Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Feature Selection and Granular SVM Classification for Protein Arginine Methylation Identification

Zejin Ding, Yan-Qing Zhang
Department of Computer Science,
Georgia State University,
Atlanta, GA, 30303, USA
{zding, yzhang}@cs.gsu.edu

Yujun George Zheng
Department of Chemistry,
Georgia State University,
Atlanta, GA, 30303, USA
yzheng@langate.gsu.edu

***Abstract*— Protein methylation modification has been discovered for half a century but still far less been studied than other modifications. Computational analysis is recently introduced to discover other unknown methylation sites based on few known ones. To effectively predict possible methylation, sophisticated classification strategy should be well devised. In this paper, we first extracted informative features from methylated fragments in many protein sequences, including the physicochemical properties, secondary structure information, evolutionary profiles, and solvent accessibility of surrounding residues. Then, an efficient feature selection method (mRMR) is applied to eliminate redundant features but keep important ones. Since methylated residues are far less than non-methylated, the collected data is relatively imbalanced. Thus, we propose to use the granular support vector machine (GSVM) which is specially designed for imbalanced classification problems. A 7-fold cross validation shows that our strategy generates comparable predication accuracy with many current methods or even better. Meanwhile, our method provides insights to identify the underlying mechanisms of protein methylation.**

***Keywords*—Protein Methylation, Imbalanced Data Mining, Granular Support Vector Machines (GSVM), Methylation Prediction, Feature Selction**

I. INTRODUCTION

Protein arginine methylation is a type of important post-translational modifications (PTMs) catalyzed by protein arginine methyltransferases (PRMTs) [1]. PRMTs mediate multiple cellular processes including transcriptional regulation, DNA repair, and signal transduction. The η-nitrogens of an arginine residue in a protein can be monomethylated or dimethylated, with either both methyl groups on one terminal nitrogen (asymmetric dimethylated arginine) or on either nitrogens (symmetric dimethylated arginine) by PRMTs. Knowing the protein substrates of a PRMT is a critical step to understand its function in vivo. However, it is experimentally challenging to identify all the methylation proteins in human, needless to mention the methylation sites in individual protein substrates, which demands tremendous time and efforts.

Fortunately, intelligent computational analysis has been recently introduced to help finding top potential candidate methylation residues for real experiments, so as to reduce cost and speed up the discovering process. So far, a limited number of papers have been published for predicting new methylated residues on unknown protein sequences. All these papers only extract the primary sequence information from PRMT substrates and then use machine learning methods to differentiate methylated and non-methylated residues [2-4]. Meanwhile, these methods are adopting the support vector machines (SVMs) to build the classifiers, due to its better generalizability and higher accuracy comparing to other classifiers, such as neural networks, and decision trees. However, all these methods either only extract basic information from primary protein sequence or did not perform any refinement on extracted features.

In this paper, we try to analyze several important factors which possibly determine the status of Arginine being methylated or non-methylated. More specifically, we first extract as many features from the sequences of known PRMTs as possible, such as physicochemical properties, secondary structures, solvent accessibility, and evolutionary profiles. In other words, not only the primary sequence information will be considered, the secondary structure and partial third structure information is also examined. Then, these features are filtered by a feature selection algorithm—minimum redundancy, maximum relevance (mRMR) [5]—to discover the very important features. Due to the fact that the number of methylated Arginine is far less than non-methylated ones, the resulting datasets are fairly imbalanced. Thus, the specially designed Granular Support Vector Machine (GSVM) [6] is performed to tackle this problem and produce both high sensitivity and specificity for classification. More clearly, the GSVM-repetitive undersampling (GSVM-RU) method is applied to our methylation dataset.

The organization of this paper is as follows. Section II introduces the background of protein methylation and several related works. Section III introduces our proposed method, including feature extraction, feature selection and granular SVM classification steps. Section IV elaborates the dataset collecting, experiment setting, and computational results with different parameters. Several metrics are used to measure the classification results with the GSVM-RU algorithm, which shows better performance. Finally, Section V discusses our study again and present some future works.

Zejin Ding is supported by Molecular Basis of Disease Program at Georgia State University, Atlanta.

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

II. Background and Related Works

A. Protein Arginine Methylation

There are mainly three forms of arginine methylation: mono-methylated, asymmetric dimethylated, and symmetric dimethylated arginines. Arginine methylation process is carried out by the special protein group called protein arginine methyltransferases (PRMTs). Eleven types of PRMTs have been identified, which forms two categories: Type I PRMTs which produce asymmetric dimethyl-arginines, and Type II PRMTs which produce symmetric dimethylarginines [13].

