected and used in the next MRFs, such as $MRF - l_{m+1}^t, \ldots, MRF - l_M^t$. This process is repeated iteratively until the convergence of the pseudo-likelihood Equation 10. We name these MRFs as collective MRFs (see Figure 2 and algorithm 1).

Algorithm 1: Collective MRFs

Input: M partial labeled network for the M SCL terms
Output: M fully assigned network for the M SCL terms
for *each SCL terms* **do**
  Initialize the x_i values of unknown proteins using MRF model without adjacent SCLs potential ϕ^A (Equation 9).
end
while *not converge* **do**
  for *each SCL terms* **do**
    Optimize the $PLF^t(\mathbf{x})$.
    Calculate acceptance probability r comparing with the $PLF^{t-1}(\mathbf{x})$.
    if $r > r^*_{unif}$ **then**
      Update the labeling of $\mathbf{x}$ according to $PLF^t(\mathbf{x})$.
    end
  end
end
$*$: r_{unif} is a random variable follows uniform distribution.

3 EXPERIMENTAL SETUP

3.1 Dataset

A recently published high-quality human protein SCL benchmark set from the subcellular localization database Compartments [2] was used to evaluate the performance of our method. In total 9 SCLs including Cytosol, Endoplasmic Reticulum, Lysosome, Extracellular space, Golgi apparatus, Mitochondrion, Nucleus, Peroxisome and Plasma membrane are used for evaluation. The dataset was created from UniprotKB/Swiss-prot and Human Protein Atlas (HPA).

The corresponding protein sequences for generating the features from YLoc were retrieved from UniprotKB/Swiss-prot (version 2016.08). The PPI data were retrieved from the interactom browser - Mentha [5] (version 2016.09). It limits itself to direct physical PPIs curated by members of the International Molecular Exchange consortium (IMEx) [19]. Each PPI is associated with a reliability score which takes the evidienes such as experimental method, size of experiments and relevant literature into account [5]. In Figure 4 we notice a dramatic reduction of PPI size with a cutoff of reliability score 0.25. We consider that with this cutoff value, we can remove most of the low quality PPIs in the network. For the remaining PPIs, we use the reliability scores to weighted our MFRs. In the filtered connected PPI network, 5496 proteins are SCL-known while 1299 protein have no SCL annotation available. Figure 3 further shows the distribution of the SCLs of our human proteins data set. As can be seen, of the 5496 proteins, 4367 are single-SCL located proteins, 1129 have from 2 to 7 SCL annotations. As shown in the pie chart, the majority of single-SCL proteins locate in the nucleus which is consistent with the distribution of the overall proteins. For the proteins locate in 2 or more SCLs, nucleus shows less and less portion in the distribution. Therefore, the single-SCL protein plays more significant roles in shaping the overall distribution of the data set. Nevertheless, the multi-SCLs proteins which takes big percentage of the population can not be ignored.

3.2 Evaluation

To evaluate the prediction performance of our method, we perform 6 fold cross validation. For 1000 out of 5496 proteins, we mask their SCL labels, and treat them as unknown protein. Hence, 2299 proteins in the network are unlabeled. And the predicted label of these masked protein are used for performance evaluation. The dataset stratification was done by using R package "utiml" [20].
The traditional performance measures are difficult to apply for multiple SCL prediction. To better reflect the multi-label capabilities of classifiers, we use the popular multi-label measures including Precision (PRC), Recall (RCL), F1-score (F1), Average Precision (AP) and Hamming Loss (HL) [20]. Except HL, for all the rest of performance measures, the higher the measures, the better the prediction performance. To keep the consistency, we show the 1-HL instead of HL for the evaluation.

3.3 Comparison partners

In our experiments, in order to investigate the effects of different potentials described in section section 2, we build 4 versions of MRFs which include different combinations of potentials, such as MRFs with PPI only (M1) ,with PPI and SCL spatial adjacency (M2), with PPI and protein features (M3), and the MRFs with all three defined as Equation 4 (M4).

