Conditional Random Fields for Predicting and Analyzing Histone Occupancy, Acetylation and Methylation Areas in DNA Sequences

Dang Hung Tran[1], Tho Hoan Pham[2], Kenji Satou[1,3], and Tu Bao Ho[1,3]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
tran@jaist.ac.jp
[2] Faculty of Information Technology, Hanoi University of Pedagogy, 136 Xuan Thuy, Cau Giay, Hanoi, Vietnam
[3] Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST), Japan

Abstract. Eukaryotic genomes are packaged by the wrapping of DNA around histone octamers to form nucleosomes. Nucleosome occupancies together with their acetylation and methylation are important modification factors on all nuclear processes involving DNA. There have been recently many studies of mapping these modifications in DNA sequences and of relationship between them and various genetic activities, such as transcription, DNA repair, and DNA remodeling. However, most of these studies are experimental approaches. In this paper, we introduce a computational approach to both predicting and analyzing nucleosome occupancy, acetylation, and methylation areas in DNA sequences. Our method employs conditional random fields (CRFs) to discriminate between DNA areas with high and low relative occupancy, acetylation, or methylation; and rank features of DNA sequences based on their weight in the CRFs model trained from the datasets of these DNA modifications. The results from our method on the yeast genome reveal genetic area preferences of nucleosome occupancy, acetylation, and methylation are consistent with previous studies.

Keywords: Histone proteins, acetylation, methylation, conditional random fields.

1 Introduction

Eukaryotic genomes are packaged into nucleosomes that consist of 145–147 base pairs of DNA wrapped around a histone octamer [9]. 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, 4, 5, 16], DNA repair, and DNA remodeling [13].

There have been recently many studies of mapping histone occupancies together with their modifications in DNA sequences and of relationship between

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 221–230, 2006.

them and various genetic activities concerning DNAs [1, 2, 5, 7, 16, 18, 19]. But most of these studies were experimentally conducted by the combination of chromatin immunoprecipitation and whole-genome DNA microarrays, or ChIP-Chip protocol.

The nucleosome occupancy as well as its modifications such as acetylation and methylation mainly depend on the DNA sequence area they incorporate in. The majority of acetylation and methylation occurs at specific highly conserved residues in the histone components of nucleosomes: acetylation sites include at least nine lysines in histone H3 and H4 (H3K9, H3K14, H3K18, H3K23, H3K27, H4K5, H4K8, H4K12, and H4K16); methylation sites include H3K4, H3K9, H3K27, H3K36, H3K79, H3R17, H4K20, H4K59, H4R3 [14]. When a nucleosome appears in a specific DNA sequence area, these potentially sites can have a certain acetylation or methylation level [5, 16].

Recently we have introduced a support vector machine (SVM)-based method to qualitatively predict histone occupancy, acetylation and methylation areas in DNA sequences [15]. In this paper, we present a different computational method for this prediction problem. We employ conditional random fields (CRF) [6], a novel machine learning technique, to discriminate between DNA areas with high and low relative occupancy, acetylation, or methylation. Our experiments showed that CRF-based method has competitive performance with SVM method. Moreover, similar to SVMs, our CRF method can extract informative k-gram features based on their weight in the CRFs model trained from the datasets of these DNA modifications. The results from our CRF-method on the yeast genome are consistent with those from the SVM method and 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 [16], we extracted 14 datasets and used 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 (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 [16] 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. Sequences with value in between 0.8 and 1.2 are ignored.

2.2 Conditional Random Fields

The sequential classification problem is well known in several scientific fields, especially computational linguistics, and computational biology [6]. There are

Table 1. Datasets of histone occupancy, acetylation, and methylation by ChIP-Chip protocol in vivo [16]

