2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

New Feature Weighting Technique for Predicting Protein Subcellular Localization

Hisham Al-Mubaid and Duong B. Nguyen
University of Houston, Clear Lake, Houston, TX 77058, USA

Abstract— **Protein subcellular localization prediction is an important task with significant applications including the discovery of molecular functions of proteins. A number of prediction techniques have been developed in the past years based on protein sequence information. In this paper, we propose a new technique for predicting subcellular localizations of proteins using improved features of proteins extracted from protein sequences. The method is effective in inducing the features from protein sequences in multiple localizations. We evaluated the method using six datasets of proteins from bacteria, gram-negative and gram-positive, plant, and non-plant proteins and compared the results with recent methods. The evaluation results with six protein localization datasets showed that the method is promising and competitive for predicting protein localizations. This method is fairly effective in extracting strong features from protein sequences which will have significant impact on great deal of research work that relies mainly on protein sequence information for discovering molecular functions of proteins, drug design, and disease-protein associations.**

Keywords-feature weighting; protein localization;

I. INTRODUCTION

Protein subcellular localization is an important task for understanding the molecular functions and biological processes of proteins. It is also important for many highly significant applications especially for the discovery of novel protein and gene disease associations. Furthermore, it is necessary for drug discovery and understanding the disease mechanisms [1, 3, 6]. Many computational techniques have been proposed during the past decade for predicting protein subcellular localizations as the experimental lab techniques for this task are very laborious, time consuming, and expensive. In this paper, we propose a new technique for protein subcellular localization prediction based on calibrating feature weight of protein features extracted from protein sequences. The proposed technique is applied to protein features extracted from the basic protein sequences. When proteins are represented with their sequences, as *n-grams* for example, then subcellular localization prediction task can be manifested as an induction problem with a very high dimensional data. The proposed technique relies on the probability distribution of the class (of subcellular localizations) conditioned on the presence versus absence of the feature. This is basically the difference in the class probability after seeing the feature (refer to equation (1)). This process allows for calibrating the weight of each feature according to its contribution in each localization leading to improving the learning and prediction. We evaluated the method using six datasets of proteins from gram-negative and gram-positive bacteria, plant, and non-plant proteins. The proposed technique produced encouraging results and proved to be effective for inducing fairly strong protein features and encoding protein sequences as vectors of features. In the experimental results, the proposed method outperformed the baseline in all class pairs of all six datasets. We also compared the performance of the method with *tfidf* technique and with recently published results. This method is capable of extracting fairly strong features from sequence information which will have positive impact and can lead to enhancing the performance of numerous applications that extensively rely on protein sequence analysis.

II. RELATED WORK

Large volumes of protein sequence data are easily available due to the great progress of the sequencing techniques which encourages more development and investigations in this domain. Research studies have shown that proteins not localized in their normal cellular regions may lead to some diseases like Alzheimer [1]. Quite a few computational techniques had been proposed in the past several years for the prediction task of protein localization including machine learning methods such as support vector machines [1, 2, 11]. In [10], Tang et al (2013) divide the protein subcellular localization methods based on their employed features into three classes: −methods that are based on sequence features and mainly amino acid sequences; −methods based on gene ontology, textual information, or keyword features and, −methods based on feature combination from the first two classes. PlantLoc [10] is a web server for predicting the multiple-site subcellular localizations of plant proteins directly from protein sequences without using any machine learning or any annotation information. In [5], Yu et. al (2010) use SVM and Blast for in PSORTb 3.0 protein subcellular localization prediction [5]. The alignment method of PSORTb [5] was examined using proteins from

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.35

Swiss-Prot and produced a precision exceeding 97% with five-fold cross validation.

As a dimensionality reduction method, feature selection is a technique that attempts to reduce the number of features in the learning task by removing all unnecessary and redundant features in the features space [2]. In general feature reduction techniques can be roughly categorized into three types: Feature selection, feature clustering, and feature hashing; and feature selection is the most widely used among these three types [1, 2, 12, 17]. Feature selection and reduction techniques have been used extensively in many bioinformatics problems including gene selection and classification of microarray data, biomedical document clustering, prediction of gene and protein function, gene-protein name disambiguation, biomedical term disambiguation, biomedical WSD, and more [2, 12, 18]. Moreover, a number of natural language processing NLP applications have benefited directly from feature reduction including information retrieval and text categorization [19].