Moreover, we compared our MRFs with state-of-art SCL prediction methods, including:

- DC-kNN proposed by [17] provides the best SCLs predicting result for human proteins based on PPI. In their study, they reported the SCLs for 4366 human proteins with no

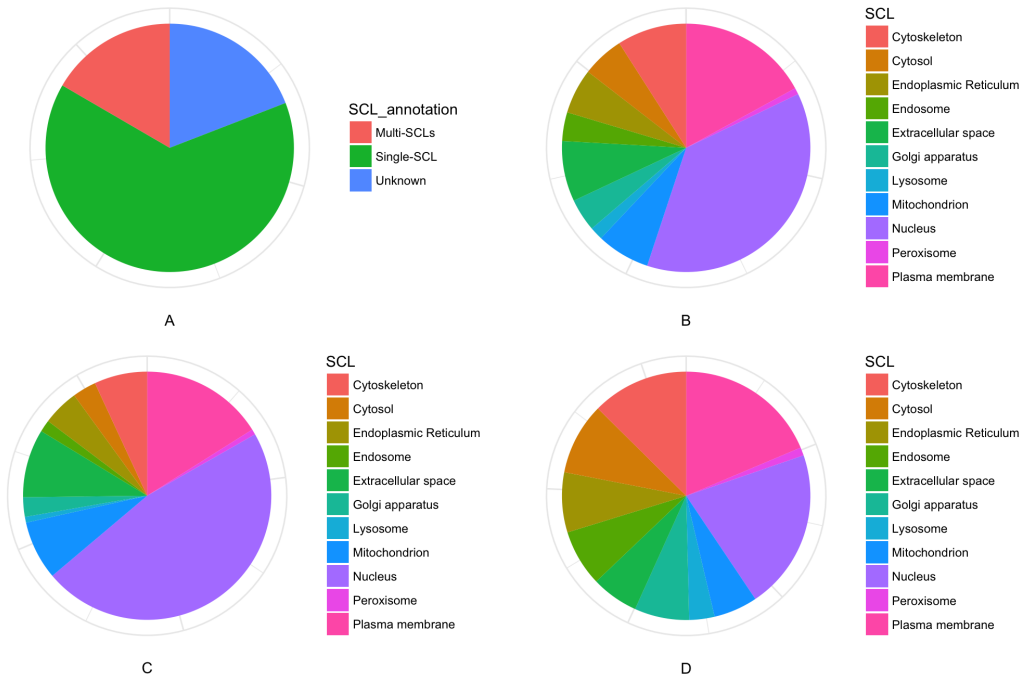

Figure 3: Information of our human protein data set. A. SCL annotation of proteins; B. Overall distribution of protein in SCL classes; C. Distribution of single-SCL protein; D. Distribution of multi-SCLs proteins.

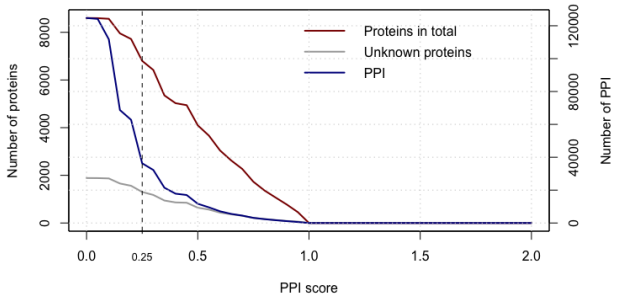

Figure 4: Protein-protein interactions of test dataset controlled by the confidential scores.

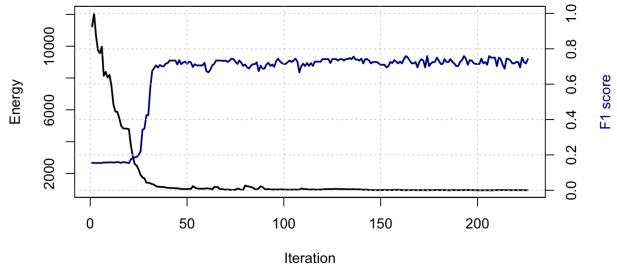

Figure 5: Relationship between the likelihood and prediction performance.

