The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009
Copyright © 2009 ORSC & APORC, pp. 254–260

Protein Subcellular Localization Prediction for *Fusarium graminearum*[*]

Chenglei Sun[1,2] Wei-Hua Tang[3] Luonan Chen[2,4]
Xing-Ming Zhao[2,3,†]

[1]Department of Mathematics, Shanghai University, China, 200444
[2]Institute of Systems Biology, Shanghai University, China, 200444
[3]Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
[4]Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan

Abstract The fungal pathogen *Fusarium graminearum* (telomorph Gibberella zeae) is the causal agent of several destructive crop diseases. Investigating subcellular localizations of *F. graminearum* proteins can provide insight into pathogenic mechanisms underlying *F. graminearum*-host interactions. In this paper, we design a novel balanced ensemble classifier based on support vector machines (SVMs) to predict *F. graminearum* proteins' subcellular localization from the primary sequence. The method is performed with a fungi dataset collected from UniProtKB database. In addition, we utilize SCL-BLAST (SubCellular Localization BLAST) to transfer annotations of homologous proteins to the target uncharacterized protein. We make three fold contributions to this filed. First, we present a new algorithm to cope with imbalance problem that arises in protein subcellular localization prediction, which can improve prediction accuracy significantly. Second, we employ feature selection techniques to find out most informative features for each compartment, and reduce computation cost and improve prediction accuracy at the same time. Third, we use BLAST to complement SVMs based methods, which makes our prediction more effective.

Keywords *Fusarium graminearum*; protein subcellular localization; re-balanced classifier

1 Introduction

F. graminearum, which cause destructive disease-*Fusarium* head blight (FHB) on wheat and barley, is a leading cause of economical loss in these crops [1]. It is estimated that *F. graminearum* causes economical losses of $3 billion in the US between 1991 and 1996 [2]. In addition, the fungus contaminates grain with toxic metabolites that are a threat to human health [3]. Therefore, it is necessary to investigate the mechanism underlying the pathogenic process of this destructive fungus, in order to facilitate the searching

[*]This work was partly supported by Innovation Funding of Shanghai University, Key Project of Shanghai Education Committee (09ZZ93), Open Funding of National Key Laboratory of Plant Molecular Genetics, SRF for ROCS, SEM.
[†]Corresponding author

for an efficient way to control it. Although some pathogenicity genes have been identified for this fungus, such as pathogenicity genes deposited in PHI-base [4], the molecular mechanisms that *F. graminearum* overcomes plant defense barriers and causes a disease is still largely unknown. Generally, proteins are transported to specific compartments in a cell to function properly. These subcellular localizations therefore provide insights into protein functions help to understand the pathogenic process of this destructive fungus. Although the whole genome of *F. graminearum* has been sequenced and partly annotated [5], there are no subcellular localization annotation available for *F. graminearum* right now.

In this work, we present a framework to predict subcellular localizations for *F. graminearum* proteins. Recently, various machine learning methods have bee developed for protein subcellular localization prediction, such as *k*-nearest neighbors(kNN) [6, 7, 8], artificial neural networks(ANNs) [9, 10, 11], support vector machines(SVMs) [12, 13, 14], and Bayesian networks[15, 16, 17]. Furthermore, many different types of features have been used for subcellular localization prediction. One popular description of protein is amino acid composition including acid pair composition (PAA) and gapped amino acid composition (GapAA) owing to its simplicity and effectiveness. In this work, SVMs and amino acid composition are used for prediction in sequel.

Despite high prediction accuracy on selected datasets, most existing methods have some limitations. Subcellular localization prediction is actually a classification problem from perspective of machine learning, where the proteins do not belong to target compartment are usually treated as negative samples. Therefore, the number of negative samples is generally much larger than that of positive samples, which leads to imbalance problem and degrade performance of classifier [18]. Under the circumstances, we present a new algorithm to cope with imbalance problem existing in protein subcellular localization prediction, which can improve prediction accuracy significantly. Furthermore, since there are a large number of features extracted for each protein in learning procedure, which leads to 'bottleneck of dimensionality' and the noise in the data will degrade performance, we employ feature selection techniques to find out most informative features for each subcellular localization, and reduce computation cost and improve prediction accuracy at the same time. In addition, we use BLAST to complement SVMs based methods, which makes our prediction more effective. The results on validation datasets demonstrate the efficiency and effectiveness of the proposed method.

