Prediction of Histone Modifications in DNA sequences

Tho Hoan Pham
Hanoi National University of Education
136 Xuan-Thuy, CauGiay, Hanoi, Vietnam
h-pham@jaist.ac.jp

Tu Bao Ho
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Japan
bao@jaist.ac.jp

Dang Hung Tran
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Japan
hungtd@jaist.ac.jp

Kenji Satou
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Japan
ken@jaist.ac.jp

Abstract

DNA molecules are wrapped around histone octamers to form nucleosome structures whose occupancy and histone modification states profoundly influence the gene expression. Depending on the DNA segment that a nuleosome incorporated, its histone proteins exihibit paticular modifications by added some functional chemical groups to specific amino acids. The key approach up to now to determining the DNA locations of histone occupancy as well as histone modifications is an experimental technique called ChiP-Chip, or Chromatin Immunoprecipitation on Microarray Chip. This experimental technique has some disadvantages such as it is tedious, wastes time and money, produces noise, and cannot provide results at an arbitrarily high resolution, especially with large genomes like human's. We have developed a computational method to determine qualitatively histone-occupied as well as acetylation and methylation locations in DNA sequences. The method is based on support vector machines (SVMs) to learn models from training data sets that discriminate between areas with high and low levels of histone occupancy, acetylation or methylation. Our computational method can give quickly the prediction at any position in a DNA sequence based on the content and context of the subsequence around that position. The prediction results on the yeast genome by three-fold cross-validation showed high accuracy and were consistent with the ones from experimental methods. Moreover, SVM-classification models in our method can present genetic preferences of DNA areas that have high modification levels.

Keywords: histone modifications, acetylation, methylation, support vector machine (SVM)

1 INTRODUCTION

Eukaryotic genomes are packaged into nucleosomes that consist of 145–147 base pairs of DNA wrapped around a histone octamer (two each of histones H2A, H2B, H3 and H4) [13]. Histone octamers are identical for all nucleosomes in all DNA sequences of a species, but 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 conserved sites in histone H2A and H2B; methylation sites include H3K4, H3K9, H3K27, H3K36, H3K79, H3R17, H4K20, H4K59, and H4R3 [16]. When a nucleosome appears in a specific DNA sequence area, these potential sites can have a certain acetylation or methylation level [10, 19].

The histone components of nucleosomes and their modification states (of which acetylation and methylation are the most important ones) can profoundly influence many genetic activities, including transcription [1, 9, 10, 19], DNA repair, and DNA remodeling [15, 12]. There is recently an explosion of studies on the histone modifications in nucleosomes as well as on the relationship between them and gene expression [4, 8, 10, 12, 19]. The key approach to determining the DNA locations of histone occupancy as well as histone modifications in these studies is an experimental technique called ChiP-Chip, or Chromatin Immunoprecipitation on Microarray Chip [3]. This experimental technique has some disadvantages such as it is tedious, wastes time and money, produces noise, and cannot provide results at an arbitrary high resolution especially with large genomes like human.

1-4244-1509-8/07/$25.00 ©2007 IEEE

In this paper, we have developed a computational method to locate qualitatively histone-occupied as well as acetylation and methylation positions in DNA sequences. The method is based on support vector machines (SVMs) to learn models from training data sets that discriminate between areas with high and low levels of histone occupancy, acetylation or methylation. The prediction results on the yeast genome by three-fold cross-validation showed high accuracy and were consistent with experimental methods. Moreover, SVM-classification models in our methods can present genetic preferences of areas with high or low modification levels.

2 METHODS

2.1 Vectorization of sequences

The histone occupancy and modification states at each position in a DNA sequence are, in our work, assumed to be influenced by two factors: (1) the subsequence of a length L equally expanding both sides from the position; and (2) genetic elements such as promoters, the begin and the end of genes, etc, around it. These two kinds of information should be represented by a numerical vector of features.

Each L subsequence can be represented by a set of k-gram features (called content-based features). We use a k-sliding window along a DNA subsequence to compute the number of occurrences of each k-gram. Each subsequence is thus represented by a 4^k-dimensional vector of the number of occurrences of all possible k-grams.

