Qualitatively Predicting Acetylation and Methylation Areas in DNA Sequences

Tho Hoan Pham[1] **Dang Hung Tran**[2]
h-pham@jaist.ac.jp tran@jaist.ac.jp
Tu Bao Ho[2,3] **Kenji Satou**[2,3] **Gabriel Valiente**[4]
bao@jaist.ac.jp ken@jaist.ac.jp valiente@lsi.upc.edu

[1] Faculty of Information Technology, Hanoi University of Pedagogy, 136 Xuan Thuy, Cau Giay, Hanoi, Vietnam
[2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[3] Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Japan
[4] Department of Software, Technical University of Catalonia, E-08034 Barcelona, Spain

Abstract

Eukaryotic genomes are packaged by the wrapping of DNA around histone octamers to form nucleosomes. Nucleosome occupancy, acetylation, and methylation, which have a major impact on all nuclear processes involving DNA, have been recently mapped across the yeast genome using chromatin immunoprecipitation and DNA microarrays. However, this experimental protocol is laborious and expensive. Moreover, experimental methods often produce noisy results. In this paper, we introduce a computational approach to the qualitative prediction of nucleosome occupancy, acetylation, and methylation areas in DNA sequences. Our method uses support vector machines to discriminate between DNA areas with high and low relative occupancy, acetylation, or methylation, and rank k-gram features based on their support for these DNA modifications. Experimental results on the yeast genome reveal genetic area preferences of nucleosome occupancy, acetylation, and methylation that are consistent with previous studies. Supplementary files are available from `http://www.jaist.ac.jp/~tran/nucleosome/`.

Keywords: histone proteins, acetylation, methylation, support vector machines

1 Introduction

Eukaryotic genomes are packaged into nucleosomes that consist of 145–147 base pairs of DNA wrapped around a histone octamer [11]. The histone components of nucleosomes and their modification state (of which acetylation and methylation are the most important ones) can profoundly influence many genetic activities, including transcription [2, 7, 8, 15], DNA repair, and DNA remodeling [12].

The mapping of nucleosome occupancy and acetylation and methylation has been recently conducted by several research groups using the combination of chromatin immunoprecipitation and whole-genome DNA microarrays, called ChiP-Chip protocol [1, 2, 8, 10, 15, 16, 17]. However, this new experimental method is laborious and expensive, and produces noisy results [15]. Hence, computational methods would be valuable.

Since histone octamers are identical for all nucleosomes in all DNA sequences of a species, the characteristics (for example, acetylation and methylation) of an individual nucleosome depend on the actual DNA sequence area incorporated. The majority of acetylation and methylation sites in histones occur at specific highly conserved residues: acetylation sites include at least nine lysines in histone H3 and H4 (H3K9, H3K14, H3K18, H3K23, H3K27, H4K5, H4K8, H4K12, and H4K16) and less

Table 1: Datasets of histone occupancy, acetylation, and methylation by ChiP-Chip protocol in vivo [15]. POS and NEG are the number of positive and negative examples, respectively.

Dataset	Full name	POS	NEG	Description
H3	H3.YPD	7,667	7,298	H3 occupancy
H4	H4.YPD	6,480	8,121	H4 occupancy
H3K9ac	H3K9acvsH3.YPD	15,415	12,367	H3K9 acetylation relative to H3
H3K14ac	H3K14acvsH3.YPD	18,771	14,277	H3K14 acetylation relative to H3
H4ac	H4acvsH3.YPD	18,410	15,685	H4 acetylation relative to H3
H3K4me1	H3K4me1vsH3.YPD	17,266	14,411	H3K4 monomethylation relative to H3
H3K4me2	H3K4me2vsH3.YPD	18,143	12,540	H3K4 dimethylation relative to H3
H3K4me3	H3K4me3vsH3.YPD	19,604	17,195	H3K4 trimethylation relative to H3
H3K36me3	H3K36me3vsH3.YPD	18,892	15,988	H3K36 trimethylation relative to H3
H3K79me3	H3K79me3vsH3.YPD	15,337	13,500	H3K79 trimethylation relative to H3

conserved sites in histone H2A and H2B; methylation sites include H3K4, H3K9, H3K27, H3K36, H3K79, H3R17, H4K20, H4K59, and H4R3 [13]. When a nucleosome appears in a specific DNA sequence area, these potential sites can have a certain acetylation or methylation level [8, 15].