Dataset	#positives	#negatives	Description
H3.YPD	7667	7298	H3 occupancy
H4.YPD	6480	8121	H4 occupancy
H3.H2O2	17971	15516	H3.H202 occupancy
H3K9acvsH3.YPD	15415	12367	H3K9 acetylation relative to H3
H3K14acvsH3.YPD	18771	14277	H3K14 acetylation relative to H3
H3K14acvsWCE.YPD	17672	16290	H3K14 acetylation relative to WCE
H3K14acvsH3.H2O2	18410	15685	H3K14 acetylation relative to H3.H2O2
H4acvsH3.YPD	18410	15685	H4 acetylation relative to H3
H4acvsH3.H2O2	18143	12540	H4 acetylation relative to H3.H2O2
H3K4me1vsH3.YPD	17266	14411	H3K4 monomethylation relative to H3
H3K4me2vsH3.YPD	18143	12540	H3K4 dimethylation relative to H3
H3K4me3vsH3.YPD	19604	17195	H3K4 trimethylation relative to H3
H3K36me3vsH3.YPD	18892	15988	H3K36 trimethylation relative to H3
H3K79me3vsH3.YPD	15337	13500	H3K79 trimethylation relative to H3

two kinds of model for solving this problem, generative models and conditional models. While generative models define a joint probability distribution of the observation and labelling sequences $p(X, Y)$, the conditional models specify the probability of a label given an observation sequence $p(Y|X)$. The main drawback in generative models is that, in order to define a joint probability distribution, they must enumerate all possible observation sequences, which may be not feasible in practice [6, 12, 21]. Our work employs conditional models, specially conditional random fields, which can overcome the drawbacks of generative models.

CRF [6] is a probabilistic framework for segmenting and labelling sequential data using conditional model [6]. It has the form of a undirected graph that defines a log-linear distribution over label sequences given a particular observation sequence. CRFs have several advantages over other models (e.g., HMMs and MEMMs) such as relaxing strong independence Markov assumptions and avoiding weakness called the label bias problem [6, 11, 12, 21].

Definition. CRFs can be represented by an undirected graphical model. According to [6], we define $G = (V, E)$ to be an undirected graph, with $v \in V$ corresponds to each of the random variables representing a label sequence Y_v from Y, and $e \in E$ corresponds to the definition of conditional independence for undirected graphical models. In other words, two vertices v_i and v_j are conditionally independent given all other random variables in the graph.

In theory, CRFs can be represented by arbitrarily structure graph, although in this work, we focus on linear-chain structure graph. Let $X = (x_1, x_2, ..., x_T)$ be an observed data sequence; S be a set of finite state machines, each is associated with a label $l \in L$; and $Y = (y_1, y_2, ..., y_T)$ be the state sequence. The linear-chain CRFs [20, 12] then define the conditional probability of a state sequence given an input sequence as follows

$$p_\theta(Y|X) = \frac{1}{Z(X)} exp(\sum_{i=1}^{T} \sum_k \lambda_k f_k(y_{i-1}, y_i, X, i))$$

where $Z(X) = \sum_{s \in S} exp(\sum_{i=1}^{T} \sum_k \lambda_k f_k(y_{i-1}, y_i, X, i))$ is a normalization factor over all state sequences, and $f_k(y_{i-1}, y_i, X, i)$ are feature functions, each of them is either a state feature function or a transition function [20, 12, 21]. A state feature captures a particular property of the observation sequence X at current state y_i. A transition feature represents sequential dependencies by combining the label l' of the previous state y_{i-1} and the label l of the current state y_i. As [6], we assume that the feature functions is fixed, and denote $\lambda = \{\lambda_k\}$ as a weight vector which to be learned through training.

Inference in CRFs. Inference in CRFs is to find a state sequence y^* which is the most likely given the observation sequence x

$$y^* = \text{argmax}_y p_\theta(y|x) = \text{argmax}_y \left\{ exp(\sum_{i=1}^{T} \sum_k \lambda_k f_k(y_{i-1}, y_i, x, i)) \right\}$$

Similarly to HMMs, CRFs use a dynamic programming method for finding y^* [6, 21, 12]. In fact, we choose the most well-known method being the Vieterbi algorithm [17]. Viterbi stores the probability of the most likely path up to time t which accounts for the first t observations and ends in state y_t. We define this probability to be $\alpha_t(y_i)$ $(0 \leq t \leq T-1)$. We set $\alpha_0(y_i)$ to be the probability of starting in state y_i. The recursion is given by

$$\alpha_{t+1} = max_{y_j} \{\alpha_t(y_j) exp\left(\sum_k \lambda_k f_k(y_j, y_i, x, t)\right)\}$$

At the end time (i.e., $t = T-1$), we can backtrack through the stored information to find the most likely sequence y^*.