SCL previously known at the time in 2008 predicted by their method. From then to 2016, 1704 of these proteins has been reported in various SCLs. We collect the SCL annotations following the same criteria as their benchmark [17].

- Hum-mPLoc 3.0 is a most recent protein feature-based SCL predictor for human proteins [26]. The predicted SCLs of 5390 human proteins from their database are used for the comparison.

4 EXPERIMENTAL RESULTS

4.1 Likelihood and prediction performance

In our method, we minimize the energy function to correctly predict the SCL label of proteins. In other words, the higher the calculated conditional probability of a protein given its interacting neighbors for a certain SCL l_m, $1 \leq m \leq M$, the more confident that this protein locates in l_m, which infers that the overall prediction performance achieves for l_m should be positively correlated with the data likelihood. Figure 5 shows that the lower the energy (the negative logarithm of the likelihood) is, the higher the F1 score is which confirms the concept.

4.2 Effects of different potentials

To investigate the effect of the potential described for the prediction, we compare the performances including of the four versions of MRFs M1, M2, M3, M4.

4.2.1 Single-SCL prediction. We firstly compare the performance of the 4 models for each SCL class individually. **M2 VS M1** : Fig 6 shows that the spatial SCL adjacency relation of interacting proteins can improve the prediction for the majority of the SCL classes, expect Lysosome and Peroxisome. There are even the decrements of prediction performance. Firstly, these two SCL classes are highly imbalanced with few positive labels (see Figure 9). Moreover, we notice that the prediction on the SCL Cytosol is quite poor. Therefore, the MRFs of Lysosome and Peroxisome can not gain the correct information from their only spatial adjacent SCL Cytosol to increase their prediction performance. In order to put the spatial adjacency to good use, it is necessary to firstly improve the overall performance. Therefore, we integrated the potential based on protein features into MRF model (M3). With regard of Cytosol, it is an intracellular fluid which comprises most of the cellular organelles, and involved in many biological processes. The low performance could be due to its complication. The features can not improve the prediction performance. Finally, adding the SCL adjacency potential above on M3, we observe the improvement of prediction performance on most of the SCL classes.

4.2.2 Multi-SCLs prediction. As can be seen from Figure 7, M2 outperforms M1 which means additional spatial adjacency can improve the performance comparing with the simple SCL inference based on PPI only. However, the improvement is limited due to that M2 cannot efficiently gain correct knowledge from the adjacent SCLs which are poorly predicted. As expected, M3 significantly improve of prediction performance by adding the features of proteins on the model of M1. M4 can achieve the best performance of all. Comparing with M3 in particular, together with the observations of single-SCL predictions, we can conclude that the improvement is owing to that the model can efficiently gain the correct knowledge from the adjacent SCLs. However, in order to show a larger improvement of performance of the multi-SCLs prediction by adding the spatial adjacency on the proteins features in the model (M4 against M3), an additioanl tuning of parameters would be necessary.

4.3 A collective process improves the performance

To demonstrate how the collective MRFs can help to improve the performance of our SCL prediction, we show the changes of performance of M4 during the 21 iterations in Figure 8. Overall, the F1 scores gradually increase from initialization (iteration 1), single MRFs (iteration 2) and collective MRFs (from the 3^{rd} iteration). The performances stay stable as the pseudo likelihood value of BCMRFs converge.

4.4 Transductive learning from imbalanced MLDs

Our human protein dataset is highly imbalanced since some of the labels are very frequent whereas most others are rarely used.

Table 1: F1 scores with/without imbalance correction.

Model	M1	M2	M3	M4
With imbalance coefficient	0.637	0.641	0.71	**0.732**
Without imbalance coefficient	0.616	0.632	0.701	0.722

Table 2: F1 scores for transductive VS conventional.

Model	M1	M2	M3	M4
Transductive learning	0.648	0.652	0.743	**0.759**
Conventional learning	0.602	0.647	0.684	0.692

The imbalance level of a MLD can be effectively measured by the imbalance ratio (*IRLbl*) [6]. Figure 9 shows that the SCLs such as Lysosome and Peroxisome are highly imbalanced compared to the other SCLs, with IRLbl of 22. 4 and 44.18 respectively.