One of the commonly used feature selection techniques in the literature is the Mutual Information MI where features with the highest average MI are used for the learning and model induction [2]. Other techniques for feature reduction include principal component analysis PCA, latent semantics analysis, and latent Dirichlet [2]. King et. al (2012) proposed an n-gram Bayesian based method for prediction of subcellular localization called ngLOC [3]. ngLOC was examined with both prokaryotes and eukaryotes proteins with overall prediction accuracy lies in the range of 89.9% to 91.4% [3]. It was able to predict 11 distinct locations each in plant and animal species as well as 4 and 5 distinct locations in gram-positive and gram-negative bacterial datasets [3]. SherLoc2 is another method for protein localization prediction that combines several feature types including text features, sequence features, Gene Ontology GO features, and phylogenetic profiles [13]. It can predict 11 eukaryotic subcellular localizations [13]. In [14], Chi et. al (2012) proposed WegoLoc, which is based on weighted GO terms and sequence similarity [14]. Another method, iLoc-Gpos [7], uses multi-layer scale along with GO and sequential evolution information in the prediction of protein subcellular localizations [7].

III. METHODS AND TECHNIQUES

In protein subcellular localization, we would like to identify the subcellular localization of a given protein using its sequence. This can be done based on a set of features of this protein. In this paper we use *n-gram* features extracted from protein sequences with learning and prediction using Support Vector Machines SVM. The main contribution of this work is the new method for extracting strong features from protein sequences for improving the predictive power of subcellular localization techniques. The method can encode protein n-gram features for effective and accurate learning and prediction. The method calibrates feature weights based on the probability distribution of the class, which is one subcellular localization, conditioned on the presence versus absence of the features. This allows for calculating the difference in the class probability after seeing the feature. The features are collected from the amino acid sequences of proteins where each feature is an *n-gram* with *n=3, 4, .etc.*. Let $F=\{f_i\}_{i=1}^{m}$ be the set of all features in a given dataset. Each feature f_i is an *n-gram* from protein sequence. For a given feature f_i we calculate the feature weight $w(f_i)$ based on probability distribution of class C_1 conditioned on the presence and absence of the feature as follows:

$$w(f_i) = P(C_1|f_i) - P(C_1|\neg f_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

and this can also be written as:

$$w(f_i) = \frac{P(fi \mid C1) .|C1|}{P(fi \mid C1) .|C1|+P(fi \mid C2) .|C2|} - \frac{P(\neg fi \mid C1) .|C1|}{P(\neg fi \mid C1) .|C1| + P(\neg fi \mid C2) .|C2|} \quad \ldots.(2)$$

where $P(fi \mid C_1)$ is probability of occurrence of f_i conditioned on class C_1; and $P(\neg fi \mid C_1)$ is probability of absence of f_i in class C_1. Assuming we have n subcellular localizations; through supervised learning process we induce n predictors, one for each localization. In equation (1) we have

$$-1.0 \leq w(f_i) \leq 1.0.$$

This implies that a feature f_i tends to incline to class C_1 (resp. to class C_2) as its weight, $w(f_i)$, is approaching 1.0 (resp. approaching –1.0).

Feature vector:– The protein sequence data will be converted into numeric vectors wherein the n-grams are encoded in these vectors. Let X_i represent the numeric vector of protein sequence S_i such that

$$X_i = \{x_{ij}\}_{i=1..n}^{j=1..m}$$

where x_{ij} is the value of jth feature f_j in the vector X_i. With this, $P(C_1|f_j)$ can be depicted as:

$$P(C_1|f_j) = \frac{|\, X_i : X_i \in C_1 \; and \; f_j \in X_i \,|}{|\, X_i : f_j \in X_i \,|}$$

We use the feature count *count(i, j)* in the *baseline* method for un-weighted features where

count(i, j) = number of occurrence of feature f_j in the instance (sequence) S_i .....(3)

then, feature weights in the baseline vectors will be

$$x_{ij} = count(i,j) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

In the proposed method, we use:

$$x_{ij} = count(i,j).w(f_j) \ldots\ldots\ldots\ldots\ldots(5)$$

This proposed technique is suitable for subcellular localization task because it is mainly based on feature appearance $\{P(C_1|f_i)\}$ versus absence $\{P(C_1|\neg f_i)\}$ rather than continuous real valued features.