Training CRFs. Let $D = \left\{(x^k, y^k)\right\}_{k=1}^{N}$ be the training data set. CRFs are trained by finding the weight vector $\theta = \{\lambda_1, \lambda_2, ...\}$ to maximize the log-likelihood

$$L = \sum_{j=1}^{N} log\left(p_\theta(y^{(j)}|x^{(j)})\right) - \sum_k \frac{\lambda^2}{2\sigma^2}$$

where the second sum is a Gaussian prior over parameters (with variance σ^2) that provides smoothing to help coping with sparsity in the training data [3].

Since the likelihood function in exponential models of CRFs is convex, the above optimization problem always has the global optimum solution, which can be found by an iterated estimation procedure. The traditional method for training in CRFs is iterative scaling algorithms [6, 21]. Sine those methods are very slow for classification [20], therefore we use quasi-Newton methods, such as L-BFGS [8], which are significantly more efficient [10, 20].

L-BFGS is a limited-memory quasi-Newton procedure for unconstrained optimization that requires the value and gradient vector of a function to be optimized. Assuming that the training labels on instance j make its state path unambiguous, let $y^{(j)}$ denote that path, then the first-derivative of the log-likelihood is

$$\frac{\delta L}{\delta \lambda_k} = \left(\sum_{j=1}^{N} C_k(y^{(j)}, x^{(j)})\right) - \left(\sum_{j=1}^{N} \sum_{y} p_\theta(y|x^{(j)}) C_k(y, x^{(j)})\right) - \frac{\lambda_k}{\sigma^2}$$

where $C_k(y, x)$, the count of feature f_k given y and x, equal to $\sum_{t=1}^{T} f_k(y_{i-1}, y_i, x, i)$, i.e., the sum of $f_k(y_{i-1}, y_i, x, i)$ values for all positions i in the training sequence. The first two terms correspond to the difference between the empirical and the model expected values of feature f_k. The last term is the first-derivative of the Gaussian prior.

2.3 Features of a DNA Sequence Area

The most important issue in CRFs learning is to select a set of features that hopefully capture the relevant relationships among observations and label sequences. CRFs have two kinds of features, state features and transition features. However, in this work we focus only on state features. Also, each observation sequence in the datasets has only one observation (L-DNA sequence area) and the label sequence is a sequence of 0 (negative class) and 1 (positive class). Our feature set to input to CRF systems is built by two steps. First, we use a k-sliding window along a DNA sequence to get binary k-grams (patterns of k consecutive nucleotide symbols). Each DNA sequence is thus represented by a binary 4^k-dimensional vector of all possible k-grams. Second, we define the unigram function for each k-gram as follows:

$$u_t(x) = \begin{cases} 1 \text{ if the t}^{th} \text{ } k\text{-gram appear in the sequence } x \\ \\ 0 \text{ otherwise} \end{cases}$$

Therefore, the relationship between the observation and two classes, positive and negative, is described in the following features:

$$f_{tP}(y, x) = \begin{cases} u_t(x) \text{ if } y \text{ belong to positive class} \\ \\ 0 \text{ otherwise} \end{cases}$$

$$f_{tN}(y, x) = \begin{cases} u_t(x) \text{ if } y \text{ belong to negative class} \\ \\ 0 \text{ otherwise} \end{cases}$$

3 Results and Discussion

3.1 Prediction of Histone Occupancy, Acetylation, and Methylation

We used CRFs with the limited-memory quasi-Newton method (Section 2.2) to perform threefold cross-validation on 14 datasets of histone occupancy, acetylation and methylation areas (Table 1). Three criteria of precision, recall and F1-measure are used to report the results:

Table 2. Results of histone occupancy, acetylation and methylation prediction

	k=5			k=6			k=4,5			k=5,6		
Dataset	*Pre.*	*Rec.*	$F1$.	*Pre.*	*Rec.*	$F1$.	*Pre.*	*Rec.*	$F1$.	*Pre.*	*Rec.*	$F1$.
H3.YPD	80.17	80.17	80.07	82.33	82.31	82.32	80.27	80.27	80.27	**82.67**	**82.46**	**82.56**
	(80.54)	(80.50)	(80.52)	(81.78)	(81.62)	(81.70)	(80.94)	(80.89)	(80.92)	(79.97)	(79.28)	(79.62)
H4.YPD	83.21	83.07	83.14	85.62	85.49	85.55	82.79	82.67	82.73	**85.57**	**85.53**	**85.55**
	(81.72)	(81.57)	(81.65)	(83.97)	(83.87)	(83.92)	(81.87)	(81.72)	(81.79)	(82.63)	(82.13)	(82.38)
H3.H2O2	82.80	82.53	82.67	82.96	82.98	82.97	82.73	82.76	82.74	**83.06**	**83.00**	**83.03**
	(80.85)	(80.79)	(80.82)	(81.51)	(81.34)	(81.43)	(81.20)	(81.12)	(81.16)	(81.45)	(81.39)	(81.42)
H3K9acvsH3.YPD	70.36	70.22	70.29	71.50	71.27	71.38	69.86	69.65	69.75	**71.58**	**71.44**	**71.51**
	(70.92)	(70.74)	(70.83)	(73.98)	(73.49)	(73.74)	(71.12)	(70.96)	(71.04)	(71.76)	(70.22)	(70.98)
H3K14acvsH3.YPD	66.58	65.99	66.28	**68.13**	**68.06**	**68.09**	66.26	65.69	65.97	68.27	67.78	68.02
	(68.55)	(67.66)	(68.10)	(73.12)	(71.68)	(72.39)	(68.80)	(67.80)	(68.30)	(71.95)	(67.44)	(69.62)
H3K14acvsWCE.YPD	62.86	62.60	62.73	63.76	63.67	63.71	65.88	65.69	65.79	**63.74**	**63.72**	**63.73**
	(64.04)	(63.95)	(64.00)	(68.04)	(67.83)	(67.94)	(64.47)	(64.37)	(64.42)	(66.52)	(65.81)	(66.16)
H3K14acvsH3.H2O2	65.42	65.41	65.41	**66.58**	**66.44**	**66.51**	66.21	66.02	66.12	66.46	66.43	66.44
	(67.14)	(67.13)	(67.14)	(69.88)	(69.44)	(69.66)	(67.35)	(67.34)	(67.34)	(69.28)	(68.85)	(69.06)
H4acvsH3.YPD	66.21	66.02	66.12	67.64	67.43	67.53	65.94	65.75	65.85	**67.85**	**67.66**	**67.75**
	(68.01)	(67.62)	(67.82)	(72.22)	(71.60)	(71.91)	(68.31)	(67.85)	(68.08)	(69.98)	(68.11)	(69.03)
H4acvsH3.H2O2	69.73	69.69	69.71	**70.44**	**70.27**	**70.35**	69.65	69.59	69.62	70.16	70.11	70.13
	(69.32)	(69.27)	(69.29)	(71.69)	(71.42)	(71.55)	(69.32)	(69.25)	(69.29)	(69.93)	(69.07)	(69.50)
H3K4me1vsH3.YPD	65.21	64.79	65.00	66.47	65.83	66.15	64.87	64.40	64.63	**67.03**	**66.34**	**66.68**
	(66.16)	(65.59)	(65.87)	(69.55)	(68.48)	(69.01)	(66.43)	(65.79)	(66.11)	(68.30)	(65.87)	(67.06)
H3K4me2vsH3.YPD	68.58	68.05	68.31	**64.78**	**62.85**	**63.80**	62.95	62.15	62.54	64.41	62.95	63.67
	(64.20)	(61.50)	(62.82)	(68.10)	(64.15)	(66.07)	(64.68)	(61.70)	(63.16)	(67.25)	(60.18)	(63.52)
H3K4me3vsH3.YPD	67.00	66.81	66.91	63.18	62.66	62.92	60.90	60.53	60.71	(64.49)	**64.23**	**64.36**
	(63.96)	(63.55)	(63.75)	(68.90)	(64.32)	(66.53)	(64.50)	(64.06)	(64.28)	(68.08)	(65.76)	(66.90)
H3K36me3vsH3.YPD	69.76	69.55	69.65	71.21	71.05	71.13	69.46	69.24	69.35	**71.44**	**70.90**	**71.17**
	(70.85)	(70.61)	(70.73)	(72.39)	(71.93)	(72.16)	(71.11)	(70.87)	(70.99)	(72.83)	(71.88)	(72.36)
H3K79me3vsH3.YPD	75.83	75.56	75.69	**78.02**	**77.87**	**77.94**	75.50	75.37	75.44	77.81	77.73	77.77
	(76.24)	(76.02)	(76.13)	(78.95)	(78.31)	(78.63)	(76.44)	(76.21)	(76.32)	(78.03)	(77.38)	(77.70)