Another kind of information that influences the histone occupancy and their modifications is genetic elements in DNAs. To capture this kind of information around the predicted position, we use four context-based features which measure how far it is from the begin and end of the nearest SGD-annotated genes [5]. These four features are defined explicitly as follows:

$$f_{bp}(position) = \begin{cases} 0 \text{ if the distance } (d_{bp}) \text{ between the position and the } \textit{begin} \text{ of the nearest gene in the } \textit{prime} \text{ DNA strand} > 500 \\ \frac{500-d_{bp}}{500} \text{ if } d_{bp} \leq 500 \end{cases}$$

$$f_{bc}(position) = \begin{cases} 0 \text{ if the distance } (d_{bc}) \text{ between the position and the } \textit{begin} \text{ of the nearest gene in the } \textit{complementary} \text{ DNA strand} > 500 \\ \frac{500-d_{bc}}{500} \text{ if } d_{bc} \leq 500 \end{cases}$$

$$f_{ep}(position) = \begin{cases} 0 \text{ if the distance } (d_{ep}) \text{ between the position and the } \textit{end} \text{ of the nearest gene in the } \textit{prime} \text{ DNA strand} > 500 \\ \frac{500-d_{ep}}{500} \text{ if } d_{ep} \leq 500 \end{cases}$$

$$f_{ec}(position) = \begin{cases} 0 \text{ if the distance } (d_{ec}) \text{ between the position and the } \textit{end} \text{ of the nearest gene in the } \textit{complementary} \text{ DNA strand} > 500 \\ \frac{500-d_{ec}}{500} \text{ if } d_{ec} \leq 500 \end{cases}$$

2.2 Binary support vector machines

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, to map the input vector x_i to a vector $\phi(x_i)$ in a richer feature space, either linearly or nonlinearly, which is relevant to the selection of the kernel function. Second, to obtain an optimized linear division within the feature space from the first step, that is, to 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:

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

Its dual is a quadratic optimization problem:

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

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 SVMs classification function, once trained, has the following form:

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

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 [11], 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*.

SVMs 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 SVMs can be found in [6, 20].

2.3 SVMs with confidence

Since the SVMs prediction of a vector x is based only on the information that which of two "semi-spaces" formed by the SVM hyperplane f the vector belongs to (i.e. whether the value of function f is greater than 0 or not). The prediction does not care the distance (or margin) between the vector and the SVM classification hyperplane, which is an important factor to infer the confidence of the prediction. The larger the margin the more confident the prediction is. So, we improve the support vector classification by adding some confidence to the prediction of SVM. The predictive confidence of a vector x is calculated as follows:

$$conf(x) = \begin{cases} 0 \text{ if } f(x) \leq T_1 \\ \frac{f(x)-T_1}{T_2-T_1} \text{ if } T_1 \leq f(x) \leq T_2 \\ 1 \text{ if } f(x) \geq T_2 \end{cases} \qquad (2)$$

where T_1 and T_2 are thresholds to decide the least and the most of confidences. In our work, T_1 and T_2 are determined from examples $x_1, \ldots, x_n$ in the training set as follows: calculate all absolute values of $|f(x_i)|, i = 1, \ldots, n$; sort them in ascending order; let $T_1 = |f(x_{[\frac{n}{20}]})|$ and $T_2 = |f(x_{n-[\frac{n}{20}]})|$, or in other words, there are 5% of examples will have a confidence of 0 and 5% will have a confidence of 1.

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 [2, 7, 17]. When SVMs 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):

$$f(X = (f_1, f_2, \ldots, f_m)) = \sum_{i=1}^{m} w_i f_i + b . \qquad (3)$$

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.

2.5 Datasets

From the genome-wide map of histone occupancy, acetylation and methylation locations reported in [19] by the ChiP-Chip method, we selected 10 datasets and used them to illustrate the performance of our method. These datasets are described in detail in Table 1. Among them, two datasets H3 and H4 are relative occupancy data of histone H3 and H4, respectively. Two datasets H3K9ac, H3K14ac are relative histone acetylation data of specific lysines K9 and K4 of histone H3; and dataset H4ac presents relative general acetylation of histone H4. Three (mono-, di-, and tri-) methylation states of lysine K4 of histone H3 are contained in three datasets H3K4me1, H3K4me2, and H3K4me3. Finally, two other tri-methylation datasets are measured at different lysines of H3: K36 and K79.

