Journal of Ambient Intelligence & Humanized Computing manuscript No.
(will be inserted by the editor)

Quantitatively assessing the effects of regulatory factors on nucleosome dynamics by multiple kernel learning

Bich Hai Ho · Ngoc Tu Le · Tu Bao Ho

Received: date / Accepted: date

Abstract Nucleosome, a nucleoprotein structure formed by coiling 147 bp of DNA around an octamer of histone proteins, is the fundamental repeating unit of eukaryotic chromatin. By regulating the access of biological machineries to underlying *cis*-regulatory elements, its mobility has been implicated in many important cellular processes. Although it has been known that various factors, such as DNA sequences, histone modifications, etc., cooperatively affect nucleosome mobility, the contribution of each factor in the common impact remains unclear. We propose, in this work, a novel computational approach based on Multiple Kernel Learning (MKL) for quantitatively assessing the effects of two important factors, i.e., genomic sequence and post-translational histone modifications (PTMs), on nucleosome dynamics. Our result on *S.cerevisiae* shows that, epigenetic feature, such as histone modifications, plays more important role than genomic sequence in regulating nucleosome dynamics. Based on that, we carried further analysis on each PTM to reveal their combinatory effects on nucleosome dynamics and found out that some pairs of PTMs such as H3K9Ac - H4H14Ac, H4K5Ac - H4K12Ac and H4K5Ac - H3K14Ac might co-operate in altering nucleosome stability in gene regulation.

Keywords nucleosome dynamics · post-translational histone modification · multiple kernel learning · genomic sequence

1 Introduction

Eukaryotic genomes are packaged inside cell nucleus under chromatin structure, which has the form like a bead-on-string fiber containing fundamental units of nucleosomes. Each nucleosome is formed by wrapping 147bp of DNA around a histone core consisting two each of four histone proteins H3, H4, H2A, H2B (Luger et al 1997). There is increasing evidence showing that chromatin plays more important role far beyond DNA compaction. By occluding the access of biological machineries to *cis*-regulatory elements and/or modifying the related epigenetic information, it can ubiquitously and profoundly affect many important biological processes, such as transcription, DNA replication, DNA repair, etc. To overcome the obstacle imposed by chromatin, cells have developed complicated pathways (Li et al 2007; Prost et al 2009). In these pathways, nucleosome must be displaced or even removed from chromatin to provide access to the underlying DNA sequences. Hence, understanding how cells regulate nucleosome stability will give us additional insights into the mechanisms of those cellular processes.

The dynamics (or stability and mobility, hereinafter used interchangeably) of nucleosome can be described as the phenomenon of nucleosome positioning being non-uniform temporally (Tanaka et al 2010). Such stability can be affected by the combinatory effects of

Bich Hai Ho*
E-mail: haihb@jaist.ac.jp

Ngoc Tu Le*
E-mail: ngoctule@jaist.ac.jp

Tu Bao Ho**
E-mail: bao@jaist.ac.jp
* School of Knowledge Science
Japan Advanced Institute of Science and Technology
Tel.: +81-761-51-1730
Present address: Asahidai 1-1 Nomi City Ishikawa 923-1292 Japan
** John von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam.

many factors, including DNA sequence, post-translational histone modifications (PTMs) or chromatin remodeling complexes (Henikoff 2008). First, nucleosomal DNA sequence preferences probably originate from the sequence dependent mechanics of the wrapped DNA itself (Windom 2001). A number of previous works both in *in vivo* and *in vitro* have suggested that genomic signatures are reliable determinants for nucleosome occupancy (Segal et al 2006; Gupta et al 2008; Tillo and Hughes 2009). In *S.cerevisiae*, more specifically, Kaplan *et al.* (2009) directly pointed out that intrinsic DNA sequence preferences play a dominant role in nucleosome organization. They are, therefore, likely to be important to favoring or disfavouring nucleosome eviction or nucleosome dynamics. Besides the DNA sequence itself, PTMs, the covalent changes occur in the histone tail domains, can also influence nucleosome positions. Segal and Windom (2009) reckoned that PTMs impacts on histone positioning may be indirect and modest, but may be substantial in influencing ATP-dependent chromatin remodelling factors, which as *trans*-factor have been shown to be closely related to nucleosome positioning (Jansen and Verstrepen 2011). Also, acetylated histones are shown to be easily dissociated from DNA (Zhao et al 2005). Last but not least, chromatin remodelling complexes usually cooperate with histone chaperones to displace histones from their original positions (Li et al 2007). These evidences lead to the consideration of combinatory effects of DNA sequence and PTMs on nucleosome stability.