Note: 1.Pre., Rec., F1. are precision, recall and F1-measure, respectively.
2.The numbers in the brackets are prediction results by using SVM [15]

$$Precision_{positive} = \frac{TP}{TP+FP}; Precision_{negative} = \frac{TN}{TN+FN}$$
$$Recall_{positive} = \frac{TP}{TP+FN}; Recall_{negative} = \frac{TN}{TN+FP}$$
$$Precision = \frac{Precision_{positive}+Precision_{negative}}{2}$$
$$Recall = \frac{Recall_{positive}+Recall_{negative}}{2}$$
$$F1 - measure = \frac{2*(Precision*Recall)}{Precision+Recall}$$

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

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). Due to the computational complexity, we have only tried with $k \leq 6$ and report here the results from sets of k-grams with k=5, k=6, k=4,5, and k=5,6.(Table 2).

The highest performance of our CRF method (at 18^{th} L-BFGS iteration) for relative histone occupancy predictions (H3, H4, H3.H2O2), and acetylation predictions (H3K9acvsH3, H3K14acvsH3, H3K14acvsWCE, H3K14acvsH3.H2O2, H4acvsH3, H4acvsH3.H2O2), as well as methylation predictions (H3K4me1vsH3, H3K4me2vsH3, H3K4me3vsH3, H3K36me3vsH3, H3K79me3vsH3.YPD) achieved when we use features of both 5-grams and 6-grams (Table 2). The numbers in the brackets are the performance of the support vector machine (SVM)-based method (which was used in [15] to address the same problem) when using the same binary k-gram features. As it can be seen, CRF method is competitive with SVM-based method. In some cases, CRFs gave better performance, but in others performance was worse. SVM method can take into account the number of k-gram occurrences that represents DNA sequence better than binary k-gram features, hence SVM method can achive better performance [15]. However, CRFs have some advantages over SVMs such as they can easily incorporate knowledge into their prediction, and in the future we will take account annotated information concerning DNA sequence into our CRF method to improve the prediction results.

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

During the training CRFs model, we reported the weight of features (i.e. weight vector, see Section 2.2). In a CRF model, features with the larger weight would be more relevant than those with lower weight. We ranked the features based on their weight supporting for either positive or negative classes in CRF models, which were trained on 14 datasets. Table 3 and Table 4 show the most informative features from a set of 4-grams and 5-grams at 18^{th} L-BFGS iteration (which did though give the best performance (Table 2), but to make later interpretation easily) for histone occupancy, acetylation, and methylation.

