Application of a Feature Selection Method to Nucleosome Data: Accuracy Improvement and Comparison with Other Methods

MASANORI HIGASHIHARA[†], JOVAN DAVID REBOLLEDO-MENDEZ*, YOICHI YAMADA*, KENJI SATOU*

† Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292
JAPAN
m-higa@jaist.ac.jp http://www.jaist.ac.jp/~m-higa/

* Graduate School of Natural Science and Technology
Kanazawa University
Kakuma-machi, Kanazawa 920-1192
JAPAN

Abstract: - In binary classification problem, data of feature vectors with binary labels are prepared in general. However, today it is well known that using all the features for discrimination does not always the best way to achieve the highest accuracy in prediction. Feature selection is a technique to find a subset of features with the highest accuracy by eliminating features harmful in prediction. Among various methods proposed, in this study we used a method which can be divided in two steps. Firstly, along the ranked features $f_1,\ldots,f_n$ based on Gini index, the feature subsets $\{f_1\},\{f_1,f_2\},...,\{f_1,\ldots,f_n\}$ are tested by SVM with RBF kernel. Secondary, variants of the best feature subset found in the first step are tested in the same way. In the application to the prediction of nucleosome occupancy and modification from genome subsequence, the method achieved a small but assured improvement from the previous study. In addition, observed ranking of features revealed some relationships between features and categories of nucleosome datasets. Finally, the method was compared with other promising methods and outperformed them.

Key-Words: - Epigenetics, Histone, Acetylation, Methylation, Feature selection, Support vector machine, Gini-index, Random forest

1 Introduction

Genes and their expressions are important concepts for understanding how an organism lives. Due to the success of various genome projects including the Human Genome Project, today it is well-understood that an organism has thousands or tens of thousands of genes in its genome sequence. Gene expression (transcription and translation resulting in protein biosynthesis) is essential for life; however it is not constant. For instance, even within the same species, gene expression can differ from each individual, tissue, or physicochemical situation including starvation, cold shock, etc. More importantly, gene expression is regulated by various factors. As summarized by O. Hobert [1], transcription factors (TFs) and microRNAs (miRNAs) regulate genes separately, collaboratively, or oppositely, and they form a complicated network of gene regulation. In addition, nucleosomes are a factor of gene regulation in eukaryotic genomes and they have gotten a lot of attention recently.

A relatively long genome sequence in eukaryote is packaged using a unit called nucleosome which consists of 4 pairs of proteins called histones (H2A, H2B, H3, H4) and 145-147 base pairs of DNA wrapped around the histone octamer (Fig.1). Among the various roles of nucleosomes including the compaction of DNA into chromosomes, gene regulation is essential, since gene transcription is prevented in regions in which the DNA is tightly packed by nucleosomes. In this sense, gene regulation by nucleosomes is on higher level than that by TF and miRNA, and occupancy of nucleosomes on DNA is an important clue to understanding the expression patterns of each gene. Additionally, chemical modification of DNA and/or histones is also related to chromatin formation (tight or loose packing) and gene regulation.

Pokholok et al. [2] reported the results of comprehensive experiment on genome-wide mapping of histone occupancy and modification (acetylation and methylation) in a yeast genome. They conducted systematic analysis and revealed a relationship between histone profiles and gene expression. It implies that if we can predict histone profiles from sequence information with a certain level of accuracy, it might be a useful hint to understanding the expression patterns of genes. Using 10 out of 14 datasets published by Pokholok et al. [2], Pham et al. attempted to predict the occupancy and modification of a DNA sequence fragment [3]. Using a Support Vector Machine (SVM) with RBF kernel and k-gram features of the sequence in various window sizes (k=3,4,5,6,etc.), they achieved a high accuracy of prediction. They also conducted feature ranking using SVM with linear kernel to identify informative features for positive or negative classes. Instead of SVM, Tran et al. used Conditional Random Field (CRF) and obtained similar results regarding accuracy and the ranking of features [4]. However, neither of them made use of the ranking of features to improve prediction accuracy.

Coupled with feature ranking, in this study we applied a technique called feature selection to this problem. Since noisy or misleading features decrease prediction accuracy, eliminating such features is a better way than simply using all the features, and frequently brings improved accuracy in practice. Using feature ranking by the Gini index and feature selection along the ranking, we achieved a small but assured improvement of prediction accuracy. Moreover, we achieved further improvement by searching variants of the feature subset with the best accuracy. Besides accuracy improvement, the results provided some insights about selection of features around the feature set with the best accuracy. In addition to accuracy improvement, the results of feature ranking revealed some relationships between features and groups of datasets.

The rest of this paper is organized as follows. In section 2, after the datasets, problem formulation, and prediction algorithm are briefly described, our method of feature selection is illustrated. In section 3, experimental results are shown with some analysis and interpretation. Finally, section 4 concludes this paper.