Recent advancement in profiling techniques, such as ChIP-Chip and ChIP-Seq, has provided unprecedented opportunities for investigating the effects of several regulatory factors on nucleosome dynamics at the same time. However, to our understanding, previous attempts on the above mentioned two factors are still modest. Previous works such as (Rippe et al 2007; Schnitzler 2008) that accessed the co-effects of DNA sequence and chromatin remodeling complexes on nucleosome dynamics were experimental. Recently, Le *et al.* (2009) proposed a computational approach to qualitatively assess the effects of DNA sequence and histone modifications on nucleosome dynamics. The problem of to what extend each factor contributes to the regulation of nucleosome dynamics has not been addressed yet.

In this work, we propose a computational approach for quantitatively assessing the effects of two important factors, i.e., genomic sequence and PTMs, on nucleosome dynamics. The problem is formulated as a prediction one and solved by the so-called Multiple Kernel Learning (MKL) framework (Sonnenburg et al 2006). Our first attempt of adapting the theoretically well-founded MKL for this particular biological problem may pave the way for more efforts in considering other regulatory elements simultaneously. Moreover, our obtained results on *S.cerevisiae* show further evidence that genomic sequence and epigenetic feature, e.g., PTMs, can together influence nucleosome stability. More importantly, the latter plays a more important role than the former in determining the stability states of nucleosome. To further confirm that conclusion, we carried an analysis on PTMs separately, based on the well-known hypothesis of *histone language* (Oliver and Denu 2011) . The results show that there exists a number of interactions among them, loosely deemed as co-operation, some of which have been reported in literature. These are encouraging findings to explaining the dynamics of nucleosome and offer an insight into epigenetic regulatory mechanisms of many important cellular processes.

The rest of this paper is organized as follows: Section II. presents the proposed approach; Section III. gives details on experiment; Section IV. discusses further biological insights along with our findings; Section V. concludes the work.

2 Methodology

The problem is formulated as a binary classification task, in which the classifiers take two kinds of features, i.e., genomic (DNA sequence) and epigenetic (PTMs), as inputs and output nucleosome states of *well-positioned* and *delocalized*, as class labels. By representing different data sources as their corresponding kernels and learning optimal weights for each, MKL is not only as efficient as Support Vector Machines (SVMs), a state-of-the-art classification technique, in the sense of classification power, but also it enables the combination of heterogeneous data into a common format and improves the interpretability of resultant classifiers (Sonnenburg et al 2006). Also, in another work by MKL author, it was shown that MKL can be successfully applied to a large number of biological sequence features (Ratsch et al 2006). In our work, we represented genomic and epigenetic data in an appropriate and reasonably informative form, and solved prediction problem under MKL framework. The aims are two-fold: assessing not only whether the features can help to predict nucleosome state and access the contribution of each kind of feature to the final outcome, the state of nucleosome. Details and background are presented in the following subsections.