2 Materials and methods

2.1 Data sets

The annotation of protein subcellular localization for fungi were downloaded from the UniProtKB database release 57.1 and used as training dataset. The number of proteins in the dataset is 23,228, of which 17,769 is annotated. After discarding those subcellular location annotations followed 'By similarity', 'Potential' and 'Probable', 10,554 proteins remain and are used by BLAST as reference dataset. Proteins for the following 9 subcellular localizations were retrieved to build the dataset: Extracellular, Cytoplasm, Nucleus, Mitochondria, Endoplasmic reticulum, Golgi apparatus, Peroxisome, Endosome and Vacuole. Furthermore, proteins localized in more than one subcellular compartment were removed, those with less than 50 amino acids in length were removed, and those with

ambiguous amino acids (B, X and Z) were also removed from the dataset. In addition, CD-HIT program was used to remove the homology bias in the dataset with a threshold identity of 40%. Finally, we got a non-redundant dataset of 4,496 proteins as training dataset for SVMs. Other subcellular localizations have been excluded because too few (less than 30) non-redundant representatives remain and are not enough for training a classifier. Table 3 shows the statistics for nine compartments used to train SVMs.

Table 1: Fungi proteins used to train SVMs.

Subcellular location	Proteins in UniProtKB[a]	Proteins_40[b]
Extracellular	286	159
Cytoplasm	1388	941
Nucleus	1540	1356
Mitochondrion	1721	951
Endoplasmic reticulum	904	572
Golgi apparatus	291	163
Peroxisome	120	82
Endosome	169	74
Vacuole	319	198
Total	6738	4496

[a] Number of proteins with unique localization found in UniProtKB.
[b] Curated dataset with pairwise sequence identity <40%.

2.2 Feature extraction and selection

In machine learning, each protein should be represented as a feature vector. The amino acid triplets (threAA) are considered here. Each protein vector is generated consisting of frequency of all possible combinations of three amino acids from 20-amino acid alphabet. Therefore, each protein contains 8,000 (20^3) features. In addition, to reduce effect of protein sequence length, each feature value is normalized as following:

$$V_{ij} = \frac{V_{ij}}{\max\{V_{ij} | j \in \{1,...,m\}\}} \tag{1}$$

where V_{ij} is the value for feature j in vector i, where $j \in \{1,...,m\}$.

The representations of each protein described above have 8,000 features, which leads to high computation cost and the noise in the data generally degrade performance of classifiers. To find out informative features and reduce computation cost, we first utilized t-test to rank the features and then employ sequential forward feature selection to select the most informative features starting from the top ranked features by t-test. The obtained feature set is used in sequel.

2.3 Re-balancing imbalanced dataset

After getting the feature vectors for protein sequences, one classifier can be designed for each protein class, and the new protein sequence can be classified into the class with the biggest decision value. However, as described previously, the imbalanced problem will arise in this case. To overcome this problem, a bagging-like re-balanced classifier is presented in this section. Figure 1 presents the schematic flowchart of the proposed

method for re-balancing the imbalanced dataset. In our work, the number of negative data is always larger than that of positive data, and the negative dataset is first under sampled and divided into m subsets, where each subset has similar size as the positive data set. After the sampling procedure, we get m training sets, where each training set consists of one subset from the negative data and the positive data, i.e. {negative subset 1, positive set}, ..., {negative subset m, positive set}. With the newly generated data sets, we train m classifiers with one for each training set. Given a new test example, the prediction results are obtained by fusing the outputs from the m classifiers.Then the results are combined by a voting scheme.