In this paper, we introduce a computational approach to qualitatively predict nucleosome occupancy, acetylation, and methylation areas in DNA sequences. Our method uses support vector machines (SVM) [5, 18], a promising machine learning technique for bioinformatics, to discriminate DNA areas with high and low relative occupancy, acetylation, and methylation. Furthermore, the SVM method can rank k-gram features based on their support for these DNA modifications. The most informative features reveal genetic area preferences of nucleosome occupancy, acetylation, and methylation that are consistent with previous studies.

2 Materials and Methods

2.1 Datasets

From the genome-wide map of nucleosome acetylation and methylation reported in [15], we extracted 10 datasets and used them to illustrate the performance of our method. These datasets are described in detail in Table 1. Each example in the datasets corresponds to a DNA sequence area (segment) with a fixed length L, called L-DNA sequence area (in our experiments, we selected $L = 200, 500, 1000, 1500$). A DNA sequence area is assigned to the positive class if the relative occupancy, acetylation, or methylation [15] measured at its middle position is greater than 1.2, and to the negative class if the relative occupancy, acetylation, or methylation is lesser than 0.8. Otherwise, noisy examples are ignored.

2.2 Features of a DNA Sequence Area

Each L-DNA sequence area needs to be represented by a set of features that can be input to a machine learning system, that is, a support vector machine in this paper. Here, we use k-grams (patterns of k consecutive nucleotide symbols) to generate these features by using a k-sliding window along a DNA sequence area to compute the number of occurrences of each k-gram. Each example is thus represented by a 4^k-dimensional vector of the number of occurrences of different k-grams.

2.3 Binary Support Vector Machine

The support vector machine (SVM) is a learning technique based on statistical learning theory. The basic idea of applying SVM to binary pattern classification can be briefly stated as follows. First, map the input vectors x_i to a vector $\phi(x_i)$ in a feature space (often with a higher dimension), either linearly or non-linearly, which is relevant to the selection of the kernel function. Second, obtain an optimized linear division within the feature space from the first step, that is, construct a hyperplane $w^T\phi(x_i) + b$ that separates the two classes.

The implementation of SVM is as follows. Let $(x_i, y_i), i = 1, \ldots, \ell$, be a training dataset, where x_i is a vector and $y_i = \pm 1$ is a class attribute. SVM training solves the following primal problem:

$$\begin{cases} \min\limits_{w,b,\xi} \dfrac{w^T w}{2} + C \sum\limits_{i=1}^{\ell} \xi_i \\ y_i(w^T\phi(x_i) + b) \geqslant 1 - \xi_i, \quad i = 1, \ldots, \ell \\ \xi_i \geqslant 0, \quad i = 1, \ldots, \ell \end{cases}$$

Its dual is a quadratic optimization problem:

$$\begin{cases} \min\limits_{\alpha} \dfrac{\alpha^T Q \alpha}{2} - e^T \alpha \\ 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, \ell \\ y^T \alpha = 0 \end{cases}$$

where e is the vector of all ones, $C > 0$ is an error penalty parameter, $y = \{y_i\}_{i=1,\ldots,\ell}$, $Q_{ij} = y_i y_j K(x_i, x_j)$, $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is a kernel function, and $\phi(x_i)$ maps x_i into a higher (maybe infinite) dimensional space. So $K(x_i, x_j)$ is a symmetric positive definite function that reflects the similarity between examples x_i and x_j. In our research, we employed a linear function $K(x_i, x_j) = x_i.x_j$ and a radial basis function (RBF) $K(x_i, x_j) = \exp(-\gamma(x_i - x_j)^2)$ as kernel functions. The SVM classification function, once trained, has the following form:

$$f(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{1}$$

where $\alpha = \{\alpha_i\}_{i=1,\ldots,\ell}$ is the solution of the above dual problem and b is in the solution of the primal problem. Based on Karush-Kuhn-Tucker theory [9], the solutions of the primal and dual problems satisfy the following equation:

$$\alpha_i \left\{ y_i(w^T\phi(x_i) + b) - 1 + \xi_i \right\} = 0\,.$$

Therefore, if $\alpha_i \neq 0$ for some i, then $y_i(w^T\phi(x_i) + b) - 1 + \xi_i = 0$. In this case, x_i is called a *support vector.*

SVM has a solid theoretical background, a good performance in practice, and a guaranteed global optimum. It can also handle large datasets and is easier to implement and train than a neural network. A more detailed description of SVM can be found in [5, 18].

2.4 SVM Method for Feature Selection

Ranking informative (discriminant) features is of fundamental and practical interest in data mining and knowledge discovery. SVM has been successfully applied to this task [4, 6, 14]. When SVM uses a linear kernel, it finds an optimal hyperplane that separates the positive from the negative class in the original space (not mapping into a higher dimensional space). This optimal hyperplane has then the following form (replacing $K(x, y) = x.y$ in Eq. 1):

$$f(X = (f_1, f_2, \ldots, f_m)) = \sum_{i=1}^{m} w_i f_i + b\,. \tag{2}$$

Table 2: Results of acetylation and methylation prediction $acc(cc)$ from a set of k-gram features and from two sets of k-gram features. Both the accuracy (acc) and the correlation coefficient (cc) are shown.

Dataset	$k = 3$		$k = 4$		$k = 5$		$k = 6$		$k = 3, 4$		$k = 4, 5$		$k = 5, 6$	
	acc	cc	acc	cc	acc	cc	acc	cc	acc	cc	acc	cc	acc	cc
H3	84.93	0.70	85.88	0.72	85.50	0.71	85.10	0.70	85.88	0.72	**86.41**	**0.73**	85.23	0.70
H4	85.91	0.71	87.14	0.74	87.77	0.75	87.95	0.75	87.32	0.74	**88.09**	**0.76**	87.87	0.75
H3K9ac	71.04	0.41	73.64	0.47	75.58	0.51	77.27	0.54	73.44	0.47	75.99	0.51	**77.54**	**0.54**
H3K14ac	68.64	0.35	71.28	0.41	73.25	0.47	76.13	0.51	71.04	0.41	74.63	0.49	**76.34**	**0.51**
H4ac	67.65	0.35	69.93	0.39	70.79	0.41	74.91	0.49	68.65	0.39	72.99	0.45	**75.57**	**0.50**
H3K4me1	66.21	0.31	68.29	0.35	70.26	0.39	72.11	0.43	67.22	0.35	70.52	0.41	**72.42**	**0.44**
H3K4me2	66.09	0.27	67.05	0.29	69.41	0.35	71.14	0.39	61.82	0.30	70.30	0.40	**71.27**	**0.39**
H3K4me3	62.37	0.24	65.09	0.30	68.06	0.36	71.56	0.42	62.83	0.29	69.01	0.39	**72.45**	**0.46**
H3K36me3	71.74	0.43	73.37	0.46	74.56	0.48	76.99	0.54	73.44	0.46	75.64	0.51	**77.12**	**0.54**
H3K79me3	78.25	0.56	79.91	0.60	80.87	0.61	82.15	0.64	79.85	0.59	81.44	0.63	**82.57**	**0.65**

Table 3: Detailed information on the best prediction results from Table 2. TP, FP, TN, FN are the number of true positive, false positive, true negative, and false negative examples, respectively.