2.1 Data representation

For training purpose, nucleosomal sequences were represented using two different approaches. The first one, called *spectrum* feature, was proposed by Peckham *et al.* (2007), which has shown competitive performance for the task of discriminating "nucleosome forming" sequences from "nucleosome inhibiting" sequences and was applied successfully on human data (Gupta et al 2008). According to this approach, each nucleosomal sequence is represented as a $2,772$-entry vector, in which each entry is a normalized count of the occurrences of a particular k-mer or its reverse complement, with k=1 up to 6. Sequence elements longer than 6 nucleotides showed no significant changes the discrimination power of their classifiers. The second approach, derived from the work of Tillo and Hughes (2009), argues that the sequence rules underlying nucleosome occupancy can be captured by a much simpler model compared to the model proposed previously by Kaplan *et al.* (2009), which used *spectrum* 5-mer features. This observation has been supported by various experimental works on the different nucleosome binding affinities of DNA sequences, i.e., some specific DNA sequence properties are more important for intrinsic nucleosome preference. For example, G+C content alone can explain 50% of the variation in nucleosome occupancy *in vitro*, consistent with the conclusion by (Peckham et al 2007). Poly-dA/dT sequences are also believed both *in vivo* to be rigid and anti-nucleosome forming (Kaplan et al 2009; Schwartz et al 2009); and *in vitro* to be highly discriminative with ROC-scores of 0.91 in (Tillo and Hughes 2009). Based on this, we represented each nucleosomal sequence as a 12-dimension vector, one is its %GC content and the other eleven are its 4-mer features (Table 1). We skipped two physical characteristics, called *propeller twist* and *slide* from original work (Tillo and Hughes 2009) in our representation. Experiments were done on both representations for our case of predicting nucleosome stability, resulting in nearly the same effectiveness to the classification performances with each kind of feature (the data were not shown); thus we used the 12-entry one for further analysis. Lastly, PTMs feature corresponding to each nucleosome was represented as a normalized vector of 9 acetylations and 3 histone methylations (Table 1).

No.	Genomic	Histone modifications
1	G+C content	H3K18Ac
2	AAAA	H4K12Ac
3	AAAT	H3K9Ac
4	AAGT	H3K14Ac
5	AATA	H4K5Ac
6	AATT	H2AK7Ac
7	AGAA	H4K8Ac
8	ATAA	H4K16Ac
9	ATAT	H2BK16Ac
10	ATTA	H3K4Me1
11	GAAA	H3K4Me2
12	TATA	H3K4Me3

Table 1 Genomic and histone modification features

2.2 Support Vector Machines (SVMs) and Kernel methods

SVMs is a binary supervised classification method which has a solid theoretical background and performs the classification task more accurately than most other algorithms in many applications. It was also successfully applied to a wide variety of problems in computational biology, such as protein function prediction, splice site recognition, etc. Given a training set containing instance-class pairs $(\mathbf{x}_i, y_i)$, $i = 1, 2, \cdots, l$ where $\mathbf{x}_i \in R^l$ and $y_i \in \{-1, 1\}$ is a class label, if the data are linearly separable, an SVM classifier is a hyperplane $\mathbf{w}^T\phi(\mathbf{x}_i) + b$, where $\phi(\mathbf{x}_i)$ is a function mapping $\mathbf{x}_i$ into a higher (maybe infinite) dimensional space, that best separates the two classes. In case of non separable data, finding the hyperplane can be transformed into following primal optimization problem:

$$\begin{array}{l} Minimize: \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_{i=1}^{l} \xi_i \\ \text{Subject to: } y_i\left(\mathbf{w}^T\phi\left(\mathbf{x}_i\right) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \; i = 1, 2, \ldots, l \end{array} \tag{1}$$

This problem is called soft margin optimization, which is able to deal with errors in the data by allowing some data points to fall on the wrong side of the separating hyperplane by introducing slack variables $\xi_i(\geq 0)$. Its dual is a quadratic optimization problem:

$$\begin{array}{l} Minimize: \frac{\alpha^T Q\alpha}{2} - \mathbf{e}^T\alpha \\ \text{Subject to: } C \geq \alpha_i \geq 0 \; i = 1, 2, \ldots, l \\ \mathbf{y}^T\alpha = 0 \end{array} \tag{2}$$

where $\mathbf{e}$ is an unit vector, $C > 0$ is an error penalty parameter, $Q_{ij} = y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right)$, $K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \langle\phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle$ is a *kernel function*, a fundamental concept of a class of machine learning techniques called *kernel methods*.