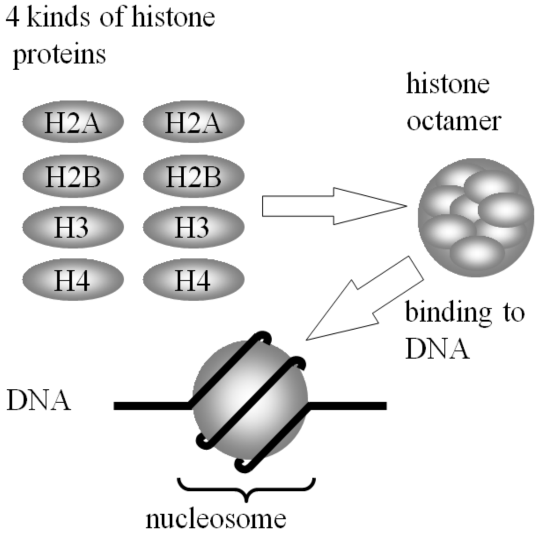

Fig.1. Histone and nucleosome

2 Materials and Methods

2.1 Preparation of positive and negative examples

Table 1 lists 10 datasets about nucleosome occupancy and modification. They are a subset of the datasets published by Pokholok et al. and have been used in this and previous studies [3,4].

In a dataset name in Table 1, “H3” or “H4” indicates the histone type, “K” and a succeeding number indicate a modified amino acid (e.g. “K9” denotes lysine as the 9th amino acid in the histone), and “ac” or “me” indicate the type of modification (acetylation or methylation). A number after “me” indicates times of methylation. A dataset is a set of pairs (position, value) where position and value indicate a specific point of genomic DNA sequence and relative occupancy or modification at this point. Similar to previous studies [3,4], we generated positive and negative examples as follows:

1. For a position with a value greater than 1.2, generate a positive example by extracting a subsequence of 500 base pairs which centers on the position.
2. Similarly, generate a negative example from a position with a value less than 0.8.
3. Others are not used.

The numbers of positive and negative examples are shown in Table 2. Generated subsequences were then converted into vectors by counting k-grams in each subsequence. For instance, if we adopt a window size k=3, frequencies for “AAA”, “AAT”, “AAC”, “AAG”, “ATA”, ... are counted. Though various window sizes were explored in the previous studies [3,4], here we assume k=3 for simplicity. An example is represented as a vector with 64 dimensions of features corresponding to possible tri-nucleotide sequences.

Table 1. Nucleosome datasets.

Dataset	Brief description
H3	H3 occupancy
H4	H4 occupancy
H3K9ac	H3K9 acetylation relative to H3
H3K14ac	H3K14 acetylation relative to H3
H4ac	H4 acetylation relative to H3
H3K4me1	H3K4 monomethylation relative to H3
H3K4me2	H3K4 dimethylation relative to H3
H3K4me3	H3K4 trimethylation relative to H3
H3K36me3	H3K36 trimethylation relative to H3
H3K79me3	H3K79 trimethylation relative to H3

Table 2. The numbers of examples.

Dataset	Positive	Negative
H3	7,667	7,298
H4	6,480	8,121
H3K9ac	15,415	12,367
H3K14ac	18,771	14,277
H4ac	18,410	15,685
H3K4me1	17,266	14,411
H3K4me2	18,143	12,540
H3K4me3	19,604	17,195
H3K36me3	18,892	15,988
H3K79me3	15,337	13,500

2.2 Prediction algorithm and implementation

Support vector machine is a promising algorithm of supervised learning [5] applicable to a huge variety of problems in classification and regression. Also in bioinformatics, SVM has been frequently applied to structure analysis [6,7], gene expression analysis [8], and protein interaction analysis [9]. We adopted SVM with RBF kernel, which was used in [3]. We used a different parameter σ = 0.05 for better accuracy of prediction. About implementation, Pham et al. used their own implementation of SVM. For reproducibility, we used the ksvm function included in the kernlab package [10] for R [11].

2.3 Feature selection and feature ranking

In the field of pattern recognition, feature selection has been actively studied. Wide variety of its application includes text classification [12], protein classification [13], intrusion detection [14], and so on. The problem can be defined as follows: given a whole set of N features, how can we find a subset which achieves the best discrimination performance in prediction? Since the size of search space is 2^N (or 2^N-1 if we exclude an empty feature set), an exhaustive search is not applicable to practical problems. To solve this problem, various methods were proposed. They are classified into three classes. A wrapper method executes learning and prediction and utilizes the result of prediction to decide a feature (or a feature set) that can remain as a candidate or should be discarded. In contrast, a filter method statistically estimates the relevance of features as a preprocessing step without learning and prediction. The third class is an embedded method. Specific to a given learning machine, it performs feature selection within the process of learning. Besides this classification, the choice of a search algorithm is an important factor to characterize a method of feature selection. In addition to heuristics like forward selection and backward elimination, many search algorithms were proposed for feature selection: best-first search, floating search, random search including Relief algorithm, genetic algorithm search, and so on. Traditional algorithms are summarized by Jain et al. [15] and Molina et al. [16]. Newer algorithms and research trends are summarized by Liu et al. [17].