Each example in the datasets corresponds to a point in DNAs that has the experimental data published in [19]. Since the data of nucleosome occupancy and histone modifications is relative measures, so we assign locations with the relative data greater than 1.0 into the positive class; otherwise, the negative class. We assume that the relative occupancy of nucleosomes as well as their modifications are influenced by two factors: (1) the subsequence (segment) with a fixed length L, called L-subsequence; and (2) around genetic elements. These two kinds of information will be converted into vectors as described in 2.1.

Table 1. Datasets

Dataset	Full name	POS	NEG
H3	H3.YPD	25,137	16,069
H4	H4.YPD	25,924	15,299
H3K9ac	H3K9acvsH3.YPD	18,726	22,241
H3K14ac	H3K14acvsH3.YPD	21,535	19,639
H4ac	H4acvsH3.YPD	20,402	20,839
H3K4me1	H3K4me1vsH3.YPD	20,344	20,401
H3K4me2	H3K4me2vsH3.YPD	21,440	19,493
H3K4me3	H3K4me3vsH3.YPD	20,686	20,508
H3K36me3	H3K36me3vsH3.YPD	20,826	20,296
H3K79me3	H3K79me3vsH3.YPD	22,574	18,536

Datasets of histone occupancy, acetylation, and methylation by ChiP-Chip protocol in vivo [19]. POS and NEG are the number of positive and negative examples, respectively.

3 RESULTS

3.1 Prediction results

We developed a machine learning method based on support vector machines to qualitatively predict locations of histone occupancy and modification states given a DNA sequence. The prediction for each point is based on two kinds of features: (1) k-gram features of the subsequence of the length L equally expanding to both sides of the DNA sequence, and (2) the information of genetic elements around that point. Both are converted into vectors (see Section 2.1) before inputting to support vector machines. We improved support vector machines by introducing the confidence of the prediction based on the distance between the predicted vector and the SVM classification hyperplane (Section 2.3). The prediction with a higher confidence would be more accurate.

We illustrated the performance of our method with 10 datasets (see Section 2.5) collected from experimental data at many locations in the yeast genome by the work of Pokholok *et al.* [19]. We follow the three-fold cross-validation procedure on each of 10 datasets and use two performance criteria of accuracy (acc) and correlation coefficient (cc) to report the prediction results [1].

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

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

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

There are two kinds of parameters of the SVM-based model in our problem that we should estimate from training datasets such that the model would predict best on testing datasets. The first is SVM parameters, i.e. the kernel type, kernel width, the penalty C for an error classification. The second is parameters k (in k-grams) and L (in L-subsequences) that are used to represent a DNA location to numerical vectors (see Section 2.1). First we fixed some values of k and L parameters and did various experiments with different values of SVM parameters. We found that SVM models reached the best prediction results with RBF kernel, kernel width 0.01 and C=1. These best SVM parameters were then fixed and we tried to find out how sensitive the parameters k and L are on prediction results (see Tables 2 and 3).

The best prediction accuracies of both H3 and H4 histones occupancy are 74.60 and 77.12, respectively, when using features of short k-grams (k=5 or 6, see Table 2) and depends on a subsequence of a short length (L=500, Table 3) around the predicted point. This reveals that nucleosomes formation (of which H3 and H4 are the main components) is not influenced by long-ranged DNA elements. However, their modifications (acetylation and methylation states) are depended on subsequences with the longer length (L=1000, see Table 3). Of which, acetylation and tri-methylation states are often easier to predict. The prediction accuracies for those states can be greater than 80% except H3K9 acetylation state (Table 3).

Moreover, our method can introduce the confidence of the prediction which are very accurate on subsequences where the confidence is greater than a certain threshold (min_conf). Table 4 showed prediction accuracies with different thresholds min_conf: 0.0 (corresponding to normal SVMs), 0.25, 0.50, and 0.75. As can be seen, the average prediction accuracy on 10 datasets increases dramatically from 78.73% at $min\_conf = 0.0$ (with the predicting coverage is 100%) up to 93.19% at $min\_conf = 0.75$ (but the predicting coverage is droped to 24.7%). Clearly, the prediction with more confidence is more accurate, but the predicting coverage of the prediction is of course smaller.