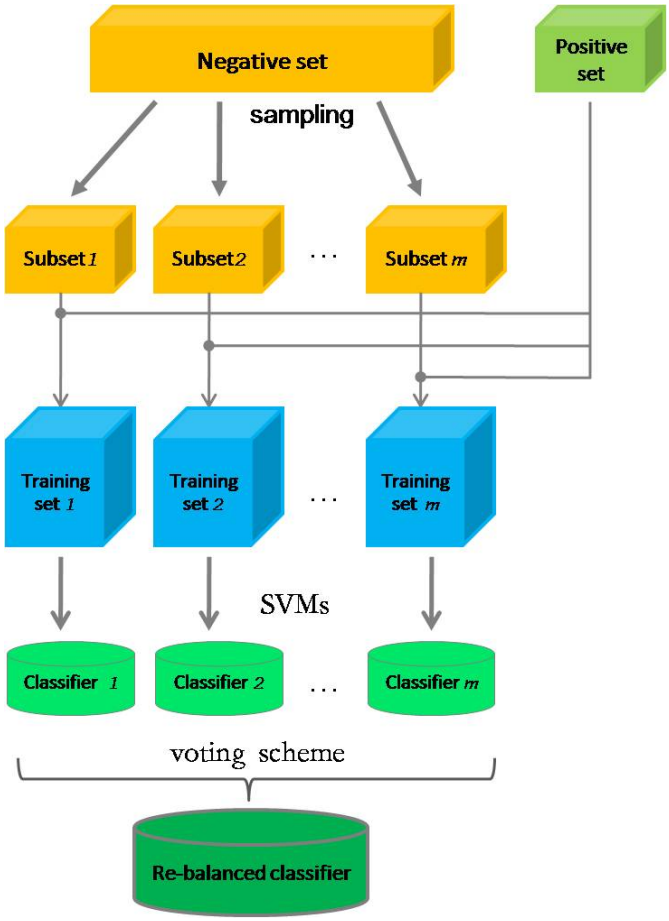

Figure 1: The architecture of Re-balanced classifier.

2.4 SCL-BLAST

Subcellular localization tends to be evolutionarily conserved, and the homologues with localization annotation appears to be a good indicator of the target protein. We there-

fore use SCL-BLAST (for SubCellular Localization BLAST) [16], in which a BLAST search of a submitted protein is carried out against our database of 10,554 proteins with known localization using an E-value cutoff of 1e-10. Then we get the subcellular location of homologous proteins for our target protein, which can perfect our prediction because all of the probable localization site are returned.

3 Results

3.1 Cross-validation

To see the performance of the proposed method, we evaluated it using 10-fold cross-validation (CV). The classifier used here is SVMs, and gaussian kernel was employed for SVMs and the parameters were optimized in CV procedure. To evaluate the performance of different methods, accuracy and AUC score (area under ROC curve) were employed in this work. Table 2 lists the results. Next, we investigated the effect of balancing and feature selections on performance of classifier. Figure 2 shows the comparison of performance of SVMs classifier without feature selection data against that with feature selection and balancing. The results on nine compartments demonstrate the efficiency and effectiveness of the proposed method.

Table 2: The results on 10-fold cross-validation.

Subcellular location	Accuracy	AUC
Extracellular	0.92	0.92
Cytoplasm	0.70	0.66
Nucleus	0.75	0.70
Mitochondrion	0.75	0.72
Endoplasmic reticulum	0.81	0.81
Golgi apparatus	0.78	0.76
Peroxisome	0.87	0.82
Endosome	0.91	0.92
Vacuole	0.75	0.70

3.2 Predicting protein subcellular localization for *F. graminearum*

After get the training dataset and trained classifier, we aim to predict protein subcellular localizations of *F. graminearum*. The prediction results for 13,321 *F. graminearum* proteins are showed in Table 3 and Table 4 respectively. From the results, we can see that our method can predict subcellular localizations of most *F. graminearum* proteins. Although these predictions are not verified in lab, we believe that the predictions can provide guidelines for future experiments and help to understand this destructive fungus.

4 Conclusions

In this work, we present a framework to predict subcellular localizations for *F. graminearum* based on protein primary structures. A new balanced classifier is presented for predictions, where no homology information can be used, i.e. sequence identity below 40% in the dataset. Furthermore, SCL-BLAST is utilized to predict subcellular localizations of *F. graminearum* proteins in the case that homology information available in the

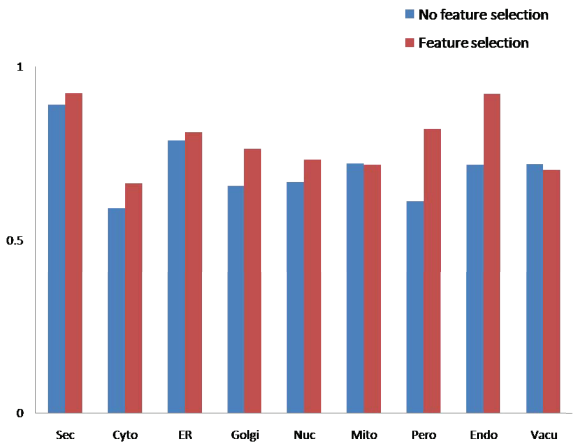

Figure 2: The comparison of performance of SVMs classifier without feature selection data against that with feature selection and balancing.