In this study, our algorithm of feature selection was divided into the following two steps:

Step 1) Along a pre-computed ranking of features like $f_1,\ldots,f_n$, where f_1 and f_n are the features at the top and bottom of the ranking respectively, all the subsets $\{f_1\},\{f_1,f_2\},\ldots,\{f_1,\ldots,f_n\}$ are tested by executing learning and prediction by SVM with RBF kernel.

Step 2) Neighbors (variants) of the feature set with the best prediction accuracy in the previous step are tested.

In step 1, we used random forest [18] to compute the Gini index, which is one of the popular measures in feature ranking as well as information gain, t-statistics, etc. Random forest is a kind of ensemble learning algorithm based on randomly generated decision trees. Though it has various advantages, we used it only for generating the Gini index to be used in features ranking. To generate the variants in step 2, we considered a set of at most eight features consecutive in the ranking, which is centered on the feature with the lowest Gini index in the best feature set in step 1. For instance, if $\{f_1,\ldots,f_k\}$ is the best feature set in step 1, union of $\{f_1,\ldots,f_{k-4}\}$ and a set $F \in$ Powerset($\{f_{k-3}, f_{k-2}, f_{k-1}, f_k, f_{k+1}, f_{k+2}, f_{k+3}, f_{k+4}\}$) is tested for every F. Since $k \leq n-4(=60)$, 2^8 feature sets are tested. Otherwise, $2^4 \sim 2^7$ feature sets were tested.

3 Experimental Results

3.1 Feature ranking by random forest

In addition to the result of learning and prediction, randomForest function [19] for R can generate a value called MeanDecreaseGini for each feature. Using this, we can rank features in the order of importance (discriminative power). Fig.2 illustrates the relationship between rank and MeanDecreaseGini normalized into the interval [0,1] in each dataset.

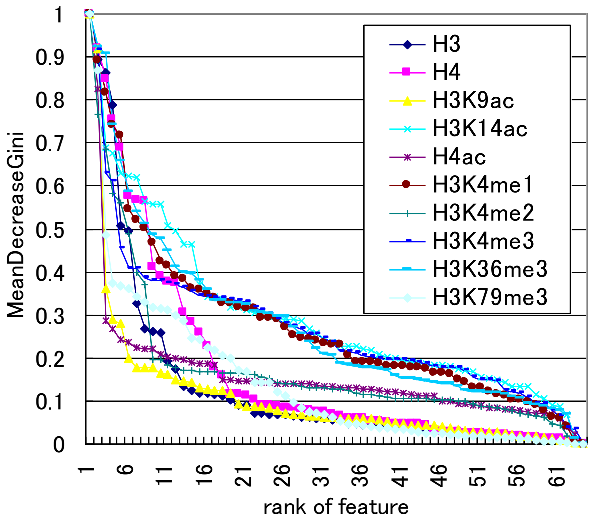

Fig.2. MeanDecreaseGini along feature ranking

Common to all the datasets, the importance of features rapidly decreases in the regions of top 2~11, then slowly in the rests. It means that the majority of features in a dataset have only a small importance. Important features, detected through visual inspection of Fig.2, are listed in Table 3. In this table, it is shown that many datasets have features which mainly consist of "T" (thymine) and "A" (adenine). In two datasets (H3K9ac and H3K36me3), the ratio of "G" (guanine) and "C" (cytosine) are relatively higher. Additional observations are as follows:

- $\boxed{\text{TTT}}$ and $\boxed{\text{AAA}}$ are important in both of H3 and H4 occupancies. It is also important in most of H3 methylations, but not in acetylations.
- **AAT** and **ATT** are important in H3 occupancy (not in H4 occupancy). They are commonly important in all acetylations including H4ac.
- $\underline{\text{ATA}}$ and $\underline{\text{TAT}}$ are important in both of H3 and H4 occupancies, but in contrast to **AAT** and **ATT**, they are selectively important in most of H3 methylations.
- H3K79me3 is special in the sense that $\boxed{\text{TTT}}$, $\boxed{\text{AAA}}$, **AAT**, and **ATT** are not so important in it.

Note that it is not clear whether an important feature is preferred in positive or negative examples.