Type I enzymes (PRMT1, 3, 4, 6 and 8) catalyze the transfer of the methyl group from S-adenosyl-L-methionine (SAM, AdoMet) to the guanidino nitrogen atoms of arginine residue to produce ω-NG monomethylarginines (MMA, L-NMMA) and ω-NG,NG-asymmetric dimethylarginines (ADMA) [14]. Type II enzymes (PRMT 5, 7 and 9) catalyze the formation of MMA and ω-NG, N'G-symmetric dimethylarginines (SDMA) [13]. Of note, the enzymatic activity of PRMT2, 10 and 11 remains uncharacterized.

B. Related Works

Plewczynski et al. published a methylation prediction tool within their AutoMotif Server [2]. They first find methylated 9-amino acid long sequence fragments (positive) from the proteins in SWISS-PORT database which are known with methylation modifications. Then, the negative instances are created by randomly choosing fragments without methylation. Then all instances are projected to a high dimensional space with different encoding schemes to feed to SVM classifier. Several amino acid representation methods are used, including orthogonal vectors, BLOSUM62, and position-specific amino acid preference, etc. However, their method ended up with a relative lower accuracy (recall: 36.28%, precision: 78.00%).

Daily et al. [3] designed another methylation prediction method by utilizing the biological hypothesis that methylation sites prefer intrinsically disordered regions in proteins. They also use SVM to build classifiers, but with an expanded set of features, such as aromaticity, net charge, hydrophobic moment, sequence complexity and beta entropy, and evolutionary profiles (Position specific scoring matrices, PSSM). This method reaches 73.6% in sensitivity and 82.2% in specificity for Arginine methylation prediction.

Another latest methylation server named MEMO also used the SVM for prediction [4]. Chen et al. [4] manually collected methylation sites from SWISS-PORT database and PubMed publication and refined them by removing homologous proteins using BLASTCLUST. Their method produced a high accuracy for Arginine prediction: 69.6% in sensitivity, 89.2% in specificity, and 86.7% in accuracy.

Most recently, Shien et al. [12] combines more structural information of the sequences to identify methylation sites. They first collect methylated lysine, arginine and other residues from MEMO and newest SWISS-PORT database (Version 53). Then each methylated fragment was encoded with their amino acid characteristics, predicted secondary structures (by PSIPRED), and predicted area accessible ability (by RVP-Net). They also use SVM as the learning classifiers. Since more methylated residues have been collected and more information has been used for representation, their method produces the highest prediction accuracy compared with other existing methods. The sensitivity they got is 82.1% and the specificity is 87.4%. Similar to previous methods, their server also only provide a general methylation prediction, and the related PRMTs are not identified.

III. Methods

Similar to other methods, we also assume that methylated Arginines are largely determined by its neighboring residues in primary sequence structures. That is, the characteristics of residues centered with an Arginine will be explored as deep and diversified as possible to find the potential relations. Each status of an Arginine (being methylated or non-methylated) is represented by a vector of features extracted from its primary neighboring residues. These features include physicochemical properties (PhCh), the multiple sequence alignment profiles (the PSSM), secondary structure (SS), and solvent accessible area (SAS). Thus, our first step is to extract these features for each Arginine in the PRMTs. Then, a feature selection is involved to discard less irrelevant and redundant features; in our case, the minimum redundancy and maximum relevance (mRMR) method is deployed. Finally, the resulting dataset from a selected feature subset is fed to the machine learning for classification and prediction. Here, we employ the granular SVM to specially handle the imbalanced classification problem. In all, our proposed method is illustrated in Fig. 1, including three main stages: 1) feature extraction, 2) feature selection and 3) methylation classification.