IV. EVALUATION AND EXPERIMENTS

In the evaluation, we used six datasets of protein subcellular localizations; see Table I. These proteins where experimentally classified into their subcellular localizations. We used LibSVM for learning and prediction with default linear kernel [20]. Within the context of this task, bacteria can be divided into two classes Gram-positive and Gram-negative based on the differences in the structure of their cell walls. In Gram stain test, the Gram positive bacterial are usually stained dark blue and retain the color whereas Gram negative bacteria cannot retain the stain and appear red or pink [7].

Datasets: the six datasets are summarized in Table I. The first dataset, *D1*, includes 1444 proteins divided into five classes. Proteins in dataset D1 are Gram negative bacteria proteins obtained from PSORTb [8, 21]. The second dataset, *D2*, contains Gram positive proteins and obtained from [7, 22]. D2 includes four main classes: cell membrane, cell wall, cytoplasm and extracellular; and was used in in the evaluation of *iLoc-Gpos* [7]. Note: after removing the repeated proteins (4 proteins) and proteins with two locations (4 proteins) the total becomes 515 proteins. The third dataset, *D3*, contains plant proteins in four different localization sites [4]. The set includes 852 proteins in the four categories; see Table I. The original dataset contains 882 proteins distributed among 13 localizations. We used the largest four categories: Cytoplasmic, Integral membrane, Secretory, and Attached to the membrane by a lipid anchor (we call it *membrane attached*). Dataset D4 includes 1931 proteins from four subcellular localizations obtained from [3]. Datasets D5 & D6 are taken from [2, 9]. D5 contains 940 proteins belonging to 4 localizations while D6 includes 2738 non-plan proteins belonging to 3 subcellular locations.

TABLE I. THE SIX DATASETS USED IN THE EVALUATION

Dataset	Protein types	Number of subcellular localizations	Number of proteins in each subcellular localization	Total number of proteins in the dataset
D1	*bacteria gram negative*	5	Cytoplasmic *CY* 278 Cytoplasmic Membrane *CM* 309 Periplasmic *PP* 276 Outer Membrane *OM* 391 Extracellular *EC* 190	1444
D2	*bacteria gram positive*	4	Cell membrane 174 Cell wall 18 Cytoplasm 208 Extra Cellular 123	523
D3	*plant proteins*	4	Cytoplasmic CY340 Integral membrane IM 402 Secretory SE 50 Membrane attached MA 60	852
D4	*plant proteins*	4	Cytoplasm C 481 Mitochondria M 469 Nuclear N 630 Plasma Membrane P 351	1931
D5	*plant proteins*	4	chloroplast CPT 141 Mitochondrial MTC 368 Secretory pathway/signal peptide SP 269 other (nuclear 54, cytosolic 108) UC 162	940
D6	*non-plant proteins*	3	Mitochondrial MTP 361 Secretory pathway/signal peptide SP 715 Others (nuclear and cytosolic) UC 1662	2738

For evaluation metric, we use the *prediction accuracy* which is the ratio of correctly predicted instances to the total number of instances. We also used *Area Under ROC curve* (AUC) results for evaluation and comparison. All the experiments are conducted with *n-gram* features with *n=3* and using 10 fold cross validation 10-FCV. The evaluation results are summarized in Table II in terms of prediction accuracy and illustrated in Figure 1. In all datasets, the proposed method outperformed the baseline, and with significant improvement over the baseline in most of the datasets as shown in Table II. We did not include the detailed results for all classes of all datasets for the interest of space; we report only the detailed results of dataset D6 in Table III since it is the largest dataset (2738 proteins). The proposed method outperformed the baseline in the three class pairs in prediction accuracy and AUC (Table III). We also report the detailed results of dataset D5, in Table VI, for the purpose of demonstrating the performance of *tfidf* vs baseline using both prediction accuracy and AUC of baseline, *tfidf*, and the proposed method.

V. DISCUSSION AND CONCLUSION

To compare with prior similar techniques, we compare and report the performance results (prediction accuracy) from published results for datasets D1, D3, D5, D6 in Table IV. As shown in Table IV, the proposed method outperforms these similar techniques in all experiments. In the published results using plant proteins D5 (940 proteins) as shown in Table IV our method outperforms the similar techniques [2, 9]. Also, in [9], Tamura and Akutsu (2007) reported other techniques with accuracy ranging between 84.8% to 88.1%.