Mathematically, for a function $K : X \times X \longrightarrow \mathbb{R}$ to be a kernel function, it must satisfy two conditions. First, it must be symmetric, that is $K(\mathbf{x}, \mathbf{x'}) = K(\mathbf{x'}, \mathbf{x})$ and second, it must be positive definite, that is

$$\sum_{i=1}^{l}\sum_{j=1}^{l} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) > 0$$

for any $l > 0$, any choice of l objects $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l \in X$ and any choice of non-zero real vector $(c_1, c_2, \ldots, c_l) \in R^l \backslash \{0\}$. There are many kernel functions, so how to select a good kernel function in an application is also a critical issue. However, there are several popular kernel functions, such as linear kernel, RBF kernel, polynomial kernel, spectrum kernel, among others. We can, therefore, choose among them in applications by heuristic selection. With the introduction of kernel functions, a linear classification algorithm can be used to build nonlinear classifiers by using "kernel trick" calculation. Moreover, by transforming each data source into a kernel matrix using a suitable kernel function and then combining those matrices into one, kernel method provides the ability to work with multiple, heterogeneous data sources.

The simplest way to combine multiple kernels is called *unweighed multiple kernel method* (UMK), in which each kernel is assigned an equal weight. It is based on positive semi-definiteness of kernel function, that is the addition of several kernel functions is also a kernel function: given kernel functions $K_1, K_2, \ldots, K_n$ and the embedding mappings $\phi_1, \phi_2, \ldots, \phi_n$ then the function $K = \sum_{k=1}^{n} K_k$ is a kernel function, too. This method, however, does not offer an easy way to interpret the decision function.

2.3 Multiple Kernel Learning (MKL)

MKL framework provides an alternative approach to combining multiple kernels. It considers convex combinations of n kernels:

$$K = \sum_{k=1}^{n} \beta_k K_k \tag{3}$$

with $\beta_k \geq 0$ and $\sum_{k=1}^{n} \beta_k = 1$, where each kernel K_k uses only a set of features. For appropriately designed subkernels K_k, the optimized combination coefficients can then be used to infer which features of the examples are of importance for discrimination. If we are able to obtain an accurate classification by a *sparse* weighting β_k then the resulting decision function is easy to interpret. This is an important characteristic missing in current kernel based algorithms.

To find the optimized coefficients β_k, Lanckriet at al. (2004) reformulated MKL problem as a convex optimization one known as quadratically-constraint quadratic program (QCQP), that can be solved by general-purpose optimization toolboxes or by an algorithm based on sequential minimization optimization (SMO) proposed by Bach *et al.* (2004). These algorithms, however, are only feasible for small problems. Sonnenburg *et al.* (2006) proposed an efficient algorithm that can work with large-scale data by reformulating the binary classification MKL problem as a *semi-infinite linear program* (SILP). This framework was used for the task of learning MKL classifiers in our work.

In the MKL problem for binary classification, given N data points $(\mathbf{x}_i, y_i)$ $(y_i \in \{-1, 1\})$ where $\mathbf{x}_i$ is transformed by n mappings: $\phi_k(\mathbf{x}) \longmapsto \mathbb{R}^{D_k}$, $k = 1 \ldots n$, from the input into n feature spaces $(\Phi_1(\mathbf{x}_i), \ldots \Phi_n(\mathbf{x}_i))$, with D_k denotes the dimensionality of the k-th feature space, we should solve the following optimization problem:

MKL Primal for Classification

$$\begin{aligned}
&Minimize: \tfrac{1}{2}(\sum_{k=1}^{n} \|w_k\|_2)^2 + C\sum_{i=1}^{N}\xi_i \\
&\text{w.r.t.: } w_k \in \mathbb{R}^{D_k}, \boldsymbol{\xi} \in \mathbb{R}^N, b \in \mathbb{R} \\
&\text{s.t. : } \xi_i \geq 0 \; and \; y_i(\sum_{k=1}^{n} \langle w_k, \Phi_k(\mathbf{x}_i)\rangle + b) \geq 1 - \xi_i
\end{aligned} \tag{4}$$

Its dual problem is derived as following (Bach et al 2004):

MKL Dual for Classification

$$\begin{aligned}
&Minimize: \gamma - \sum_{i=1}^{N}\alpha_i \\
&\text{w.r.t.: } \gamma \in \mathbb{R}, \boldsymbol{\alpha} \in \mathbb{R}^N \\
&\text{s.t. : } \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1C}, \; \sum_{i=1}^{N}\alpha_i y_i = 0 \\
&\tfrac{1}{2}\sum_{i,j=1}^{N}\alpha_i y_i \alpha_j y_j K_k(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma, \; k = 1, \ldots, n
\end{aligned} \tag{5}$$

where $K_k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_k(\mathbf{x}_i), \Phi_k(\mathbf{x}_j)\rangle$. Finding the solution for problem (5) is equivalent to solving the following semi-infinite linear program (Sonnenburg et al 2006):