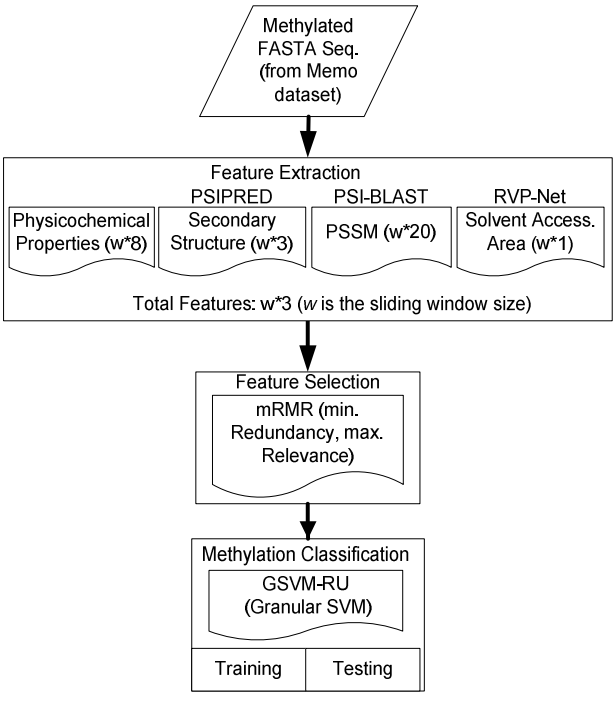

Figure 1. The flowchart of proposed there-step method. Only if the central residue in a sliding window (size w) is Arginine, we shall extract its features.

A. Feature Extraction

To better understand the underlying mechanism of protein Arginine methylation, enough information has to be collected for further machine learning task. So far, most studies of identifying methylation sites only use the primary structure information: the properties of amino acids and the evolutionary profiles. However, it's well known that the protein secondary and tertiary structures have significant impact on many PTMs. Therefore, structure information should also be used to differentiate the methylation status. In this paper, the secondary structure and solvent accessible area information are extracted along with physicochemical properties and PSSM to represent the characteristics of neighboring residues. We believe it's the first time that such comprehensive information has been explored for protein methylation identification.

The physicochemical (PhCh) properties of each amino acid include the status of being polar, charged, positive/negative, aromatic, small, tiny, hydrophilic/hydrophobic, and aliphatic. Each property is represented with a binary bit, so 8 bits are used for PhCh feature.

The evolutionary profiles of each protein sequence are generated by PSI-BLAST [7] against the NCBI non-redundant database with three iterations and a cutoff E-value 10e-3. The numbers in PSSM matrix produced by PSI-BLAST will be scaled in the range [0, 1].

Not all PRMT sequences have been discovered with 3D structures so far, thus the predicted structure information is used for feature extraction. For secondary structures of each sequence, the PSIPRED server [8] can provide reliable three-status (helix, coil, and strand) prediction for each residue of a query sequence. The real values of three kinds of secondary structures are directly incorporated in the feature list.

For solvent accessible area, we also use the predicted information, which is obtained from RVP-net [9]. The RVP-net can output real-valued predictions of accessible surface area for each amino acid; hence, it should bring more information for further classification than binary or discrete predictions.

B. Feature Selection

The features collected above may contain irrelevant and redundant information, which will deteriorate the performance of classifier. Thus, selecting an informative feature subset from original set is necessary. Usually, there are two approaches for feature selection problem: the filter method and wrapper method. Independent to classification task, filter methods use evaluation metrics like statistical correlation to determine the usefulness of every feature. On the other hand, wrapper methods utilize the performance of classification algorithms to measure the worth of a feature. The accuracy changes by including or excluding a specific feature in the learning process indicates its importance in the classification task. Features selected via wrapper methods are highly biased by the specific algorithms, thus, here, filter methods are considered in our feature selection process.

Among several filter feature selection metrics, the mRMR method proposed by Peng et al. [5] has shown better performance for feature selection problems in different domains, e.g., the microarray gene selection. Based on mutual information theory, mRMR criteria try to maximize the dependency between features and target classes. Two directions to reach this goal are minimizing the redundancy between every pairs of features and maximizing the relevance between features and targets, as described in following formula [5]:

$$Max\left(\frac{Relevance}{Redundancy}\right) = Max\left(\frac{\frac{1}{|S|}\sum_{i\in S} MI(h,i)}{\frac{1}{|S|^2}\sum_{i,j\in S} MI(i,j)}\right)$$

where S is the set of selected features, MI(i, j) is the mutual information between feature i and j, and h is the target classes.

C. Granular SVM with Repetitive Undersampling

As known to all, the methylated residues are far less than the non-methylated ones; thus, traditional SVM will be significantly skewed to the minority methylated class so that it always make non-methylated predictions (majority class) on any unknown sequences. Here, the granular SVM repetitive undersampling (GSVM-RU) method [6] is incorporated to handle the imbalance learning problem. Since a single SVM learning normally cannot create a "good" hyperplane for classification, multiple SVM learning by discarding previous extracted support vectors from majority class may push the hyperplane away from minority class. Through this way, the hyperplane will be more evenly located in the middle two classes. More precisely, all samples from minority class are kept during all SVM learning iterations, while the support vectors from majority class are excluded during learning process. Gradually, the samples in majority class will be less and less and then the classification hyperplane is adjusted into an appropriate position.