Dataset	TP	FP	TN	FN
H3	6,490	1,177	6,441	857
H4	5,631	849	7,231	890
H3K9ac	12,523	2,892	9,020	3,347
H3K14ac	15,063	3,168	9,625	4,652
H4ac	14,701	3,709	11,064	4,621
H3K4me1	13,817	3,449	9,125	5,286
H3K4me2	15,675	2,468	6,192	6,348
H3K4me3	15,289	4,315	11,371	5824
H3K36me3	15,322	3,570	11,594	4,394
H3K79me3	13,161	2,176	10,650	2,850

We can change the sign of the weights $w_i, i = 1, \ldots, m$, and b in the above function such that if $f(X) > 0$ then X would be classified as a positive example and otherwise, as a negative example. It can be clearly seen that if w_i is positive, then feature i would support the positive class. Otherwise, this feature would support the negative class (or prevent the positive class), and the larger the absolute value of w_i, the stronger feature i supports (or prevents) the respective class. From this remark, we define the weight w_i as the *support* of feature i.

3 Results and Discussion

3.1 Prediction of Histone Occupancy, Acetylation, and Methylation

We used SVM with a RBF kernel ($\gamma = 0.001$) (Section 2.3) to do threefold cross-validation on 10 datasets (Table 1). The accuracy (acc) and correlation coefficient (cc) criteria were used to report the results:

$$acc = \frac{(TP + TN)}{(TP + FP + TN + FN)}, \quad cc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where TP, TN, FP, FN are the number of true positive, true negative, false positive, and false negative examples, respectively.

Table 4: Most informative features for positive class, from a set of 3-gram and 4-gram features. TF is the number of transcription factor binding motifs containing the feature.

	Feature	Weight	TF	Feature	Weight	TF	Feature	Weight	TF
H3	TGGC	2.35	13	GCGA	2.18	15	CCTG	2.16	13
	CCT	1.99	51	GGGA	1.87	13	TGCG	1.79	14
	CAG	1.78	35	TGTG	1.74	16	CGTT	1.68	20
H4	CAAA	2.78	20	TATC	2.41	15	ATC	2.21	55
	TTTG	2.01	22	CCA	1.87	57	GATA	1.80	19
	GGA	1.57	52	TGG	1.56	41	ACAG	1.55	9
H3K9ac	CCGG	3.12	22	TATA	2.56	21	GGTC	2.20	8
	CTTG	2.13	7	GGCC	1.93	12	GGGC	1.73	7
	CGGT	1.70	12	GCCC	1.63	11	TCGC	1.59	7
H3K14ac	TATA	2.32	21	CCGG	2.19	22	TCGC	2.04	7
	TAAG	2.01	13	GTCC	1.92	6	GGCC	1.87	12
	TAGT	1.73	10	GGTC	1.71	8	TTTT	1.69	33
H4ac	TCGC	3.12	7	GCGA	3.00	15	CCGG	2.45	22
	GGCC	2.13	12	CCCG	2.11	14	GTCC	2.03	6
	GGTC	2.01	8	CGGT	1.92	12	TATA	1.75	21
H3K4me1	TATC	2.26	15	GACG	1.97	18	CCGC	1.67	27
	CAAA	1.57	20	CGTC	1.56	16	TTTG	1.49	22
	CCA	1.47	57	CAT	1.45	67	GATA	1.42	19
H3K4me2	TAAG	1.69	13	CGGT	1.62	12	CGA	1.32	60
	GTCC	1.28	6	CCCG	1.23	14	TCGC	1.23	7
	TGAG	1.20	9	CACT	1.18	17	CCGG	1.12	22
H3K4me3	CCGG	3.09	22	TCGC	3.04	7	GCGA	2.79	15
	CCCG	2.78	14	TATA	2.69	21	GTCC	2.26	6
	GGCC	2.11	12	TAAG	2.03	13	ACCC	1.96	23
H3K36me3	CGTC	2.24	16	CACC	1.84	23	GACG	1.78	18
	ACGA	1.77	23	ACC	1.72	59	TGG	1.59	41
	CTTC	1.52	15	CAAA	1.51	20	GATA	1.49	19
H3K79me3	CAAA	2.44	20	ATC	2.27	55	GATA	2.24	19
	TATC	2.24	15	GGTA	1.93	13	TTTG	1.85	22
	TACC	1.66	15	ATCC	1.56	13	TGGA	1.55	11