3.2 Informative features

We used SVMs with a linear kernel to rank features based on their support for histone occupancy, acytelation, and methylation (see Section 2.4). Tables 5 and 6 show

[1] the dataset is partitioned into 3 subsamples. Of the 3 subsamples, a single subsample is retained as the validation data for testing the SVM model that was trained on the remaining 2 subsamples. The cross-validation procedure is repeated 3 times (the folds), with each of the 3 subsamples used exactly once as the validation data. The 3 results (i.e. acc and cc) from the folds then is then averaged to produce a single estimation

Table 2. Prediction results depending on features of k-grams

Dataset	$k=4$		$k=5$		$k=6$		$k=7$		$k=8$		$k=9$		$k=10$	
	acc	*cc*	*acc*	*cc*	*acc*	*cc*	*acc*	*cc*	*acc*	*cc*	*acc*	*cc*	*acc*	*cc*
H3	74.06	0.44	**74.60**	**0.45**	74.29	0.45	74.35	0.45	73.78	0.43	71.90	0.39	68.68	0.32
H4	76.42	0.48	77.05	0.49	**77.12**	**0.50**	76.86	0.49	75.54	0.46	73.55	0.41	70.33	0.33
H3K9ac	68.00	0.35	69.95	0.39	71.65	0.43	73.13	0.46	74.26	0.48	**74.28**	**0.49**	71.61	0.44
H3K14ac	66.82	0.33	68.91	0.38	71.54	0.43	73.89	0.48	76.47	0.53	78.93	0.58	**80.34**	**0.61**
H4ac	66.83	0.34	68.67	0.37	71.07	0.42	73.90	0.48	75.90	0.52	78.25	0.56	**79.19**	**0.59**
H3K4me1	64.65	0.29	66.54	0.33	68.16	0.36	69.73	0.39	71.02	0.42	72.61	0.45	**73.55**	**0.47**
H3K4me2	62.19	0.24	64.22	0.28	66.33	0.32	67.65	0.35	69.45	0.39	**70.96**	**0.42**	70.87	0.42
H3K4me3	63.11	0.26	66.21	0.32	69.21	0.38	72.64	0.45	75.57	0.51	78.24	0.57	**79.95**	**0.60**
H3K36me3	70.37	0.41	71.83	0.44	74.16	0.48	75.61	0.51	76.91	0.54	77.85	0.56	**78.59**	**0.58**
H3K79me3	73.87	0.47	75.30	0.50	76.13	0.52	**76.87**	**0.53**	76.54	0.53	75.25	0.50	73.17	0.47

Results of acetylation and methylation prediction $acc(cc)$ from a set of k-gram features ($L = 500$). Both the accuracy (*acc*) and the correlation coefficient (*cc*) are shown.

Table 3. Prediction results depending on different L

Dataset	$L=300$		$L=500$		$L=1000$		k
	acc	*cc*	*acc*	*cc*	*acc*	*cc*	
H3	71.18	0.37	**74.60**	**0.45**	74.17	0.44	5
H4	74.27	0.43	**77.12**	**0.50**	74.18	0.43	6
H3K9ac	61.5	0.24	74.28	0.49	**75.52**	**0.51**	9
H3K14ac	60.91	0.21	80.34	0.61	**82.90**	**0.66**	10
H4ac	60.79	0.22	78.25	0.56	**82.24**	**0.64**	9
H3K4me1	61.18	0.23	72.61	0.45	**74.65**	**0.50**	9
H3K4me2	59.54	0.18	70.96	0.42	**74.65**	**0.49**	9
H3K4me3	59.58	0.19	79.95	0.60	**82.96**	**0.66**	10
H3K36me3	64.60	0.29	78.59	0.58	**82.27**	**0.65**	10
H3K79m3	67.88	0.35	76.54	0.53	**80.29**	**0.60**	8

Results of acetylation and methylation prediction $acc(cc)$ with different values of L's. The parameter k (in k-grams) is fixed, corresponding to the best cases in Table 2. Both the accuracy (*acc*) and the correlation coefficient (*cc*) are shown.