Semi-Infinite Linear Program (SILP)

$$\begin{aligned}
&Maximize : \theta \\
&\text{w.r.t.: } \theta \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^n \\
&\text{s.t. : } 0 \leq \boldsymbol{\beta},\ \sum_k \beta_k = 1 \ and \ \sum_{k=1}^{n} w_k S_k(\alpha) \geq \theta \qquad (6) \\
&\ \text{for all } \boldsymbol{\alpha} \in \mathbb{R}^N,\ \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C1} \ and \ \sum_i y_i \alpha_i = 0
\end{aligned}$$

where $S_k(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i y_i \alpha_j y_j K_k(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N} \alpha_i$. The optimal weights can be obtained by solving (6).

2.4 Random Forest for feature selection

Random Forest, proposed by L.Breiman and A.Cutler (2001), is an ensemble of decision tree classifiers. Each tree is grown as follows:

- If the number of cases in the training set is N, sample with replacement N cases at random from the original data. This bootstrapped data is used for training a tree.
- Given M input variables, a number $m << M$ is specified such that at each node, m variables are selected at random out of M and the best split on these m variables is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible, with no pruning.

To classify a new object, the input vector is put through each of the trees, which gives a classification, i.e., a vote. The classification having the most votes over all the trees in the forest is output. In addition to classification, Random Forests prove to be useful for feature selection, especially with biological data of heterogenous attributes (Reif et al 2006; Bryan et al 2001). In this work, we focused on the capacity of evaluating feature importance and feature interaction of this method.

For each tree, a number of cases are left out of training, called the "out-of-bag", which are used to estimate feature importance. If randomly permuting values of a particular feature m does not affect the predictive ability of trees on out-of-bag cases, m is assigned a low importance score, and vice versa. Subtract the number of votes for the correct class in the variable-m-permuted out-of-bag data from the number of votes for the correct class in the untouched out-of-bag data. The average of this number over all trees in the forest is the raw importance score for variable m. Assuming values of score from tree to tree are independent, z-score is computed by dividing raw score by standard error, with corresponding significance level (Breiman 2001).

Also, as defined by Breiman (2001), variables m and k interact if a split on one variable, say m, in a tree makes a split on k either systematically less possible or more possible. This method is reportedly able to uncover interactions among genes, proteins, and/or environmental factors that do not exhibit strong marginal effects (Reif et al 2006).

3 Experiment

3.1 Data preparation

Experimental data were retrieved from Yuan *et al.* (2005) and Liu *et al.* (2005), which covered nearly 4% of yeast genome including chromosome III and 223 additional promoter regions. The data from Yuan contain 50 bp DNA fragments tilled every 20bp. For each fragment we extracted its genomic sequence and hand-called state showing whether it is *well-positioned*, *delocalized* nucleosomal sequence or linker region. The data extracted from Liu contain measured levels of 12 different histone modifications, including acetylations of H3K9, H3K14, H3K18, H4K5, H4K8, H4K12, H4K16, H2AK7, H2BK16 and mono-, di- and tri-methylations of H3K4. We then filtered out the data corresponding to linker regions to keep only nucleosomal data. Too short nucleosomal sequences of less than 4 fragments were also removed. The resulted sequences were then truncated/extended, centered on the nucleosome, to windows of 150 bp if they were longer/shorter than that. Nucleosomes that miss histone modification values on all their fragments were removed. Ones with some fragments of missing values were filled with median of the rest. Each nucleosome was assigned either as *well-positioned* if its wrapped DNA stretches from 6 to 8 fragments or as *delocalized* if its wrapped DNA stretches more than 9 fragments. After these preprocessing steps, we obtained a dataset containing 1949 *well-positioned* and 297 *delocalized* nucleosomes, which were used for further analysis.