Through various experiments we found that our method gave the best results when predicting nucleosome occupancy, acetylation, and methylation for DNA sequence areas of length $L = 500$ (data not shown). Table 2 shows the accuracy *acc* and the correlation coefficient *cc* of relative histone occupancy predictions (H3, H4), acetylation predictions (H3K9ac, H3K14ac, H4ac), and methylation predictions (H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3) with several different types of k-gram features. Due to the time complexity, we have tried $k \leqslant 6$ only. As can be seen, our method offers the best results for predicting histone H3 and H4 occupancy when we use both features of 4-grams and 5-grams, and for predicting all acetylation and methylation sites when using both features of 5-grams and 6-grams. The detailed information on the best prediction results is reported in Table 3.

In general, when using more features (increasing k in k-grams), our method would give better prediction accuracy. But we think that when k is large enough, accuracy would reach the highest (similar to the case of H3 and H4 occupancy, which reaches the highest accuracy when using both features of 4-grams and 5-grams) and after that it would deacrease because of the over-representation (or overfitting) problem. In the future, we will use a more powerful computer or cluster to find the optimal value of k for the best acetylation and methylation prediction with our method.

Table 5: Most informative features for negative class, from a set of 3-gram and 4-gram features. TF is the number of transcription factor binding motifs containing the feature.

	Feature	Weight	TF	Feature	Weight	TF	Feature	Weight	TF
H3	CGCG	-5.33	18	GCGC	-3.32	11	CGC	-2.46	51
	TTTT	-2.40	33	GCG	-2.11	53	CGTG	-1.98	22
	CGG	-1.88	60	GCGG	-1.87	18	CCG	-1.86	80
H4	CGCG	-3.57	18	TTTT	-3.14	33	TATA	-3.07	21
	AAAA	-2.96	32	GCG	-2.84	53	CGC	-2.49	51
	CGTG	-2.47	22	GCGC	-2.15	11	GGCC	-1.78	12
H3K9ac	TTTG	-2.43	22	CAAA	-2.43	20	GGGG	-2.14	18
	CCGC	-2.10	27	TACC	-2.05	15	AAT	-2.04	86
	GCGG	-1.81	18	CCCC	-1.81	15	CGTC	-1.77	16
H3K14ac	CGGC	-2.64	12	CCCC	-2.35	15	TATC	-2.28	15
	CAAA	-2.27	20	CGTC	-2.18	16	TTTG	-2.17	22
	CCGC	-1.98	27	GGGG	-1.80	18	CGCC	-1.51	20
H4ac	CGGC	-2.94	12	GCGG	-2.62	18	CCCC	-2.46	15
	CCGC	-2.43	27	GCCG	-2.34	26	CGTC	-2.17	16
	GGGG	-2.02	18	CGCC	-1.96	20	AAT	-1.86	86
H3K4me1	CGCG	-2.68	18	TATA	-2.29	21	CACG	-2.19	20
	CCGG	-1.81	22	CGTG	-1.58	22	TCGC	-1.55	7
	TTTT	-1.52	33	CTTA	-1.39	15	CCCG	-1.34	14
H3K4me2	CGGC	-1.68	12	CCGC	-1.60	27	ATAT	-1.59	28
	CCCC	-1.47	15	ATT	-1.39	102	TCAA	-1.28	15
	ACAT	-1.14	23	TTAA	-1.05	22	ATTA	-1.04	39
H3K4me3	CGGC	-3.44	12	CCCC	-3.26	15	CCGC	-3.18	27
	GGGG	-2.69	18	GCGG	-2.28	18	TCGG	-2.06	12
	GCCG	-2.06	26	CGTC	-1.94	16	CAAA	-1.68	20
H3K36me3	CTTA	-2.08	15	TCGC	-2.03	7	TAAG	-1.96	13
	TCTC	-1.80	16	GCCA	-1.69	11	TATA	-1.66	21
	GATC	-1.63	7	CCCT	-1.60	11	GCG	-1.54	53
H3K79me3	TATA	-3.90	21	CGC	-2.73	51	GCG	-2.57	53
	CGCG	-2.01	18	AAAA	-1.89	32	TTTT	-1.80	33
	ATGT	-1.76	24	CCC	-1.65	50	CATG	-1.57	11