Table 4. The prediction accuracy (*acc*) and coverage (*cover*) at different confidence levels (*conf*)

Dataset	$conf \geq 0.00$		$min \geq 0.25$		$conf \geq 0.50$		$conf \geq 0.75$		Params	
	acc	*cover*	*acc*	*cover*	*acc*	*cover*	*acc*	*cover*	*L*	*k*
H3	74.60	100	82.06	68.55	87.65	41.62	92.50	18.00	500	5
H4	77.12	100	83.80	70.89	87.92	44.59	90.54	19.87	500	6
H3K9ac	75.52	100	82.63	67.78	86.84	41.07	89.85	18.90	1000	9
H3K14ac	82.90	100	90.38	74.82	94.59	52.87	96.95	28.12	1000	10
H4ac	82.24	100	89.44	74.87	93.39	53.69	95.55	28.60	1000	9
H3K4me1	74.75	100	82.28	69.78	87.48	46.14	90.86	23.75	1000	9
H3K4me2	74.65	100	81.43	71.09	86.22	48.13	88.61	25.17	1000	9
H3K4me3	82.96	100	90.70	74.57	94.84	52.80	96.99	28.21	1000	10
H3K36me3	82.27	100	89.45	74.78	93.48	52.67	95.51	27.96	1000	10
H3K79me3	80.29	100	87.44	74.90	91.90	53.75	94.57	28.45	1000	8
Average	**78.73**	**100**	**85.96**	**72.20**	**90.43**	**48.73**	**93.19**	**24.70**		

the most informative positive and negative features, respectively, from a set of 4-gram and context-based features, together with their support for histone occupancy, acetylation, and methylation.

Context-based features *bp*, *bc*, *ep*, and *ec* (see Section 2.1) are often among the most informative features to discriminate two classes of low and high occupancy/acetylation/methylation subsequences (Tables 5 and 6). Areas near to starting sites of genes were often acetylated since the features *bp* and *bc* strongly support all H4, H3K9 and H3K14 acetylation (Table 5). The end of genes were, vice versa, often non-acetylated since the features *ep* and *ec* were consistently the most negative ones for all H4, H3K9, and H3K14 acetylation (Table 6).

Different from consistent acetylation pattern which is always related to the begin and the end of genes, each methylation pattern depends on both different state (mono, di, tri-) of methylation and amino acid position in histone proteins. While H3K4 is often mono-methylated around the end of genes and non-mono-methylated at the promoter of genes, its di- and tri-methylation are conversely often occurred at the promoter of genes (Tables 5 and 6) and not occurred at the end of genes. The tri-methylation of H3K4, K36 and K79 are also different.

The most informative feature to recognize non-histone-occupied (both H3 and H4) areas is CGCG (Table 6). This suggests that the CpG-rich areas are often nucleosome-low.

4 DISCUSSTION

4.1 Prediction of histone occupancy, acetylation and methylation

Up to now, an experimental technique called ChiP-Chip (Chromatin Immunoprecipitation on Microarray Chip) [3] is still the most favourable technique to determine the DNA locations of histone occupancy as well as histone modifications in these studies [10, 12, 19]. This experimental technique has some disadvantages such as it is tedious, wastes time and money, produces noise, and cannot provide results at an arbitrary high resolution especially with large genomes like human. With this work, we offered a computational approach to indicate DNA locations where are nucleosome-occupied and contain modification states. Though our method could not determine quantitatively like the experimental method, it can show very accurately some areas of DNAs that have or have not nucleosomes as well as their modifications. For example, the average prediction accuracy on 10 datasets of Histones occupancy and modifications is up to 93.19% for 24.7% locations of DNAs where the prediction confidence are greater than 0.75 (see Table 4). The qualitative prediction of histone occupancy, acetylation and methylation from our method would be useful for experimental studies as guidances to focus or ignore some areas in DNA sequences.

The work in this paper is the continuing one of our conference paper [18]. There are three basic improvements here: (1) we used additional context-based features which capture the information of genetic elements around the predicted location in the DNA sequence (see Section 2.1); (2) we did investigate more deeply so that the best parameters of k-gram features have been found (Tables 2, 3, and 4); and (3) we improved the SVMs method to introduce some confidence for SVM prediction (Section 2.3), the prediction with higher confidence is often with higher accurate (Table 2). Therefore prediction results at this time are much better than previously-published ones.

The performance of our computational method is evaluated on the experimental data (microarray data), which is often noisy due to the present technology problems. Hence the prediction accuracies reported in this paper might be under-estimated since we have computed the accuracies by comparison between the predicted results and noisy experimental ones.

4.2 Genetic preferences of histone occupancy, acetylation, and methylation

Informative content-based and context-based features to discriminate two classes of high and low histone occupancy/acetylation/methylation areas are useful to uncover genetic preferences of these areas. For example, we found that both H3 and H4-occupancies in CpG-rich areas are low (Table 6). This agrees with previous studies: CpG islands are often nucleosome-free [1, 14].