The G-means value $\sqrt{sensitivity * specificity}$ [10] is used to terminate SVM learning process; when G-means values is getting lower, we stop removing SVs from majority class and use current dataset as reduced training dataset to build optimized classifiers for testing dataset.

IV. Datasets and Experiments

The dataset containing methylated and non-methylated Arginine residues in our experiment is collected from MeMo. This MeMo dataset is produced by scanning possible sites in SWISS-PORT database, refining via checking publications on PubMed, and lastly filtering homogeneous sequences. Their data contain 92 proteins with 255 Arginine methylation sites and around 2700 non-methylation sites. We first collected the features of each Arginine residues with a fix windows size 15 (we tried different sizes from 9 to 21, but 15 has shown best results). As described above, the PSSM matrix of each protein is obtained from PSI-BLAST [7]; the PhCh features are from http://prowl.rockefeller.edu/aainfo/pchem.htm; the SS features are from PSIPRED [8]; and the SAS feature is generated by RVP-Net [9].

We then run the feature selection step and GSVM training step. The reduced number of features has to be predefined in mRMR selection, so we choose various reduced feature size from 10 to 480 (the original size). The LIBSVM Toolbox [11] for Matlab version is used as the basic software to implement the GSVM-RU method. For GSVM-RU training, RBF kernel is

used to remove skewed support vectors from majority class, and then the final SVM model is optimized by grid-searching different cutoff *C* and *gamma*. A 7-fold cross validation is performed to verify the effectiveness of our method. We use the sensitivity (sn), specificity (sp), accuracy (acc.), g-means, and area under ROC (Receiver operating characteristic, AUROC) to measure the classifier performance. Table 1 summarizes the results of different accuracy measurements under different selected feature sizes with a windows size 15.

TABLE I. CLASSIFICATION PERFORMANCE WITH A 7-FOLD CV WITH SVM AND GSVM-RU UNDER DIFFERENT SELECTED FEATURES.

Num. Features	Sn	Sp	Acc.	g-means	auroc
(MeMo.)	0.6960	0.8920	0.8670	N/A	
480 (SVM)	0.3529	0.9626	0.9096	0.5829	0.7909
10	0.6941	0.8684	0.8532	0.7764	0.8036
50	0.7412	0.8288	0.8212	0.7838	0.8363
100	**0.7255**	**0.8804**	**0.8669**	**0.7992**	**0.8447**
150	0.7137	0.8807	0.8662	0.7929	0.8282
200	0.7020	0.8819	0.8662	0.7868	0.8333
250	0.7020	0.8893	0.8730	0.7901	0.8347
300	0.7098	0.8845	0.8693	0.7923	0.8224
350	0.7176	0.8804	0.8662	0.7949	0.8254
400	0.7020	0.8729	0.8580	0.7828	0.8190
480	0.7059	0.8673	0.8532	0.7824	0.8134

We also compare the classification performance over different windows size from 9 to 21. The g-means and AUROC results are summarized in Figure 2 and Figure 3. From these results, we observed that if the number of selected features exceeds certain number around 150, the g-means values are decreased gradually and the AUROC values are also relatively lower. This suggests that some features are important and enough to differentiate the methylated residues; if more features are included, even selected by the mRMR method, they started to bring hurdles to the classification methods. Analyzing these features will reveal some underlying information about the contribution of different groups of features to methylation mechanisms.

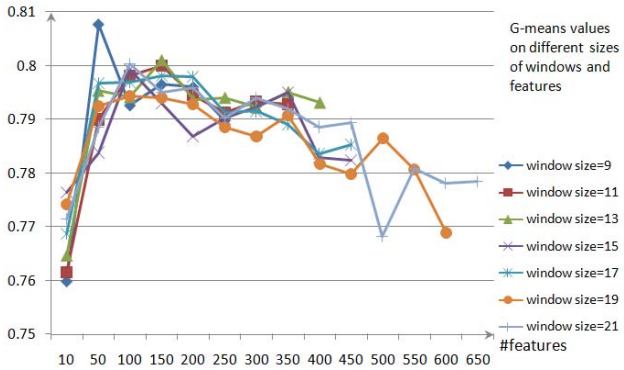

Figure 2. Different g-means values over different windows sizes and extracted features.