Facing the imbalance problem, the popular solution is data resampling including under-sampling and over-sampling [6]. However, in our case the re-sampling techniques cannot be applied due to our method being highly sensitive to the topology of PPI network. The inference of SCL in this approach depends on the number of physical interactions. Under-sampling and over-sampling are based on the deletion of true interactions or repetition of existing interactions which can largely change the topology of the network and thus mislead the MRFs. Therefore, in this study we handle the imbalanced MLD problem by introducing imbalance coefficient (Equation 11). We compare the prediction performances of the BCMRFs with and without the imbalance coefficient. The results in Table 1 shows that the MRFs with the imbalance coefficient can improve the performance.

Furthermore, we compare the prediction results of the BCMRFs built on the complete PPI network including the unknown proteins against the BCMRFs built only on the sub-network of the annotated proteins. As we can see from Table 2, the MRFs of transductive learning outperforms the MRFs of the conventional learning.

4.5 Comparison with existing methods

To further demonstrate the performance of our method, we compare our BCMRFs with the only available PPI-based approach for predicting human protein SCLs, DC-*k*NN [17] and the state-of-art protein feature-based method Hum-mPLoc 3.0 [26]. DC-*k*NN is a physical PPI-based prediction method using a *k*-nearest neighbors classification with binary reference approach. Due to the unavailability of the program and of its prediction results, the dataset we use to compare our methods only contains 1704 human proteins (see subsection 3.3). For these 1704 human proteins, we evaluate the prediction results of DC-*k*NN and the results of our method. Table 3 shows that our method significantly outperforms DC-*k*NN overall.

Hum-mPLoc 3.0 [26] is the state-of-the-art feature-based SCL predictor specifically for human proteins. It predicts SCLs based on the amino acid sequence of proteins through modeling the hidden

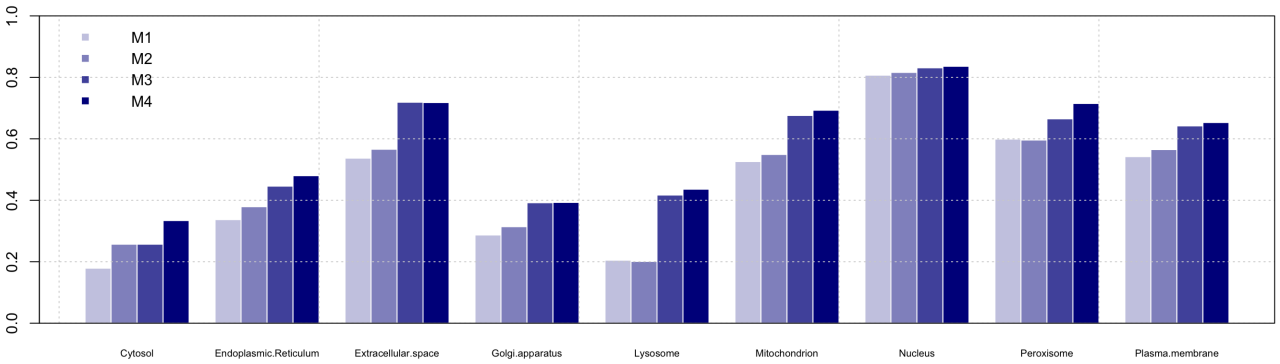

Figure 6: Comparison of four models for single label prediction.

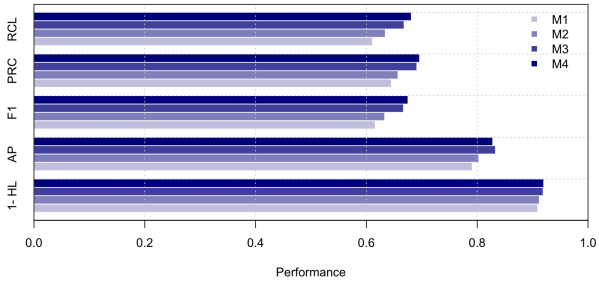

Figure 7: Performances of four models for multi-label classification.

correlations of gene ontology and functional domain features. The comparison of multi-SCL prediction results from Table 4 demonstrate that our method achieves better performance.