The context-based features have been often found in the most informative features of positive or negative acetylation and methylation subsequences. This confirms that there exists a relationship between acetylation/methylation and gene expression. The information of the distance of a location and the nearest genes in both prime and complementary DNA strands (*bp*, *bc*, *ep*, and *ec*) are often most important to know acetylation and methylation of that location. We found that locations near to gene starting sites are often acetylated (Table 5) and those near to the end of genes are non-acetylated (Table 6), which consist with previous results (see [14] and references therein). We also reported that the begin area of genes (with features *bp* or *bc*) are H3K4-di- and tri-methylated (but not mono-methylated) as well as H3K9, H3K14 and H4 acetylated (Tables 5 and 6). This agrees with the evidence in the work of Metthew et al. (2004) that there is a relation between H3K4 di&tri-methylation and acetylation of H3K9, H3K14.

Table 5. Most informative features for positive class from 4-gram and context-based features

	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
H3	bc	0.27	TCGC	0.25	GCTC	0.24	CAGG	0.22	bp	0.21
	GTAC	0.21	CCTC	0.21	CCGC	0.19	TAGG	0.19	TCGG	0.18
H4	CAAA	0.16	TGAC	0.13	CCTC	0.12	TGAT	0.12	CTGG	0.12
	GCTC	0.12	AGCC	0.11	ACTC	0.10	GTGG	0.10	TGAA	0.10
H3K9ac	bc	0.45	bp	0.34	TCCG	0.33	TCCC	0.31	ACCG	0.30
	TCCT	0.29	GTAC	0.27	TCGA	0.26	CGTG	0.26	GCGA	0.25
H3K14ac	bc	0.81	bp	0.75	CCGG	0.29	ACGG	0.29	GCGG	0.28
	TCGG	0.24	ACCG	0.23	CTAC	0.21	AAGA	0.21	AGGA	0.19
H4ac	bc	0.84	bp	0.59	ACCG	0.36	ACCC	0.30	CGTG	0.28
	TCCG	0.27	GTAC	0.27	CTAC	0.24	TCCC	0.23	ACGG	0.22
H3K4me1	ec	0.50	ep	0.42	CCCA	0.31	GCCG	0.27	CCCG	0.26
	GCCA	0.24	CCCT	0.23	GCCT	0.19	GATG	0.19	ACGC	0.18
H3K4me2	CCCA	0.32	CCCG	0.30	GGCG	0.26	bp	0.25	TCGG	0.23
	CAGC	0.23	ACGG	0.23	bc	0.22	GTGT	0.22	ACCA	0.22
H3K4me3	bc	1.01	bp	0.81	ACCG	0.30	CCCG	0.28	CGCC	0.24
	GTAC	0.22	CACA	0.21	CTCA	0.21	GTGG	0.19	CACG	0.19
H3K36me3	ec	0.42	ep	0.40	TACC	0.21	ACAC	0.18	GCCT	0.18
	CTTC	0.17	TTGA	0.17	ACCT	0.16	GAAA	0.15	GACC	0.15
H3K79me3	GCGT	0.17	CGCA	0.17	ACAC	0.17	TCGT	0.15	CAGG	0.15
	GGGA	0.14	TCCC	0.14	CCAC	0.14	TCTT	0.14	GGGC	0.14

Most informative features for positive class from a set of 4-gram and context-based features.

Table 6. Most informative features for negative class from 4-gram and context-based features