3.2 MKL classifiers

In our work, combined kernel was formed as following:

$$K = \beta_1 \times K_1 + \beta_2 \times K_2 \qquad (7)$$

where K_1 and K_2 are sub-kernel functions employed on genomic and PTMs features, whose representations are described in *Data representation.* For the purpose of model selection, we carried pilot experiments to select the most suitable kernel function for each kind of feature, including linear, polynomial (degree 2, 3, and

4), and RBF kernels. For both 2 types of features, linear kernel and *Radial Basis Function* (*RBF*) kernel ($K(\mathbf{x}_i, \mathbf{x}_j) = exp(-(\mathbf{x}_i - \mathbf{x}_i)^2)$) performed relatively better than the others (data not shown). Therefore, we then took linear and RBF as sub-kernels in MKL. Shogun toolbox (Sonnenburg et al 2006) was used for the tasks of training and testing MKL classifiers.

3.3 Performance evaluation

To evaluate the performances of the resulting classifiers, we used 10-fold cross-validation procedure. According to this procedure, each dataset was divided randomly into 10 subsets. The classifiers were trained on 9 subsets and tested on the remaining one. This training-testing procedure was repeated 10 times using a different hold-out set at each time. To measure the performances of the classifiers, we utilized the *receive-operator-characteristic* (ROC) curve. The quality of the classifier can be evaluated by calculating the *area-under-the-curve* (AUC), i.e., the “ROC score”, which is also considered a reasonable measure for cases of imbalanced data. A random classifier achieves the ROC score of 0.5 and a perfect classifier achieves the ROC score of 1.0.

4 Results and discussion

4.1 Genomic sequence and histone modifications partially contribute to affect nucleosome dynamics

Before going ahead with MKL, we carried two baseline experiments: SVMs separately for each feature sets and MKL for both without any weight, equivalent to simple feature concatenation. The purpose is to confirm the advantages of our proposed method in comparing the effect of each feature type. For both experiments, we employed 10-fold cross-validation as mentioned in Section 3.3 for evaluation. The results (details not included here) showed that (1) the average ROC-score of genomic-sequence-based SVM was 0.583, and PTMs-based one 0.605; (2) the MKL without weight reached ROC-score 0.611 ($p < 0.0005$ by one-sample t-test). From this, it is possible to see that using two separate classifiers does not tell much about the different effects and unweighed MKL dose not give very good performance. Lastly, we also tried to train a MKL with each single feature as a kernel; however, the output βs are minimally different thus hard to interpret (details not included here). Therefore, we used for this work 2 kernels for two types of genomic sequences and PTMs.

Our MKL approach proved to be better in both terms of performance and differentiation. 10-fold cross validation results with different combinations of kernel functions are given in Table 2. The average ROC score was 0.63 ($SD \approx 0.04$, $p < 0.0001$ by one-sample t-test) with RR combination of kernel functions. This is significantly higher than the performance of the random classifier. The result showed that the combination of genomic sequence and PTMs can be used to predict nucleosome stability state to some extend. This is consistent with previous investigations (Henikoff 2008; Segal and Widom 2009), reporting that *in vivo* besides DNA sequence and covalent modifications of histone proteins, among many other factors such as chromatin remodeling complexes, histone variants, etc., also contribute to stabilizing/destabilizing nucleosomes.

Turn	LL	LR	RR	RL
0	0.5	0.685	0.709	0.662
1	0.5	0.614	0.558	0.575
2	0.5	0.648	0.675	0.516
3	0.5	0.719	0.661	0.592
4	0.5	0.616	0.649	0.659
5	0.5	0.5	0.577	0.598
6	0.5	0.613	0.629	0.564
7	0.5	0.5	0.601	0.632
8	0.5	0.622	0.626	0.578
9	0.5	0.5	0.617	0.631
Avg	0.5	0.6017	0.6302	0.6007

LL, LR, RR, RL: Kernel functions for DNA sequence and post-translational modifications (PTMs) (L: Linear, R:RBF)

Table 2 ROC scores by 10-fold cross validation with different combination of kernel functions

4.2 The effect of histone modifications on nucleosome dynamics dominates that of genomic sequence

To derive optimal weights for genomic sequence and histone modifications, we used the combination of kernel functions that gave the best ROC score. According to the average ROC score given by 10-fold cross validation, RBF kernel was chosen for both kinds of features. Table 3 shows the learnt optimal weights in 10-fold cross validation settings. These results show that, the effect of post-translational modifications (PTMs)on nucleosome dynamics seems to dominate that of genomic sequence.