Table 3. List of important features

Dataset	Range of rank	Features with high importance listed in descending order of rank
H3	1~10	$\boxed{\text{TTT}}$, $\boxed{\text{AAA}}$, TAA, TTA, $\underline{\text{ATA}}$, $\underline{\text{TAT}}$, **AAT**, CCA, **ATT**, TGG
H4	1~8	$\boxed{\text{TTT}}$, $\underline{\text{ATA}}$, $\boxed{\text{AAA}}$, $\underline{\text{TAT}}$, CCA, ATC, TAA, TGG
H3K9ac	1~5	**ATT**, **AAT**, CGC, GCG, TTA
H3K14ac	1~2	**AAT**, **ATT**
H4ac	1~2	**AAT**, **ATT**
H3K4me1	1~5	$\boxed{\text{TTT}}$, $\underline{\text{TAT}}$, CCA, $\underline{\text{ATA}}$, $\boxed{\text{AAA}}$
H3K4me2	1~8	**ATT**, **AAT**, TTA, $\boxed{\text{TTT}}$, $\underline{\text{TAT}}$, $\underline{\text{ATA}}$, TAA, $\boxed{\text{AAA}}$
H3K4me3	1~5	**ATT**, **AAT**, $\boxed{\text{AAA}}$, $\boxed{\text{TTT}}$, $\underline{\text{TAT}}$
H3K36me3	1~11	CCA, $\underline{\text{ATA}}$, TGG, $\underline{\text{TAT}}$, CAA, $\boxed{\text{TTT}}$, $\boxed{\text{AAA}}$, TCA, TTG, ATC, TGA
H3K79me3	1~3	$\underline{\text{ATA}}$, $\underline{\text{TAT}}$, ATC

3.2 Prediction accuracy of feature subsets selected along the ranking

In step 1 described in subsection 2.3, 64 different feature subsets $\{f_1\},\{f_1,f_2\},\ldots,\{f_1,\ldots,f_n\}$ along the ranking were tested for each dataset using SVM with RBF kernel. The results of the prediction are summarized in Fig.3. Subsets with the best prediction accuracy are listed in Table 4.

In Fig.3, prediction accuracy in each dataset moderately increases to the bottom of the feature ranking. It implies that even if the given feature set includes features with low importance, SVM could utilize most of features for better discrimination. Actually, in Table 4, the best feature sets were identical to the full feature sets (size=64) in 3 datasets (H4ac, H3K4me2, H3K36me3). For other datasets, small but assured improvement of accuracy was observed. The accuracy was computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. Similar to Pham et al. [3], threefold cross-validation was adopted for computing accuracy.

3.3 Prediction accuracy of neighbors around the feature subset with best accuracy

Using the best feature subset listed in Table 4, step 2 of our feature selection was conducted for each dataset. The result is also shown in Table 4. Similar to step 1, step 2 achieved a small but assured improvement of accuracy.

Aside from prediction accuracy, it is notable that a feature with high (low) rank is not always included (excluded) in the best feature subset in step 2. For instance, in the dataset H4, GGG(59) was included in the best set, while TAG(56), GCC(57), and AGT(58) were not (see Table 5). A more extreme case is H3K4me1 in which, among 6 features with ranks of 59~64, only the feature with the lowest rank survived and the others were discarded. It implies that the feature selection in step 1 along feature ranking is not sufficient, and a limited but comprehensive search in step 2 could compensate for the disadvantage.

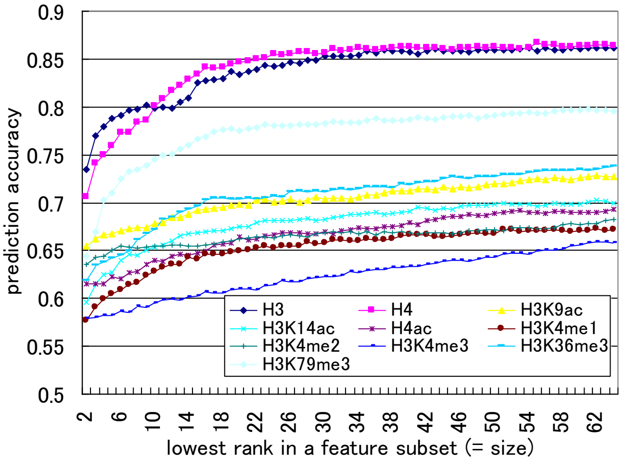

Fig.3. Effects of feature selection along the ranking

Table 4. Prediction accuracies computed by Pham et al. [3], no feature selection, step 1, and step 2.