Informative features ranked by our CRF-based method agree with those from the previous SVM-based method [15]. They can be useful to analyze the genetic area preferences of histone occupancy, acetylation, and methylation. For example, CG (CpG) is a dinuceotide that appears very often in the most informative

Table 3. Most informative features selected from CRFs model for positive class with k-grams=4 and k-grams=5

Dataset	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
H3.YPD	CTTCA	0.16	CTTTA	0.15	TGCAG	0.14	ACAGC	0.14
	CGGC	0.13	TGAAG	0.13	GTTTG	0.13	GCGA	0.13
	GTGAT	0.13	TCATC	0.13	TGGC	0.13	CAGC	0.13
H4.YPD	TAAT	0.26	CTTCA	0.23	CAAAT	0.22	GCCAC	0.20
	GGATC	0.20	CTGGT	0.19	TTGGG	0.19	ATTTG	0.18
	ATCAG	0.18	GCAG	0.18	TATA	0.18	TTTA	0.17
H3.H2O2	CCGC	0.21	GCGC	0.21	CGGC	0.20	CGGG	0.19
	GCCG	0.18	CGCG	0.18	GGCC	0.18	CATGG	0.17
	CCGG	0.16	CCCT	0.15	CGCC	0.15	CCACC	0.15
H3K9acvsH3.YPD	CATGC	0.11	CAGGG	0.10	GTTCG	0.10	GCGAG	0.10
	CTTAG	0.09	TCTCG	0.09	TACC	0.09	GATAC	0.09
	CCCCG	0.09	AGGCG	0.09	GCCGG	0.09	CACCG	0.09
H3K14acvsH3.YPD	GCGTG	0.12	TTTTT	0.10	TAGTC	0.09	CTCGC	0.09
	CTCAT	0.09	CACC	0.08	TCTCT	0.08	ATATA	0.08
	CTTTT	0.08	AAAAA	0.08	AGCGG	0.08	TTTTC	0.08
H3K14acvsWCE.YPD	ACGGT	0.10	TCTCT	0.10	AGCCT	0.09	CTCAT	0.09
	CGGA	0.09	CGGC	0.09	CACC	0.09	TCCG	0.09
	AGTCG	0.08	TTGCT	0.08	ATGCG	0.08	GGAGT	0.08
H3K14acvsH3.H2O2	AGGGG	0.12	CCCCT	0.11	TAGTC	0.10	CACC	0.10
	CGAGG	0.09	CACAC	0.09	CGTAC	0.09	CCCGG	0.08
	ATGCG	0.08	TTAGT	0.08	TCTCT	0.08	CGTGC	0.08
H4acvsH3.YPD	CTCAT	0.12	AGCAA	0.10	CACAC	0.10	CACC	0.09
	GAAAA	0.09	GATAC	0.08	CATGC	0.08	TACCC	0.08
	TAGTC	0.08	TTAT	0.08	TCTCT	0.07	CAAGT	0.07
H4acvsH3.H2O2	AGGGG	0.18	GGGGG	0.14	AAAAG	0.13	CCCCT	0.12
	GTGGC	0.11	AAGGG	0.10	CTCCC	0.09	CTTGT	0.09
	ACACG	0.09	GATAC	0.09	GGGAG	0.09	CCTCG	0.08
H3K4me1vsH3.YPD	GGCA	0.08	TATC	0.08	CCAG	0.08	CTTGA	0.08
	TTAA	0.08	TGCGG	0.08	TGCAT	0.07	CCTCA	0.07
	TCCAA	0.07	AACCC	0.07	AGTT	0.07	GGTTG	0.07
H3K4me2vsH3.YPD	CTCAT	0.06	ATGAG	0.06	GGGAA	0.06	CTTGT	0.06
	AGACA	0.06	GATCT	0.05	CACTT	0.05	ACCAC	0.05
	AGTCC	0.05	GCTTA	0.05	AAAGA	0.05	GTCCA	0.05
H3K4me3vsH3.YPD	CACC	0.10	ACCCG	0.09	AGCCA	0.09	CAAGT	0.08
	GTCCA	0.08	GTCAA	0.08	TCTCT	0.08	GAAAA	0.07
	GCGTG	0.07	CTCAT	0.07	TAGTC	0.07	TCACT	0.07
H3K36me3vsH3.YPD	AAAA	0.14	TACT	0.12	ATAT	0.10	TTTT	0.10
	GTGA	0.10	CCTCC	0.09	TAAT	0.09	CGTCC	0.09
	CATCA	0.09	AGTT	0.09	AACA	0.09	GGACG	0.09
H3K79me3vsH3.YPD	TATA	0.22	TAAT	0.22	TAAA	0.16	ATAT	0.16
	TATT	0.14	ATTA	0.14	CATCA	0.14	TTAGA	0.14
	TGCA	0.13	TACT	0.13	TTTA	0.13	GATTT	0.11