5 CONCLUSION AND FUTURE WORK

Protein subcellular localization prediction is an imbalanced multi-label classification problem. This paper proposes bayesian collective MRFs to predict multi-SCLs of human proteins. This is done by building the weighted MRFs based on the PPI network and then performing SCL label propagation to predict the SCLs of unknown proteins. We performed comprehensive experiments to evaluate the performance on human protein SCL datasets. The transductive learning from the re-balanced MLD proved to be more efficient to correctly assign SCLs. Owing to the collective MRFs which connect the binary MRFs by their spatial adjacency among SCLs, our method can achieve a higher performance for predicting the multi-SCLs comparing with the state-of-the-art methods of DC-kNN and HumPLoc 3.0.

Interestingly, neither the present approach nor the previous state-of-the-art method for SCL prediction perform as effectively for human as for other organisms (such as bacteria: precision > 0.95 and recall >0.93 for single-SCL prediction) [25]. One explanation could be that the cell structures of the bacteria (5 and 6 SCCs in total) are much simpler than mammalian cells. The the activities of human cells, such as the interactions among proteins and with other molecules, the translocation of proteins, the functions of proteins, and the biological environment of the cell are also more complicated. Therefore, there may still be room for improvement of the SCL prediction of human proteins.

All PPI data used in this study are static data reported from different studies and techniques with a huge diversity. During different biological processes, one protein can play different roles and functions, for instance by interacting with different target proteins. However, the available PPI datasets do not differentiate them according to the biological contexts. Since a single protein cannot physically interact with tens or hundreds of partners at the same time, this presents a future challenge: can we determine which interactions occur simultaneously and which are mutually exclusive? And how can we explore this knowledge to make context-specific SCL predictions?

ACKNOWLEDGMENTS

The authors would like to thank Dr. William Duddy for the English corrections. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. This work is funded by the International DFG Research Training Group GRK 1906/1.

REFERENCES

[1] Barry C. Arnold and David Strauss. 1991. Pseudolikelihood Estimation: Some Examples. *SankhyÄ: The Indian Journal of Statistics, Series B (1960-2002)* 53, 2 (1991), 233–243. http://www.jstor.org/stable/25052695

[2] Janos X Binder, Sune Pletscher-Frankild, Kalliopi Tsafou, Christian Stolte, Seán I O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. 2014. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database : the journal of biological databases and curation* 2014 (jan 2014), bau012. DOI: http://dx.doi.org/10.1093/database/bau012

[3] Torsten Blum, Sebastian Briesemeister, and Oliver Kohlbacher. 2009. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC bioinformatics* 10 (jan 2009), 274. DOI: http://dx.doi.org/10.1186/1471-2105-10-274

[4] Sebastian Briesemeister, Jörg Rahnenführer, and Oliver Kohlbacher. 2010. YLoc–an interpretable web server for predicting subcellular localization. *Nucleic acids research* 38, Web Server issue (jul 2010), W497–502. DOI: http://dx.doi.org/10.1093/nar/gkq477

[5] Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. 2013. mentha: a resource for browsing integrated protein-interaction networks. *Nature methods* 10, 8 (aug 2013), 690–1. DOI: http://dx.doi.org/10.1038/nmeth.2561