3.2 Genetic Area Preferences of Histone Occupancy, Acetylation, and Methylation

We used the SVM with a linear kernel to rank features based on their support for histone occupancy, acytelation, and methylation (see Section 2.4). Tables 4 and 5 show the most informative positive and negative features, respectively, from a set of 3-gram and 4-gram features, together with their support for histone occupancy, acetylation, and methylation. Similarly, tables 6 and 7 show the most informative positive and negative features from a set of 4-gram and 5-gram features, together with their support for histone occupancy, acetylation, and methylation. The full and detailed information on the ranking of features is available from the supplementary files.

The most informative features from the set of 3,4-grams (Tables 4 and 5) and from the set of 4,5-grams (Tables 6 and 7) are consistent with each other. As can be seen, CCTG and CAG appear in both tables of the most informative features to recognize H3 histone occupancy. Also, CCA, TGG, TATC, and GGA are in the most informative positive features for H4 histone occupancy from both the set of 3,4-grams and the set of 4,5-grams. Therefore, our method is self-validated and good for this feature selection task.

The informative features to discriminate positive from negative classes produced by our method will

Table 6: Most informative features for positive class, from a set of 4-gram and 5-gram features. TF is the number of transcription factor binding motifs containing the feature.

	Feature	Weight	TF	Feature	Weight	TF	Feature	Weight	TF
H3	CCTG	1.98	13	GCATT	1.70	4	AGTGC	1.60	3
	ACCTG	1.58	5	TCCTG	1.56	1	CTCTT	1.55	5
	TTCAC	1.54	4	GGGTT	1.54	8	ACAGC	1.52	4
H4	CCAGT	1.85	4	TCAGG	1.73	1	TGGA	1.65	11
	TTGGG	1.55	2	CGTTA	1.51	3	TATC	1.49	15
	GGAT	1.49	14	GGATC	1.48	1	TATTG	1.47	5
H3K9ac	CCGG	2.04	22	ACCCG	1.97	6	GCCCA	1.87	3
	GGTGG	1.78	2	CAGCG	1.77	3	GCGAG	1.76	7
	CCGGG	1.71	8	GCCGG	1.71	7	GGCCG	1.66	5
H3K14ac	GCGTG	2.21	1	ACCCG	2.06	6	TAGTC	2.02	2
	GGTGG	1.85	2	ATGAG	1.85	1	GGGTA	1.85	5
	CTCGA	1.84	1	TAGGA	1.72	4	AAGGC	1.70	1
H4ac	GCCGG	2.42	7	CTCAT	2.35	4	ACCCG	2.35	6
	GGGTA	1.96	5	ATGAG	1.90	1	GGTGG	1.90	2
	ACCGC	1.89	5	GCGA	1.87	15	TCGC	1.86	7
H3K4me1	GGGTC	2.31	2	CGGGA	1.96	2	AATTC	1.93	4
	CGGAG	1.87	3	AGTTT	1.75	2	TCGTA	1.71	1
	AACCC	1.68	4	ACCCA	1.63	13	ATCGA	1.62	0
H3K4me2	ACCAC	2.23	1	GTTGC	2.18	2	GCGAG	2.03	7
	CCAGC	1.97	4	CACTT	1.94	1	ACCGC	1.90	5
	ATTCC	1.73	5	GAGGG	1.68	3	GTCCA	1.64	1
H3K4me3	ACCCG	3.39	6	CGGGT	2.18	6	ATGAG	2.13	1
	TGCGA	2.06	3	GGTGG	1.92	2	CTCGC	1.91	4
	CTCAT	1.90	4	GTTGC	1.84	2	GCGA	1.82	15
H3K36me3	GGCGA	1.82	3	CGTCC	1.72	3	ACCCA	1.68	13
	GGAAC	1.59	1	AATTC	1.58	4	GATTT	1.58	5
	CAACG	1.51	7	TGGAT	1.48	3	CGCAG	1.48	0
H3K79me3	GATTT	2.09	5	TAATG	1.85	3	CATTA	1.81	6
	GCGTA	1.66	1	CTCTA	1.63	5	ACAGC	1.55	4
	CAGCA	1.54	1	CAATA	1.47	7	TAGGG	1.47	5