Dataset	Accuracy (%) in [3], no feature selection, size=64	Accuracy (%), no feature selection, size=64	The best feature subset in step 1 with features and their ranks in ()	Accuracy (%), the best feature subset in step 1	The best feature subset in step 2 with features and their ranks in ()	Accuracy (%), the best feature subset in step 2	Sum of accuracy improvements (%) in step 1 and 2
H3	84.93	86.19	{TTT(1),...,CTA(59)}	86.21	{TTT(1),...,TCG(55), AAC(56), CTA(59), ACT(60), TAG(62)}	**86.45**	0.26
H4	85.91	86.36	{TTT(1),...,CCC(55)}	86.64	{TTT(1),...,AGC(51),ACG(52),TGC(53),CAC(54),CCC(55),GGG(59)}	**86.67**	0.31
H3K9ac	71.04	72.68	{ATT(1),...,CCC(62)}	72.86	{ATT(1),...,TCC(58), AGG(59), GTC(60), CCC(62), GGA(63)}	**72.90**	0.22
H3K14ac	68.64	69.98	{AAT(1),...,GCC(62)}	70.23	{AAT(1),...,GGG(58), GGC(60), CCC(61), GCC(62)}	**70.36**	0.38
H4ac	67.65	69.26	{AAT(1),...,CGG(64)}	69.26	{AAT(1),...,GGA(60), GGG(61), CCG(63), CGG(64)}	**69.32**	0.06
H3K4me1	66.21	67.13	{TTT(1),...,GCG(62)}	67.28	{TTT(1),...,CGT(58), CGG(64)}	**67.56**	0.44
H3K4me2	66.09	68.23	{ATT(1),...,CGG(64)}	68.23	{ATT(1),...,CCC(60), CCG(61), GCG(63), CGG(64)}	**68.28**	0.05
H3K4me3	62.37	65.79	{ATT(1),...,CCG(63)}	65.87	{ATT(1),...,GGC(59), GCC(60), GCG(61), CGC(62), CGG(64)}	**65.92**	0.14
H3K36me3	71.74	**73.80**	{CCA(1),...,CGG(64)}	**73.80**	{CCA(1),...,GGC(60), GCC(61), CGT(62), CCG(63), CGG(64)}	73.80	0.00
H3K79me3	78.25	79.56	{ATA(1),...,CGA(61)}	79.77	{ATA(1),...,AGG(57), GCT(59), CGA(61), TCG(63), CGT(64)}	**79.94**	0.38

Table 5. Features included in the best feature subset in step 2. The rank of an underlined feature is the lowest rank in the best feature subset in step 1. Features with light-gray backgrounds were considered in step 2 to generate variants of it. Features with dark-gray backgrounds were included in all the variants (features with ranks 1~49 were also included, but omitted in this table). Features written in bold face were adopted in the best feature subset in step 2. Features with white backgrounds were not used in any variants.

Dataset \ Rank	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
H3	CGA	GTG	GAG	TGT	AGT	TCG	**AAC**	GGG	GTA	**CTA**	**ACT**	TAC	**TAG**	ACG	CGT
H4	CGT	AGC	**ACG**	**TGC**	**CAC**	**CCC**	TAG	GCC	AGT	**GGG**	GTG	GCT	CGA	TCG	GGC
H3K9ac	TAG	CAG	ACT	CGA	AGT	GGT	ACC	GGG	TCC	**AGG**	**GTC**	CCT	**CCC**	**GGA**	GAC
H3K14ac	TGC	CTG	AGG	GAC	GTC	CGT	CCT	ACG	GGG	TCG	**GGC**	**CCC**	**GCC**	CGG	CCG
H4ac	CTA	AGG	TCG	CTG	GTC	TCC	GCC	CCT	CAG	CCC	GGA	**GGG**	GAC	**CCG**	**CGG**
H3K4me1	GAC	GTC	AGG	CCC	CGA	ACG	TCG	GGG	CGT	CGC	GGC	GCC	GCG	CCG	**CGG**
H3K4me2	GAC	CAG	AGC	TCC	TCG	GCC	CCT	GGG	ACG	CGT	CCC	**CCG**	CGC	**GCG**	**CGG**
H3K4me3	GTC	CGA	GGG	CCT	CGT	GAC	TCG	CCC	ACG	GGC	**GCC**	**GCG**	**CGC**	CCG	**CGG**
H3K36me3	CAG	CTG	CCT	GAC	CCC	CGA	AGG	GTC	ACG	TCG	GGC	**GCC**	**CGT**	**CCG**	**CGG**
H3K79me3	CTA	GGG	GAG	CGG	GAC	CTC	AGC	AGG	ACG	**GCT**	GTC	**CGA**	CCT	**TCG**	**CGT**

3.4 Comparison with other feature selection methods in wrapper approach

To compare our method with popular methods in wrapper approach, we conducted comprehensive experiment of feature selection using Weka [20]. In this experiment, we tested all the combinations of the six classifiers (BayesNet, NaiveBayes, SMO, J48, AdaBoostM1, RandomForest) and five search methods (BestFirst, GeneticSearch, GreedyStepwise, LinearForwardSelection, RankSearch). After that, prediction accuracy of each feature subset was calculated in the same way as subsections 3.2 and 3.3. The results are shown in Tables 6~11. In these tables, prediction accuracy of each feature subset is shown with the size of the feature subset in (). Though some combinations were not successfully generated feature subsets (indicated as '-' in these tables), most of the feature subsets generated by the above classifiers and search methods achieved lower prediction accuracies in comparison with Table 4 by our methods.

Table 6. Prediction accuracies of feature subsets obtained by BayesNet classifier.