negative features (Table 4). In other words, CG-rich DNA sequence areas are often free of histone occupancy, acetylation, or methylation. We all knew that CpG islands are usually near to gene starts. So we can infer from our results that promoter regions are often not occupied by nucleosomes. This is consistent with previous results by experimental approaches in vivo [16].

4 Conclusion

We have introduced a conditional model based method to predict qualitative histone occupancy, acetylation, and methylation areas in DNA sequences. We have selected a basic set of features based on DNA-sequence. Moreover, our model can evaluate the informative features to discriminate between DNA areas

Table 4. Most informative features selected from CRFs model for negative class with k-grams=4 and k-grams=5

Dataset	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
H3.YPD	CGCGC	0.15	TTTTT	0.13	AAAAA	0.12	GCGCG	0.11
	CGCGG	0.09	CCGCG	0.09	CGGGC	0.09	CGTGC	0.08
	GCGGG	0.08	TTATA	0.08	TTTTA	0.07	GGCCG	0.07
H4.YPD	AAAAA	0.38	TTTTT	0.29	AGAAA	0.27	GCGCG	0.27
	CGGAC	0.26	TTATA	0.26	TATAT	0.25	CGTGC	0.23
	CGCGC	0.23	CCCGG	0.22	GGCT	0.22	CGCGG	0.21
H3.H2O2	CGCGC	0.35	GCGCG	0.27	GCGGG	0.23	CGCGG	0.22
	CCGCG	0.22	TTTTT	0.20	AGGT	0.18	CTTC	0.16
	CCCCC	0.16	GGGCG	0.15	CCGGG	0.15	ACCA	0.14
H3K9acvsH3.YPD	GCCGC	0.13	GCAC	0.10	TCCAA	0.10	CCTCC	0.10
	ATTTG	0.09	AAAG	0.09	TTCTG	0.09	CAAAT	0.09
	TCTT	0.09	ATATT	0.09	GCAG	0.09	GCTG	0.09
H3K14acvsH3.YPD	GCCGC	0.11	CCAAT	0.09	TTATC	0.08	CTCGT	0.08
	ATTTG	0.08	ATTCA	0.08	TGATG	0.08	AAATT	0.08
	CCAAA	0.07	TCTAA	0.07	CATCA	0.07	TCAG	0.07
H3K14acvsWCE.YPD	CGCGG	0.14	GCCGC	0.12	AAGC	0.12	GCGGC	0.11
	CTTA	0.11	TCTT	0.10	CGCGC	0.10	TCAG	0.10
	AACA	0.10	CAAG	0.09	CTCT	0.09	GTCC	0.09
H3K14acvsH3.H2O2	AAATT	0.12	TAGT	0.11	TACG	0.08	GTGGG	0.08
	CTTA	0.08	TATTA	0.08	GCGTC	0.08	GTGA	0.08
	AAGC	0.08	CATA	0.08	TCCC	0.07	AATCA	0.07
H4acvsH3.YPD	GCCGC	0.15	TATTA	0.11	CTCT	0.11	GCGGC	0.10
	CAAG	0.09	TCGGA	0.09	TTATC	0.09	ATATT	0.08
	TTTGA	0.08	AACA	0.08	TTCTT	0.08	CCAAA	0.08
H4acvsH3.H2O2	TATTA	0.10	TCGT	0.09	TGGAT	0.09	ATATT	0.09
	TTTTG	0.09	TAATT	0.08	AATTT	0.08	ACAG	0.08
	AAGC	0.08	TACG	0.08	CCATA	0.08	TAAAA	0.08
H3K4me1vsH3.YPD	GAAG	0.10	TACAC	0.10	CCGAG	0.09	TATGT	0.08
	CAATT	0.08	ATAGT	0.08	CCGGC	0.08	CGAGG	0.08
	ACCCG	0.07	GCGTG	0.07	TGGG	0.07	TCCTA	0.07
H3K4me2vsH3.YPD	ATATT	0.09	TATTA	0.08	TGAAG	0.07	AATAT	0.06
	TAATA	0.06	TAATT	0.05	TTAAT	0.05	GTAAT	0.05
	CTAAA	0.05	GCCGC	0.05	AACAT	0.05	ATCAT	0.05
H3K4me3vsH3.YPD	GCCGC	0.14	CGCGG	0.09	CAAG	0.09	TCAG	0.09
	ACCCC	0.09	GCGGC	0.09	CTTA	0.09	AAGC	0.08
	GCGCG	0.08	CTCT	0.08	AACA	0.07	GCGTC	0.07
H3K36me3vsH3.YPD	GAAG	0.18	ATAGT	0.12	TATAT	0.12	AAAAG	0.11
	TAGGA	0.10	CTTAA	0.10	CTCGA	0.09	CACC	0.09
	CTCAT	0.09	GTACT	0.09	ACCCG	0.09	ACATA	0.09
H3K79me3vsH3.YPD	TATAT	0.22	ATATA	0.20	ACATA	0.19	AAAAA	0.17
	TTGT	0.16	TTATA	0.16	TATAA	0.16	GCCGC	0.15
	ACGTA	0.13	GAAG	0.13	GCCCG	0.13	CTTC	0.12