	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
H3	CGCG	-0.36	CGCT	-0.24	CGGA	-0.23	CGTG	-0.22	AGGA	-0.21
	CGGG	-0.20	ep	-0.20	CGGC	-0.18	CGCC	-0.18	CGTA	-0.17
H4	GCGC	-0.16	GCCG	-0.15	CGCG	-0.15	CGGG	-0.15	GTAA	-0.14
	bc	-0.13	TACA	-0.11	GACA	-0.11	TATA	-0.11	GTAG	-0.11
H3K9ac	CTCC	-0.32	TACC	-0.31	ATCC	-0.28	GTCC	-0.26	CCGC	-0.24
	TTGA	-0.22	CCGT	-0.22	CACC	-0.20	TTCC	-0.19	GTCG	-0.17
H3K14ac	CGGC	-0.44	ep	-0.42	ec	-0.36	CGGG	-0.31	CGGT	-0.29
	CGGA	-0.28	CACC	-0.26	AGAA	-0.26	TACC	-0.25	CCGT	-0.23
H4ac	CCGT	-0.41	CCGC	-0.34	TACC	-0.32	ep	-0.30	ec	-0.28
	CGGC	-0.27	CACC	-0.26	CGGG	-0.23	CGGA	-0.22	TGGG	-0.21
H3K4me1	bc	-0.77	bp	-0.73	TCCC	-0.26	AGGG	-0.25	ACCC	-0.24
	AGGA	-0.21	CGCC	-0.20	TGAC	-0.19	CGCA	-0.19	TGAT	-0.18
H3K4me2	GCGT	-0.39	ACAG	-0.28	AGCA	-0.27	CCGT	-0.27	CGGC	-0.27
	GCGC	-0.25	TCCC	-0.25	ACAT	-0.24	CGGG	-0.24	CCAG	-0.23
H3K4me3	ep	-0.47	ec	-0.40	CCGC	-0.39	CTTC	-0.28	CCGT	-0.26
	ACTC	-0.25	ACGC	-0.24	GCGC	-0.23	TACC	-0.21	GGAT	-0.21
H3K36me3	bc	-0.70	bp	-0.67	CCTG	-0.26	CCTA	-0.26	CCTT	-0.25
	CACA	-0.23	CACG	-0.20	AAAG	-0.19	TAAG	-0.19	TTCG	-0.18
H3K79me3	ACGC	-0.27	AGGG	-0.23	AGCG	-0.23	CACA	-0.21	CGTA	-0.19
	CCGC	-0.18	ep	-0.17	GTCC	-0.17	ACTA	-0.17	ACGA	-0.16

Most informative features for negative class, from a set of 4-gram and context-based features.

5 CONCLUSIONS

Histone occupancy as well as modification states can be qualitatively predicted by computational models. Our computational method can give quickly the prediction at any position in a DNA sequence based on the content and context of the subsequence around that position. It is also used to extract informative characteristics of areas in DNA sequences with high or low histone modifications. The qualitative prediction of histone occupancy, acetylation and methylation from our method would be useful for experimental studies as guidances to focus or ignore some areas in DNA sequences.

ACKNOWLEDGEMENTS

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 Japan Advanced Institute of Science and Technology, project No. 2 002 06 of Ministry of Science and Technology of Vietnam. The first author has been supported by Vietnamese government scholarship from the Ministry of Education and Training of Vietnam to study in Japan. The third author has been supported by Japan government scholarship (Monbukagakusho) to study in Japan.

References

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

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

[3] M. J. Buck and J. D. Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83:349–360, 2004.

[4] M. J. Carrozza, B. Li, L. Florens, T. Syquanuma, S. K. Swanson, K. K. Lee, W. J. Shia, S. Anderson, J. Yates, M. P. Washburn, and J. L. Workman. Histone h3 methylation by set2 directs deacetylation of coding regions by rpd2s to suppress spurious intragenic transcription. *Cell*, 123(18):581–592, 2005.

[5] J. Cherry, C. Adler, C. Ball, S. Chervitz, S. Dwight, E. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, and D. Botstein. Sgd: Saccharomyces genome database. *Nucleic Acids Res.*, 26:73–79, 1998.

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

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

[8] M. C. Keogh, S. K. Kurdistani, S. A. Morris, and colleagues. Cotranscriptional set2 methylation of histone h3 lysine 36 recruits a repressive rpd3 complex. *Cell*, 123(18):593–605, 2005.

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

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

[11] K. Lange. *Optimization*. Springer Texts in Statistics. Springer-Verlag, 2004.

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

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

[14] C. L. Matthew, D. R. Dickerson, , M. Schmitt, and M. Groudine. Intragenic dna methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature Structural & Mole. Biology*, 11(11):1068–1075, 2004.

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

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

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

[18] T. H. Pham, D. H. Tran, T. B. Ho, K. Satou, and G. Valiente. Qualitatively predicting acetylation and methylation areas in dna sequences. In *Proc. 16th Int. Conf. International Conference on Genome Informatics*, pages 3–11, 2005.

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

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