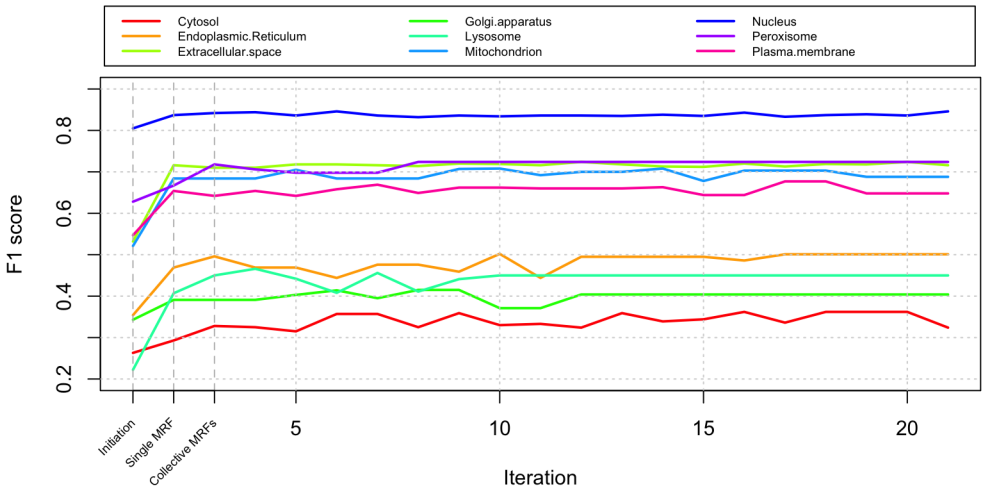

Figure 8: Performances of BCMRFs during iterations.

Table 3: Comparison with the method of DC-*k*NN - Multi-SCL prediction.

Method	Precision	Recall	F1 score	Average precision	Hamming loss
DC-*k*NN	0.502	0.472	0.474	0.672	0.119
BCMRFs	**0.674**	**0.621**	**0.633**	**0.899**	**0.073**

Table 4: Comparison with the method of Hum-mPLoc 3.0 - Multi-SCL prediction.

Method	Precision	Recall	F1 score	Average precision	Hamming loss
Hum-mPLoc 3.0	0.68	0.688	0.660	0.735	0.090
BCMRFs	**0.702**	0.67	**0.673**	**0.862**	**0.078**

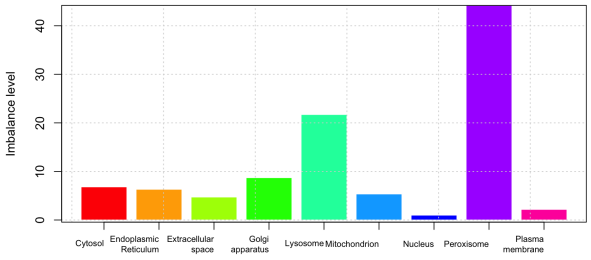

Figure 9: Imbalance level of each SCL class.

[6] Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. 2014. Concurrence among Imbalanced Labels and Its Influence on Multilabel Resampling Algorithms. Springer International Publishing, 110–121. DOI: http://dx.doi.org/10.1007/978-3-319-07617-1_10

[7] J. L. Gardy. 2003. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Research* 31, 13 (jul 2003), 3613–3617. DOI: http://dx.doi.org/10.1093/nar/gkg602

[8] Jennifer L Gardy and Fiona S L Brinkman. 2006. Methods for predicting bacterial protein subcellular localization. *Nature reviews. Microbiology* 4, 10 (oct 2006), 741–51. DOI: http://dx.doi.org/10.1038/nrmicro1494

[9] Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative Methods for Multi-labeled Classification. Springer Berlin Heidelberg, 22–30. DOI: http://dx.doi.org/10.1007/978-3-540-24775-3_5

[10] Xiaotong Guo, Fulin Liu, Ying Ju, Zhen Wang, and Chunyu Wang. 2016. Human Protein Subcellular Localization with Integrated Source and Multi-label Ensemble Classifier. *Scientific Reports* 6, February (2016), 28087. DOI: http://dx.doi.org/10.1038/srep28087

[11] Jianjun He, Hong Gu, and Wenqi Liu. 2012. Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites. *PLoS ONE* 7, 6 (2012). DOI: http://dx.doi.org/10.1371/journal.pone.0037155

[12] Mien-Chie Hung and Wolfgang Link. 2011. Protein localization in disease and therapy. *Journal of cell science* 124, Pt 20 (oct 2011), 3381–92. DOI: http://dx.doi.org/10.1242/jcs.089110

[13] Jonathan Q Jiang and Maoying Wu. 2012. Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study. *BMC bioinformatics* 13 Suppl 1, Suppl 10 (jan 2012), S20. DOI: http://dx.doi.org/10.1186/1471-2105-13-S10-S20