	BayesNet				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7899(6)	0.8507(29)	0.7899(6)	0.8258(20)	0.8607(56)
H4	0.8453(20)	0.8475(34)	0.8453(20)	0.8399(19)	0.8463(24)
H3K9ac	0.6984(23)	0.7141(41)	0.6984(23)	0.6919(21)	0.7163(51)
H3K14ac	0.6767(20)	0.6921(44)	0.6767(20)	0.6680(18)	0.6794(26)
H4ac	0.6511(17)	0.6716(34)	0.6511(17)	0.6576(20)	0.6892(53)
H3K4me1	0.6448(19)	0.6556(33)	0.6448(19)	0.6425(15)	0.6657(42)
H3K4me2	0.6475(7)	0.6600(30)	0.6475(7)	0.6464(5)	0.6491(7)
H3K4me3	0.5997(10)	0.6354(35)	0.5997(10)	0.6013(10)	0.6090(17)
H3K36me3	0.7098(28)	0.7182(38)	0.7098(28)	0.7024(19)	0.7352(58)
H3K79me3	0.7653(20)	0.7846(45)	0.7653(20)	0.7768(26)	0.7913(43)

Table 7. Prediction accuracies of feature subsets obtained by NaiveBayes classifier.

	NaiveBayes				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8394(26)	0.8445(28)	0.8174(14)	0.8386(21)	0.7909(8)
H4	0.8568(30)	0.8554(36)	0.8568(30)	0.8395(18)	0.8463(24)
H3K9ac	0.7002(22)	0.7100(36)	0.6945(18)	0.7018(24)	0.7151(50)
H3K14ac	0.6663(18)	0.6888(35)	0.6620(15)	0.6765(24)	0.6709(16)
H4ac	0.6655(22)	0.6774(33)	0.6655(22)	0.6628(26)	0.6926(64)
H3K4me1	0.6391(19)	0.6568(31)	0.6390(16)	0.6427(18)	0.6639(41)
H3K4me2	0.6443(6)	0.6577(23)	0.6450(6)	0.6422(4)	0.6491(7)
H3K4me3	0.6231(32)	0.6378(34)	0.6054(10)	0.6092(15)	0.6579(64)
H3K36me3	0.7123(28)	0.7151(33)	0.7123(28)	0.6992(24)	0.7126(30)
H3K79me3	0.7890(43)	0.7850(36)	0.7727(18)	0.7829(34)	0.7926(46)

Table 8. Prediction accuracies of feature subsets obtained by SMO classifier.

	SMO				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8382(24)	0.8581(41)	0.8228(12)	0.8563(41)	0.8539(44)
H4	0.8527(47)	0.8613(48)	0.8494(31)	0.8478(24)	0.8625(62)
H3K9ac	0.7072(40)	-	0.6954(20)	0.6994(23)	0.7142(44)
H3K14ac	0.6786(26)	0.6892(41)	-	-	0.6984(58)
H4ac	0.6669(25)	0.6827(47)	0.6667(23)	0.6684(27)	0.6880(52)
H3K4me1	0.6484(32)	0.6614(39)	0.6457(19)	0.6459(29)	0.6717(61)
H3K4me2	0.6623(14)	0.6688(37)	0.6623(14)	0.6606(22)	0.6810(62)
H3K4me3	0.6271(27)	0.6460(43)	0.6134(16)	0.6042(13)	0.6573(59)
H3K36me3	0.7058(26)	0.7208(44)	0.7036(23)	0.7238(45)	0.7368(63)
H3K79me3	0.7825(35)	0.7953(44)	0.7816(29)	0.7880(45)	0.7961(61)

Table 9. Prediction accuracies of feature subsets obtained by J48 classifier.

	J48				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.8072(8)	0.8505(36)	0.8016(6)	0.8072(8)	0.7909(8)
H4	0.8231(10)	0.8482(26)	0.8231(10)	0.8132(9)	0.8373(21)
H3K9ac	0.6698(4)	-	0.6698(4)	0.6698(4)	0.6659(4)
H3K14ac	0.6373(5)	-	0.6373(5)	0.6275(4)	0.6424(10)
H4ac	0.6190(4)	0.6368(12)	0.6190(4)	0.6190(4)	0.6210(3)
H3K4me1	0.6205(7)	-	0.6111(5)	0.6137(6)	0.5993(4)
H3K4me2	0.6475(5)	-	0.6418(3)	0.6475(5)	0.6491(7)
H3K4me3	0.5858(4)	-	0.5858(4)	0.5858(4)	0.5804(4)
H3K36me3	-	-	0.6549(5)	0.6549(5)	0.6788(10)
H3K79me3	-	-	0.7302(7)	-	0.7226(7)

Table 10. Prediction accuracies of feature subsets obtained by AdaBoostM1 classifier.

	AdaBoostM1				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7873(5)	0.8204(21)	0.7873(5)	0.7873(5)	0.7909(8)
H4	0.8110(10)	0.8527(32)	0.8110(10)	0.8134(9)	0.8286(16)
H3K9ac	0.6784(8)	0.6909(27)	0.6784(8)	0.6784(8)	0.6709(5)
H3K14ac	0.6569(10)	0.6726(29)	0.6569(10)	0.6569(10)	0.6737(21)
H4ac	0.6289(6)	0.6666(29)	0.6289(6)	0.6289(6)	0.6709(39)
H3K4me1	0.6308(10)	0.6485(30)	0.6308(10)	0.6259(9)	0.6590(29)
H3K4me2	0.6471(5)	0.6611(21)	0.6471(5)	0.6471(5)	0.6526(10)
H3K4me3	0.5918(6)	0.6188(29)	0.5918(6)	0.5918(6)	0.6169(26)
H3K36me3	0.6765(8)	0.7016(33)	0.6765(8)	0.6614(6)	0.7118(27)
H3K79me3	0.7500(11)	0.7706(32)	0.7500(11)	0.7500(11)	0.7642(19)

Table 11. Prediction accuracies of feature subsets obtained by RandomForest classifier.