with high and low occupancy, acetylation, or methylation. In the near future, we plan to incorporate features related to sequence motifs into our method in order to capture more faithfully the constrains on the model.

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, and by COE project JCP KS1 of the Japan Advanced Institute of Science and Technology. We also would like to thank Hieu P.X. for providing FlexCRFs package and sharing with us his experience in machine learning area.

References

1. B. E. Bernstein, E. L. Humphrey, R. L. Erlich, R. Schneider, P. Bouman, J. S. Liu, T. Kouzarides, and S. L. Schreiber. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc. Natl. Acad. Sci. USA*, 99(13):8695–8700, 2002.
2. 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.
3. S. F. Chen and R. Rosenfeld. *A gaussian prior for smoothing maximum entropy models.* Technical report CMU-CS-99-108, 1999.
4. T. Kouzarides. Histone methylation in transcriptional control. *Curr. Opin. Genet. Dev.*, 12(2):198–209, 2002.
5. S. K. Kurdistani, S. Tavazoie, and M. Grunstein. Mapping global histone acetylation patterns to gene expression. *Cell*, 117(6):721–733, 2004.
6. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conference on Machine Learing*, 2001.
7. 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.
8. D. Liu and J.Nocedal. On the limited memory bfgs method for large-scale optimization. *Mathematical Programming*, 45:503–528, 1989.
9. 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 A resolution. *Nature*, 389(6648):251–260, 1997.
10. R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. Proceeding CoNLL*, 2002.
11. A. McCallum. Maximum entropy markov models for information extraction and segmentation. In *Proc. 15th International Conference on Machine Learing*, 2000.
12. A. McCallum. Efficiently inducing features of conditional random fields. In *Proc. 19th Conference on Uncertainy in Artificial Intelligence*, 2003.
13. 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.
14. C. L. Peterson and M. A. Laniel. Histones and histone modifications. *Curr. Biol.*, 14(14):R546–R551, 2004.
15. 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 International Conference on Genome Informatics*, 2005.
16. 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.
17. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. Proceeding of IEEE*, pages 257–286, 1989.
18. B. Ren, F. Robert, and et al. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, 2000.
19. D. Robyr, Y. Suka, I. Xenarios, S. K. Kurdistani, A. Wang, N. Suka, and M. Grunstein. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell*, 109(4):437–446, 2002.
20. F. Sha and F. Pereira. Shalow parsing with conditional random fields. In *Proc. 15th Proceeding of Human Language Technology*, 2003.
21. H. Wallach. *Efficient Training of Conditional Random Fields.* Master thesis, 2002.