In vivo, beyond the difference in DNA sequence, nucleosomes also differ from each other by their histone

Turn	Sequence	HMs
0	0.33821	0.66179
1	0.331211	0.668789
2	0.412726	0.587274
3	0.351609	0.648391
4	0.353926	0.646074
5	0.357932	0.642068
6	0.370565	0.629435
7	0.328962	0.671038
8	0.360982	0.639018
9	0.382251	0.617749
Avg	0.3588374	0.6411626

Table 3 Optimal weights corresponding to PTMs and DNA sequence

compositions, either by PTMs or histone variants. Such changes can directly or indirectly affect nucleosome stability. For example, lysine acetylation is thought to destabilize nucleosome by directly changing its net charge (Waterborg 2002). Methylation of H3K27 or H3K9 are known to indirectly increase the stability of H3-containing nucleosomes by enabling the bindings of non-histone proteins PRC1 and heterochromatin protein 1 (HP1) to the methylated sites, respectively (Henikoff 2008). Previous investigation showed that, hyper-acetylation or complete removal of histone tails results in a small yet significant increase in the accessibility (Anderson et al 2001; Polach et al 2000) and stability (Widlund et al 2000) of nucleosomal DNA and in small sequence-dependent changes in positions of some nucleosomes, although sequences with high affinity are not affected (Yang et al 2007). This led to the conclusion that, histone modifications might have only modest direct effect on nucleosome dynamics. Taken together, our obtained result provides more evidence to elucidate the biological roles of histone modifications in regulating nucleosome stability. That is, in most cases, histone modifications may serve as the substrates for recruiting ATP-remodeling factors, which can then play a critical role in regulating nucleosome organization (Wan et al 2009).

4.3 Post-translational histone modifications individually and co-operatively regulate nucleosome stability

Post-translational histone modifications have long been known to be a *trans*-factor regulating various cellular processes, most importantly repression and activation of gene expression. Henikoff, in (2008), suggested that nucleosome destabilization plays an important role in gene regulation was affirmed; moreover, it occurs with the co-operation of epigenetic factors, hence, the hypothesis that these PMTs take part in gene regulation through regulating nucleosome stability. More concretely, they modulate chromatin structure, for instance, in that histone hyperacetylation may weaken DNA-histone contacts by neutralizing the positive charge of the histone tails and decreasing their affinity for negatively charged DNA; and conversely histone deacetylation is believed to prevent the access of biological machineries by restoring positive charge and strengthening the interactions between DNA and histones (Kurdistani and Grunstein 2003). These relationships among gene expression, nucleosome dynamics, and PTMs are possibly inferred either experimentally or computationally. Using Random Forest method, we evaluated the importance of each histone modification and uncovered pairwise interactions among them in discriminating nucleosome dynamics.

From Figure 4.3, it is clear that the difference between the most and the least important one, 28.599 and 42.872 respectively, is considerable (about 50%). Notably, H4K16Ac highly deacetylated can be found in open reading frames (ORFs) of highly-expressed genes (Wirén et al 2005). Both H3K4Me3 H2BK16Ac serve as binding marks for NF-Y, a binding factor transcriptional regulates genes of Drosophila and mouse (Tue et al 2011; Hou et al 2010).

From the results in Figure 4.3, we realized that H3K9Ac - H4H14Ac, H4K5Ac - H4K12Ac, H4K5Ac - H3K14Ac, H3K9Ac - H4K12Ac pairs might tightly cooperate in affecting nucleosome dynamics. Evidences for the cooperations of 3 first pairs among 4 can be found in literature as the following. Leroy *et al.* LeRoy et al (2008) has shown that, H4K5Ac, H4K12Ac, and H3K14Ac are binding marks of two closely associated the double bromodomain proteins Brd2 and Brd3, which are components of chromatin remodeling complexes (CRCs). Thus, we assume that these PTMs might cooperate to recruit bromodomain proteins, and CRCs thereof. Moreover, both Brd2 and Brd3 facilitate RNA polymerase II (PolII) to transcribe through nucleosomes (LeRoy et al 2008), which means that as components of CRCs they help delocalize nucleosome for the passage of PolII. Therefore, we can infer that the co-operations of these two PTM pairs (H4K5Ac - H4K12Ac and H4K5Ac - H3K14Ac with interaction levels of 25 and 67 respectively) are related to the instability of nucleosome.