	RandomForest				
	BestFirst	GeneticSearch	GreedyStepwise	LinearForwardSelection	RankSearch
H3	0.7296(2)	0.8501(33)	0.7296(2)	0.7296(2)	0.8553(48)
H4	0.8434(21)	0.8508(32)	0.6967(2)	0.8492(21)	0.8551(32)
H3K9ac	0.6544(2)	-	0.6544(2)	0.6544(2)	0.7058(35)
H3K14ac	0.5993(2)	-	0.5993(2)	0.5993(2)	0.6950(49)
H4ac	0.6148(2)	-	0.6148(2)	0.6148(2)	0.6926(63)
H3K4me1	0.5878(2)	-	0.5878(2)	0.5878(2)	0.6674(47)
H3K4me2	0.6390(2)	-	0.6390(2)	0.6390(2)	0.6742(47)
H3K4me3	0.5785(2)	-	0.5785(2)	0.5785(2)	0.6577(60)
H3K36me3	0.6187(2)	-	0.6187(2)	0.6187(2)	0.7218(37)
H3K79me3	0.6659(2)	-	0.6659(2)	0.6659(2)	0.7903(41)

3.5 Effect of longer window size (k=4)

In the above experiments, we basically adopted only one window size, that is, k=3. It is convenient to save the number of features and possible space of feature subsets. To demonstrate that our method is also useful for larger number of features, we conducted the same experiments with k=4. Though the number of features increased to 256, Table 12 shows that our method is still effective.

4 Conclusion

In this study, we attempted to improve the accuracy of predicting occupancy, acetylation, and methylation of nucleosomes. First, the importance of features was computed through learning and prediction by random forest. Features were ranked using the computed value MeanDecreaseGini. The results of feature ranking implied that some features are selectively important in some groups of datasets, e.g. 3 datasets about acetylation. Then, we conducted feature selection in two steps. Step 1 searches the best feature subset along the ranking. Compared with full feature sets, feature subsets with better prediction accuracy were found by this feature selection. It was also revealed that the majority of features have only a small importance; however, SVM attempts to utilize them as much as possible. Finally in step 2, variants of the best feature subset in step 1 were tested. As a result, the accuracy was improved again. From the best feature subsets found in step 2, it was suggested that step 2 might work complementarily. Similar experiments were conducted also for k=4, and it was shown that our method is still effective for the data with 256 features.

To compare our method with other methods, we conducted comprehensive experiments of feature selection using Weka. All the combinations of the six classifiers and five search methods were tested, and our method outperformed them.

Table 12. Prediction accuracies computed by Pham et al. [3], no feature selection, step 1, and step 2 (k=4).

Dataset	Accuracy (%) in [3], no feature selection, size=256	Accuracy (%), no feature selection, size=256	The best feature subset in step 1 with features and their ranks in ()	Accuracy (%), the best feature subset in	The best feature subset in step 2 with features and their ranks in ()	Accuracy (%), the best feature subset in	Sum of accuracy improve-ments (%) in step 1
H3	85.88	86.43	{AAAA(1),...,ACGC(243)}	86.47	{AAAA(1),...,GGGT(239), GACG,GGGC,GCGT,ACGC}	**86.47**	0.04
H4	87.14	87.04	{TATA(1),...,AGGG(244)}	87.24	{TATA(1),...,CCTA(240),CGCC,GGCG,AGGG,GGCC}	**87.32**	0.29
H3K9ac	73.64	74.98	{AATT(1),...,GACG(250)}	75.07	{AATT(1),...,ACCC(246), CGGG,GGAC,CGGA,GACG,TCCG	**75.08**	0.10
H3K14ac	71.28	**73.28**	{TATA(1),...,GGGG(256)}	**73.28**	{TATA(1),...,CCGG(252), CGGC,CGGG,GGGG}	**73.28**	0.00
H4ac	69.93	**72.06**	{AATT(1),...,CCCC(256)}	**72.06**	{AATT(1),...,GCGG(252), GCGC,CGGG,GGGG,CCCC}	**72.06**	0.00
H3K4me1	68.29	69.53	{TATA(1),...,CGGC(251)}	69.64	{TATA(1),...,CGCC(247), GCGG,CCGC,CGGC,CCGG,GGGG	**69.71**	0.18
H3K4me2	67.05	68.89	{ATTT(1),...,CGGG(255)}	68.89	{ATTT(1),...,CCCG(251), GGGG,GCGC,CGGG,CGCG}	**68.97**	0.08
H3K4me3	65.09	68.38	{AATT(1),...,GCGC(254)}	68.46	{AATT(1),...,CCGG(250), CCCG,CCCC,CGGG,CGCG}	**68.57**	0.19
H3K36me3	73.37	75.09	{TATA(1),...,CGGG(256)}	75.09	{TATA(1),...,CGGC(252), CCGG,CCCC}	**75.19**	0.09
H3K79me3	79.91	80.39	{TATA(1),...,CGGG(244)}	80.53	{TATA(1),...,ACCC(240), CGGT,CCCC,CGGG,GGCG,TCCG,	**80.58**	0.19