[14] T. S. Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S. Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C. J. Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K. Kashyap, Riaz Mohmood, Y. I. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. 2009. Human Protein Reference Database -

2009 update. *Nucleic Acids Research* 37, November 2008 (2009), 767–772. DOI: http://dx.doi.org/10.1093/nar/gkn892

[15] Pushmeet Kohli, M. Pawan Kumar, and Philip H. S. Torr. 2007. P3 & Beyond: Solving Energies with Higher Order Cliques. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8. DOI: http://dx.doi.org/10.1109/CVPR.2007.383204

[16] Yiannis A. I. Kourmpetis, Aalt D. J. van Dijk, Marco C. A. M. Bink, Roeland C. H. J. van Ham, and Cajo J. F. ter Braak. 2010. Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. *PLoS ONE* 5, 2 (2010), e9293. DOI: http://dx.doi.org/10.1371/journal.pone.0009293

[17] Kiyoung Lee, Han-Yu Chuang, Andreas Beyer, Min-Kyung Sung, Won-Ki Huh, Bonghee Lee, and Trey Ideker. 2008. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research* 36, 20 (nov 2008), e136. DOI: http://dx.doi.org/10.1093/nar/gkn619

[18] Noah Lee, Andrew F. Laine, and R. Theodore Smith. 2009. Bayesian Transductive Markov Random Fields for Interactive Segmentation in Retinal Disorders. Springer Berlin Heidelberg, 227–230. DOI: http://dx.doi.org/10.1007/978-3-642-03891-4_61

[19] Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona Brinkman, Gianni Cesareni, Andrew Chatr-aryamontri, Emilie Chautard, Carol Chen, Marine Dumousseau, Johannes Goll, Robert Hancock, Linda I Hannick, Igor Jurisica, Jyoti Khadake, David J Lynn, Usha Mahadevan, Livia Perfetto, Arathi Raghunath, Sylvie Ricard-Blum, Bernd Roechert, Lukasz Salwinski, Volker Stümpflen, Mike Tyers, Peter Uetz, Ioannis Xenarios, and Henning Hermjakob. 2012. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods* 9, 6 (2012), 626–626. DOI: http://dx.doi.org/10.1038/nmeth0612-626a

[20] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the Stratification of Multi-label Data. Springer Berlin Heidelberg, 145–158. DOI: http://dx.doi.org/10.1007/978-3-642-23808-6_10

[21] Chang Jin Shin, Simon Wong, Melissa J Davis, and Mark A Ragan. 2009. Protein-protein interaction as a predictor of subcellular location. *BMC systems biology* 3 (jan 2009), 28. DOI: http://dx.doi.org/10.1186/1752-0509-3-28

[22] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. 2010. Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology* 28, 12 (dec 2010), 1248–1250. DOI: http://dx.doi.org/10.1038/nbt1210-1248

[23] Shibiao Wan, Man-Wai Mak, and Sun-Yuan Kung. 2017. Transductive Learning for Multi-Label Protein Subchloroplast Localization Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14, 1 (jan 2017), 212–224. DOI: http://dx.doi.org/10.1109/TCBB.2016.2527657

[24] Zhi Wei and Hongzhe Li. 2007. A Markov random field model for network-based analysis of genomic data. *Bioinformatics* 23, 12 (2007), 1537–1544. DOI: http://dx.doi.org/10.1093/bioinformatics/btm129

[25] Nancy Y. Yu, James R. Wagner, Matthew R. Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S Cenk Sahinalp, Martin Ester, Leonard J. Foster, Fiona S L Brinkman, S. Cenk Sahinalp, Martin Ester, Leonard J. Foster, and Fiona S L Brinkman. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 13 (jul 2010), 1608–1615. DOI: http://dx.doi.org/10.1093/bioinformatics/btq249

[26] Hang Zhou, Yang Yang, and Hong-Bin Shen. 2016. Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics* (dec 2016), btw723. DOI: http://dx.doi.org/10.1093/bioinformatics/btw723