TABLE II. THE PERFORMANCE RESULTS OF ALL DATASETS

Dataset	Number of s.c. localiza-tions	Total number of proteins in the dataset	Prediction performance (accuracy)		AUC	
			baseline	proposed	baseline	Proposed
D1	5	1444	0.929	0.959	0.925	0.958
D2	4	523	0.869	0.920	0.707	0.776
D3	4	852	0.983	0.990	0.950	0.951
D4	4	1931	0.920	0.933	0.920	0.932
D5	4	940	0.874	0.951	0.839	0.938
D6	3	2738	0.875	0.903	0.815	0.852
		Overall	0.908	**0.943**	0.859	**0.901**

TABLE III. DETALIED PERFORMANCE RESULTS OF D6

Class pair	Accuracy %		AUC %	
	Baseline	Proposed	Baseline	Proposed
MTP – SP	88.86	93.37%	86.81	92.18
MTP – UC	87.34	90.31%	74.48	78.18
SP – UC	86.39	87.33%	83.19	85.09
Overall	87.53	**90.34%**	81.49	**85.15**

Effectively, feature weighting is more comprehensive and more powerful than feature selection as in feature weighting we can assign the weight 0 to unselect features. As mentioned earlier, most of the research in this domain is focused on feature selection rather than feature weighting [23, 24]. The most common and widely used feature weighting technique is the *tfidf* which can be stated as:

$$x_{ij} = tf(i,j).idf(f_j) \text{.....(6)}$$

$$tf(i,j) = count(i,j) \text{.....(7)}$$

$$idf(f_j) = log(N/df_j) \text{.....(8)}$$

and df_j is basically number of instances X_i in which feature f_j occurs.

Using SVM, which is a state-of-the-art learner, *tfidf* did not produce any significant improvement over the baseline feature count (equation (4)); see Table VI (*in fact it produced lower performance sometimes*). Therefore, we used feature count as shown in equation (4) as the *baseline* to compare our proposed technique. Table V contains the prediction accuracy of dataset D5 using baseline (equation (4)), *tfidf*, and the proposed technique. In this case, baseline produced slightly better overall performance (*87.45%*) than *tfidf* (*87.21%*) and proposed method gave the highest performance (*95.12%*) in all class pairs.

In conclusion: we proposed and evaluated a prediction technique for protein subcellular localization. The proposed technique is effective in calibrating feature weights of protein features extracted from protein sequences; and converting protein features into vector for learning and prediction. We presented the experimental results of the technique on six protein localization datasets from prokaryotes and eukaryotes. We showed that the method outperformed the baseline technique on all datasets. We also showed that it outperformed recently published techniques (Table IV) on the same proteins evaluated in this research. With its competitive and impressive experimental results with six protein datasets, the method will have significant contribution into advancing great deal research work that relies extensively on protein and gene sequence for various applications like discovering molecular functions of a protein.

TABLE IV. COMPARISON WITH SIMILAR METHODS

Data set	Method	Accuracy %
D1	Feature hashing (Ref [2])	86.00
	Feature selection (Ref [2])	86.00
	PLSpred (Ref [6])	91.20
	Proposed (*this paper*)	**95.93**
D3	(Rshpid et al. 2007; ref[4]) 852 proteins, 4 s.c. localizations	89.00
	Proposed (*this paper*) 852 proteins, 4 s.c. localizations	**98.97**
D5	(Caragea et al. 2007; ref[2]) 940 proteins, 4 s.c. localizations	78.40
	(Tamura and Akutsu (2007); ref [9]) 940 proteins, 4 s.c. localizations	90.96
	Proposed (*this paper*) 940 proteins, 4 s.c. localizations	**95.12**
D6	(Caragea et al. 2007; ref[2]) 2738 proteins, 3 s.c. localizations	82.83
	Proposed (*this paper*) 2738 proteins, 3 s.c. localizations	**90.34**

TABLE V. PERFORMANCE RESULTS USING D5

Class	Accuracy %		
	Baseline	tfidf	Proposed
CPT_MTC	88.60	86.25	97.84
CPT_SP	90.00	90.98	95.61
CPT_UC	81.19	79.87	97.03
MTC_SP	93.88	94.35	96.39
MTC_UC	82.83	83.40	92.64
SP_UC	88.17	88.40	91.18
Overall	87.45	87.21	**95.12**
Class	**AUC %**		
	Baseline	**tfidf**	**Proposed**
CPT_MTC	81.18	76.27	96.76
CPT_SP	87.65	88.74	94.63
CPT_UC	81.03	79.56	97.04
MTC_SP	93.35	93.61	95.97
MTC_UC	74.68	74.74	89.17
SP_UC	85.86	86.04	89.38
Overall	83.96	83.16	**93.83**

REFERENCES

[1] Y. Hu, T. Li, J. Sun, S. Tang, W. Xiong, D. Li, G. Chen, and P. Cong. Predicting Gram-positive bacterial protein subcellular localization based on localization motifs. Journal of Theoretical Biology 308 (2012) 135–140.