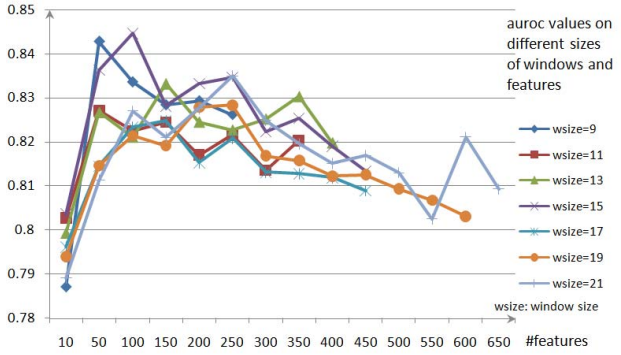

Figure 3. Different auroc values over different windows sizes and extracted features.

V. DISCUSSION

We presented a systemic method to extract features and build classifiers to distinguish Arginine methylation status. A wide range of features are designed and extracted from different biological aspects with different underlying meaning. A feature selection step is involved to remove irrelevant and redundant information, in order to find the contribution of each feature and to improve the performance of classification. A GSVM-RU specially designed for imbalanced data learning is used for final classification step. Our results are comparable to previous publications and even better. Future work will focus on analyzing the selected features to find which group of extracted features could make more impact. We will compare the selected features and find the underlying relations between these features and the methylation mechanism. Besides, we will also consider using other classification method, i.e., decision trees, to generate understandable rules to discover new methylation substrates. Moreover, newest methylation data will be collected to make accurate predictions.

ACKNOWLEDGMENT

We thank the authors of MeMo [4] for providing the methylation dataset for our study. The authors also want to thank Dr. Yuchun Tang for providing the GSVM-RU algorithm and many useful suggestions.

REFERENCES

[1] C. Walsh,"Posttranslational modification of proteins: expanding nature's inventory," Ch. 5. Englewood, CO: Roberts and Co. Publishers, 2005.

[2] D. Plewczynski, A. Tkacz, L. Wyrwicz, and L. Rychlewski, "AutoMotif server: prediction of single residue post-translational modifications in proteins," Bioinfo., vol. 21(10), pp. 2525–2527, 2005.

[3] K. Daily, P. Radivojac, and A. Dunker, "Intrinsic disorder and protein modifications: building an SVM predictor for methylation," In 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, San Diego, CA, pp. 475–481, 2005.

[4] H. Chen, Y. Xue, N. Huang, X. Yao, and Z. Sun, "MeMo: a web tool for prediction of protein methylation modifications," Nucleic Acids Res., vol. 34, pp. 249–253, 2006.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27(8), pp. 1226–1238, 2005.

[6] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs Modeling for Highly Imbalanced Classification," IEEE Tran. on Sys., Man, and Cyber. (Part B), vol. 39(1), pp. 281–288, 2009.

[7] S. Altschul, T. Madden, A. Schaffer, Z. Zhang, W. Miller, and D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acid Res., vol. 25, pp. 3389–3402, 1997.

[8] D. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," J. Mol. Biol., vol. 292, pp. 195–202, 1999.

[9] S. Ahmad, M. Gromiha, and A. Sarai, "RVP-net: online prediction of real valued accessible surface area of proteins from single sequences," Bioinfo., vol. 19, pp. 1849–1851, 2003.

[10] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One sided selection," In: 14th Int'l Conf. on Machine Learning, Nashville, TN, pp. 179–186, 1997.

[11] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001. Available at http://www.csie.ntu.edu.tw/cjlin/libsvm

[12] D.-M. Shien, T.-Y. Lee, W.-C. Chang, J.-B. Hsu, J.-T. Horng, P.-C. Hsu, T.-Y. Wang, and H.-D. Huang, "Incorporating structural characteristics for identification of protein methylation sites," J. Comput. Chem., vol. 30(9), pp. 1532–1543, July 2009.

[13] S. Pal and S. Sif, "Interplay between chromatin remodelers and protein arginine methyltransferases," J. Cell Physiol., vol. 213(2), pp. 306–315, Nov. 2007.

[14] J. Bełtowski and A. Kedra, "Asymmetric dimethylarginine (ADMA) as a target for pharmacotherapy," Pharmacol. Rep., vol. 58(2), pp. 159–178, Mar. 2006.