be useful to analyze the genetic area preferences of histone occupancy, acetylation, and methylation. For example, CG (CpG) is a dinucleotide that appears very often in the most informative features of DNA sequence areas that are neither occupied by histones nor acetylated or methylated (Tables 5 and 7). This agrees with previous studies that show CpG islands are mostly non-methylated [3].

4 Conclusion

We have introduced the use of support vector machines to qualitatively predict histone occupancy, acetylation, and methylation areas in DNA sequences. The method is the first one to date for this kind of problem, and empirical results show a high accuracy. Moreover, the support vector machine method can evaluate the informative features to discriminate between DNA areas with high and low occupancy, acetylation, or methylation. In the near future, we will improve the method to give a prediction confidence for each DNA area according to the *margin* of the corresponding feature example to the classification hyperplane of the support vector machine.

Table 7: Most informative features for negative class, from a set of 4-gram and 5-gram features. TF is the number of transcription factor binding motifs containing the feature.

	Feature	Weight	TF	Feature	Weight	TF	Feature	Weight	TF
H3	CGCG	-4.57	18	GCGC	-2.94	11	GCGCG	-2.41	2
	CGGGC	-2.21	3	GGCCG	-2.11	5	GCGG	-2.05	18
	CGCGC	-2.01	3	GCGGC	-1.87	4	GCAGC	-1.67	1
H4	CGCG	-3.46	18	GCGC	-2.42	11	TTTTT	-2.31	8
	AAAAA	-1.96	13	TATA	-1.93	21	CGTG	-1.88	22
	CCCGG	-1.82	6	GCGCG	-1.70	2	TAATT	-1.64	18
H3K9ac	GCCGC	-3.45	16	CCATA	-2.03	5	ACCCC	-2.00	2
	AAACC	-1.97	5	GCGGC	-1.86	4	CAATA	-1.84	7
	GGCGG	-1.76	5	GATTT	-1.74	5	TGTAA	-1.67	5
H3K14ac	GCCGC	-3.30	16	ACCCC	-2.13	2	TACCA	-1.88	1
	GACGT	-1.87	6	TGGTA	-1.85	2	AATTC	-1.80	4
	TCTAA	-1.78	2	GATCC	-1.76	2	CTCGG	-1.66	2
H4ac	GCCGC	-4.95	16	GCGGC	-2.61	4	ACCCC	-2.35	2
	GTTGT	-1.92	1	GCGTC	-1.92	5	GTGGG	-1.88	5
	AATTC	-1.78	4	TACGA	-1.78	3	TACCA	-1.77	1
H3K4me1	GGGTA	-2.20	5	TCCTA	-2.05	6	TACAC	-2.01	5
	TGCGA	-1.94	3	TGTCC	-1.87	0	ACCCG	-1.85	6
	CTCGA	-1.83	1	TACCC	-1.82	6	ATCGC	-1.80	0
H3K4me2	GCCGC	-2.73	16	CCGCC	-2.32	13	GGTGC	-2.13	3
	GACCG	-2.10	2	GGGAT	-1.93	2	CGCGG	-1.82	4
	GTTCC	-1.82	3	ACCGG	-1.64	4	CCGAG	-1.63	4
H3K4me3	GCCGC	-4.27	16	ACCCC	-2.74	2	CGCGG	-2.62	4
	CACGC	-2.15	3	GCGTC	-2.13	5	GTGGG	-2.09	5
	GCGGC	-2.05	4	CCGC	-2.03	27	AGCCG	-1.98	12
H3K36me3	TAGGA	-2.58	4	ACCCG	-2.39	6	CTCGA	-2.32	1
	AACCG	-2.13	5	CCCTT	-1.82	3	CTCAT	-1.68	4
	CATCG	-1.60	3	ATGCG	-1.59	6	GTATA	-1.53	2
H3K36me3	CGCG	-2.36	18	GCGC	-2.22	11	TATA	-2.22	21
	ACATA	-2.20	2	ACGTA	-2.19	3	TAATT	-2.06	18
	ATGTA	-1.92	8	AATTA	-1.91	21	TACAT	-1.80	10