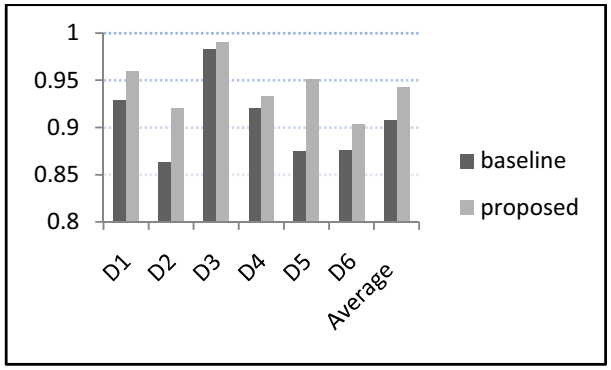

Figure 1. Accuracy of the baseline and proposed method on six datasets

[2] C. Caragea, A. Silvescu, P. Mitra. Protein Sequence Classification Using Feature Hashing. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2011.

[3] King BR, Vural S, Pandey S, Barteau A, Guda C. ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. BMC Research Notes (2012) 5:351.

[4] Rashid M, Saha S, Raghava GPS, Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics, (2007), 8:337.

[5] N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics, (2010), 26 (13):1608-1615

[6] Bhasin, Manoj, Garg, Aarti and Raghava, G. P. S. PSLpred: prediction of subcellular localization of bacterial proteins, Bioinformatics 21 , no. 10 (2005): 2522-2524.

[7] Wu, Zhi-Cheng, Xuan Xiao, and Kuo-Chen Chou. iLoc-Gpos: A multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins, Protein and Peptide Letters (2012): 19.1, 4.

[8] Gardy JL, Laird MR, Chen F, Rey S, et al. PSORTb v2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics. (2005);21:617–623.

[9] T. Tamura and T. Akutsu (2007). Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. BMC Bioinformatics 2007, 8:466, 2007

[10] PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif, Nucleic Acids Research, 2013, 1–7

[11] Bhasin, M.,Raghava,G.P.,2004.ESLpred:SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic AcidsRes.32,W414-9.

[12] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. Bioinformatics, Vol. 23 no.19, 2007.

[13] S. Briesemeister, T .Blum, S. Brady, Y. Lam, O. Kohlbacher, and H. Shatkay. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. Journal of proteome research 2009, 8:5363–6.

[14] Chi S-M, Nam D: WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms. Bioinformatics 2012, 28:1028–1030.

[15] H. Al-Mubaid and S.A. Umair. A New Text Categorization Technique Using Distributional Clustering and Learning Logic. IEEE Trans on Knowledge and Data Eng. vol.18, no. 9, 2006.

[16] V. Vapnik, The Nature of Statistical Learning Theory. Springer, 1995

[17] H. Al-Mubaid, “Data mining and analysis of material data using feature clustering for radiation shielding”, proc. of Int’l conf on Comp. and App. in Industry and Eng. CAINE-2010, Nov. 2010.

[18] H. Al-Mubaid and S. Gungu. A Learning Based Approach for Biomedical Word Sense Disambiguation. TSWJ journal, vol. 2012, PMID: 22666174; 2012.

[19] H. Al-Mubaid. A Learning-Classification Based Approach for Word Prediction. International Arab Journal on Information Technology IAJIT, Vol.4 No.3, July 2007.

[20] LibSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[21] PSORTb http://www.psort.org/dataset/datasetv2.html

[22] Gram-positive bacterial proteins with single and multiple sites:iLoc-Gpos http://www.jci-bioinfo.cn/iLoc-Gpos/data

[23] N.A. Zaidi, J. Cerquides, M.J. Carman, and G.I. Webb. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. JMLR 14 (2013) 1947-1988.

[24] J. Wu and Z. Cai. Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB). Journal of Computational Information Systems 7:5 (2011) 1672-1679.