The method proposed in this paper is classified as a wrapper method since it executes learning and prediction, and utilizes the results of a prediction to find a better feature subset. In general, one execution of random forest and at most $n + 2^i$ executions of SVM are needed in this method, where in this study, *n* and *i* are 64~256 and 8, respectively. However, in the case that *n* is large (e.g. ~20,000 in a gene selection problem on human microarray data), *n* times execution may be impractical. If we can reduce the search space into a few candidates using the importance of features, step 1 becomes more practical. On the other hand, step 2 currently tests all the 2^i feature subsets. Application of a more efficient feature selection method to this limited search space will be considered as one subject of future work.

References:

[1] O. Hobert, Gene Regulation by Transcription Factors and MicroRNAs, *Science*, Vol.319, No.5871, 2008, pp. 1785-1786.

[2] D.K. Pokholok et al., Genome-wide Map of Nucleosome Acetylation and Methylation, *Cell*, Vol.122, pp.517-527.

[3] T.H. Pham, D.H. Tran, T.B. Ho, K. Satou and G. Valiente, Qualitatively Predicting Acetylation and Methylation Areas in DNA sequences, *Genome Informatics*, Vol.16, No.2, 2005, pp.3-11.

[4] D.H. Tran, T.H. Pham, K. Satou and T.B. Ho, Conditional Random Fields for Predicting and Analyzing Histone Occupancy, Acetylation and Methylation Areas in DNA Sequences, *Applications of Evolutionary Computing, Lecture Notes in Computer Science*, Vol.3907, 2006, pp.221-230.

[5] V. Vapnik, *Statistical Learning Theory*, Wiley, 1998

[6] J. J. Ward, L. J. McGuffin, B. F. Buxton and D. T. Jones, Secondary structure prediction with support vector machines, *Bioinformatics*, Vol.19, No.13, 2003, pp.1650-1655.

[7] X.D. Sun and R.B. Huang, Prediction of protein structural classes using support vector machines, *Amino Acids*, Vol.30, No.4, 2006, pp. 469-475.

[8] Y. Lee and C.K. Lee, Classification of multiple cancer types by multicategory support vector

machines using gene expression data, *Bioinformatics*, Vol.19, No.9, 2003, pp. 1132-1139.

[9] A. Ben-Hur and W.S. Noble, Kernel methods for predicting protein-protein interactions, *Bioinformatics*, Vol.21, Suppl.1, 2005, pp. i38-i46.

[10] A. Karatzoglou, A. Smola, K. Hornik and A. Zeileis, kernlab – An S4 Package for Kernel Methods in R, *Journal of Statistical Software*, Vol.11, No.9, 2004, pp.1-20.

[11] R. Ihaka and R. Gentleman, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, Vol.5, No.3, 1996, pp.299-314.

[12] M. Ikonomakis, S. Kotsiantis and V. Tampakas, Text Classification Using Machine Learning Techniques, *WSEAS Transactions on Computers*, Issue 8, Volume 4, 2005, pp. 966-974.

[13] R. Rakotomalala and F. Mhamdi, Supervised and Unsupervised Feature Reduction for Protein Classification, *WSEAS Transactions on Information Science and Applications*, Issue 12, Volume 3, 2006, pp.2448-2455.

[14] Kun-Ming Yu and Ming-Feng Wu, Protocol – Based With Feature Selection in Intrusion Detection, *WSEAS Transactions on Computer Research*, Issue 3, Volume 3, 2008, pp.135-146.

[15] A. Jain and D. Zongker, Feature Selection: Evaluation, Application, and Small Sample Performance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.19, No.2, 1997, pp. 153-158.

[16] L.C. Molina, L. Belanche and A. Nebot, Feature Selection Algorithms: A Survey and Experimental Evaluation, *Proc. of International Conference on Data Mining 2002 (ICDM 2002)*, 2002, pp. 306-313.

[17] H. Liu et al., Evolving Feature Selection, *IEEE Intelligent Systems*, Vol.20, No.6, 2005, pp. 64-76.

[18] L. Breiman, Random Forests, *Machine Learning*, Vol.45, No.1, 2001, pp. 5-32.

[19] A. Liaw and M. Wiener, Classification and Regression by randomForest, *Rnews 2002*, Vol.2, 2002, pp. 18-22.

[20] I.H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.