Table 3: Distribution of predicted subcellular localization for FG based on SVMs.

Subcellular location	Predictions	Subcellular location	Predictions
Extracellular	245	Golgi apparatus	970
Cytoplasm	2504	Peroxisome	1059
Nucleus	953	Endosome	544
Mitochondrion	2306	Vacuole	862
Endoplasmic reticulum	1555	Total	7910

Table 4: Distribution of predicted subcellular localization for FG based on *SCL-BLAST*.

Subcellular location	Predictions	Subcellular location	Predictions
Secreted	262	Peroxisome	154
Bud	11	Multi-pass membrane protein	968
Bud neck	36	Single-pass membrane protein	229
Bud tip	6	Golgi apparatus	246
Cell membrane	346	Endoplasmic reticulum	520
Lipid-anchor	61	Endosome	52
Cytoplasm	2050	Peripheral membrane protein	280
Prospore membrane.	4	Vacuole	61
Nucleus	1858	Vacuole membrane	203
Centromere	23	cytoskeleton	88
Kinetochore	28	Spindle	48
Telomere	19	Preautophagosomal structure membrane	4
Mitochondrion	952	Total	4897

We consider 25 categories, of which some are not the subcellular location, but appear in the UniProtKB protein annotation.

reference dataset. The two methods complement each other and therefore make the predictions more effective. We believe that the predictions can provide guidelines for future experiments and help to understand this destructive fungus. We designed this new method aiming at solving the special biological problem. To show its generalized prediction ability, we will compare it with some existing methods in the near future.

References

[1] Goswami RS, Kistler HC. Heading for disaster: Fusarium graminearum on cereal crops. Molecular Plant Pathol 2004, 5:515.

[2] Priest FG, Campbell I. Brewing Microbiology, volume 3. Springer, 2002.

[3] Bennett JW, Klich M. Mycotoxins. Clinical Microbiology Reviews 2003, 16:497-516.

[4] Winnenburg R, Baldwin TK, et al. PHI-base: a new database for pathogen host interactions. Nucleic Acids Res 2006, 34(suppl 1):D459-464.

[5] Christina AC, et al. The Fusarium graminearum Genome Reveals a Link Between Localized Polymorphism and Pathogen Specialization. Science 2007, 317:1400-1402.

[6] Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 1999, 24:34-35.

[7] Huang Y, Li YD. Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 2004, 20:21-28.

[8] Lee KY, Chuang HY, Beyer A, Sung MK, Huh WK, Lee B, Ideker T. Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. Nucleic Acids Res 2008, 10:1-13.

[9] Nair R, Rost B. Better prediction of sub-cellular localization by combining evolutionary and structural information. Proteins 2003, 53:917-930.

[10] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998, 26:2230-2236.

[11] Emanuelsson O, Nielsen H, Brunak S, Heijne G. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. Journal of Molecular Biology 2000, 300:1005-1016.

[12] Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 2001, 43:246-255.

[13] Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 2003, 19:1656-1663.

[14] Chang JM, et al. PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. Proteins 2008, 72:693-710.

[15] Scott MS, Thomas DY, Hallett MT. Prediction subcellular localization via protein motif co-occurrence. Genome Research 2004, 14:1957-1966.

[16] Gardy JL, et al. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 2003, 31:3613-3617.

[17] Garga P, Sharmaa V, Chaudharia P and Roy N. SubCellProt: Predicting Protein Subcellular Localization Using Machine Learning Approaches. In Silico Biology 2009, 9:35-44.

[18] Zhao XM, Li X, Chen L, Aihara K. Protein classification with imbalanced data. Proteins 2008, 4:1125-1132.