Acknowledgments

The research described in this paper was partially supported by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, by COE project JCP KS1 of the Japan Advanced Institute of Science and Technology, by the Spanish CICYT, project GRAMMARS (TIN2004-07925-C03-01), and by the Japan Society for the Promotion of Science through Long-term Invitation Fellowship L05511 for visiting the Japan Advanced Institute of Science and Technology.

References

[1] Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P. Liu, J. S., Kouzarides T., and Schreiber, S. L., Methylation of histone H3 Lys 4 in coding regions of active genes, *Proc. Natl. Acad. Sci. USA*, 99(13):8695–8700, 2002.

[2] Bernstein, B. E., Liu, C. L., Humphrey, E. L., Perlstein, E. O., and Schreiber, S. L., Global nucleosome occupancy in yeast, *Genome Biol.*, 5(9):R62, 2004.

[3] Bird, A. and Boyes, J., The essentials of DNA methylation, *Cell*, 70(1):5–8, 1992.

[4] Brank, J., Grobelnik, M., Milic-Frayling N., and Mladenic, D., Feature selection using support vector machines, *Proc. 3rd Int. Conf. Data Mining Methods and Databases for Engineering, Finance and Other Fields*, 261–273, 2002.

[5] Cristianini, N. and Taylor, J. S., *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[6] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Selection for cancer classification using support vector machines, *Machine Learning*, 46(1/3):389–422, 2002.

[7] Kouzarides, T., Histone methylation in transcriptional control, *Curr. Opin. Genet. Dev.*, 12(2):198–209, 2002.

[8] Kurdistani, S. K., Tavazoie, S., and Grunstein, M., Mapping global histone acetylation patterns to gene expression, *Cell*, 117(6):721–733, 2004.

[9] Lange, K., *Optimization*, Springer-Verlag, 2004.

[10] Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D., and Lieb, J. D., Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nat. Genet.*, 36(8)900–905, 2004.

[11] Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J., Crystal structure of the nucleosome core particle at 2.8 Å resolution, *Nature*, 389(6648):251–260, 1997.

[12] Narlikar, G. J., Fan, H. Y., and Kingston, R. E., Cooperation between complexes that regulate chromatin structure and transcription, *Cell*, 108(4):475–487, 2002.

[13] Peterson, C. L. and Laniel, M. A., Histones and histone modifications, *Curr. Biol.*, 14(14):R546–R551, 2004.

[14] Pham, T. H., Satou, K., and Ho, T. B., Support vector machines for prediction and analysis of beta and gamma-turns in proteins, *J. Bioinform. Comput. Biol.*, 3(2):343–358, 2005.

[15] Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K., and Young, R. A., Genome-wide map of nucleosome scetylation and methylation in yeast, *Cell*, 122(4):517–527, 2005.

[16] Ren, B., Robert, F., *et al.*, Genome-wide location and function of DNA binding Proteins, *Science*, 290(5500):2306–2309, 2000.

[17] Robyr, D., Suka, Y., Xenarios, I., Kurdistani, S. K., Wang, A., Suka, N., and Grunstein. N., Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases, *Cell*, 109(4):437–446, 2002.

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