Histone deacetylases Clr3 and Sir2 cooperatively functions throughout the genome, including the silent regions. And, the most significant acetylation sites are H3K14Ac for Clr3 and H3K9Ac for Sir2 at their genomic targets (Wirén et al 2005). These two PTMs are co-regulated, thus they assumably co-operatively function. It was also observed in (Wirén et al 2005) that H3K9Ac tends to be enriched and H4K14Ac to be low in intergenic regions (IGRs) and open reading frames (ORFs) of highly expressed genes. Our result showed that H4K14Ac and H3K9Ac (interaction level of 34) in concert affect nucleosome states. Hence, we suggest that these two paired PMTs promote gene expression through regulating nucleosomes in corresponding regions.

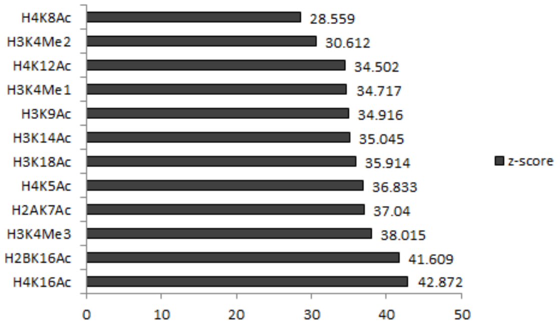
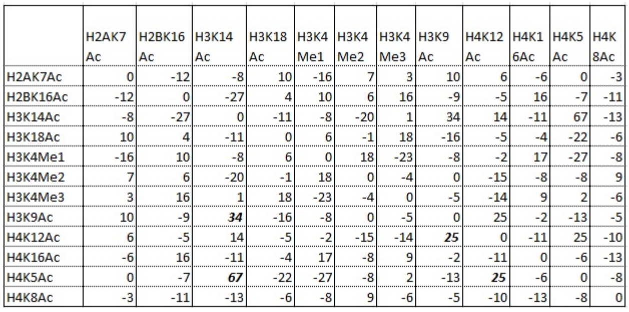


Fig. 1 Importance level of histone modifications



	H2AK7	H2BK16	H3K14	H3K18	H3K4	H3K4	H3K4	H3K9	H4K12	H4K1	H4K5	H4K
	Ac	Ac	Ac	Ac	Me1	Me2	Me3	Ac	Ac	6Ac	Ac	8Ac
H2AK7Ac	0	-12	-8	10	-16	7	3	10	6	-6	0	-3
H2BK16Ac	-12	0	-27	4	10	6	16	-9	-5	16	-7	-11
H3K14Ac	-8	-27	0	-11	-8	-20	1	34	14	-11	67	-13
H3K18Ac	10	4	-11	0	6	-1	18	-16	-5	-4	-22	-6
H3K4Me1	-16	10	-8	6	0	18	-23	-8	-2	17	-27	-8
H3K4Me2	7	6	-20	-1	18	0	-4	0	-15	-8	-8	9
H3K4Me3	3	16	1	18	-23	-4	0	-5	-14	9	2	-6
H3K9Ac	10	-9	***34***	-16	-8	0	-5	0	25	-2	-13	-5
H4K12Ac	6	-5	14	-5	-2	-15	-14	***25***	0	-11	25	-10
H4K16Ac	-6	16	-11	-4	17	-8	9	-2	-11	0	-6	-13
H4K5Ac	0	-7	***67***	-22	-27	-8	2	-13	***25***	-6	0	-8
H4K8Ac	-3	-11	-13	-6	-8	9	-6	-5	-10	-13	-8	0

Fig. 2 Interactions among histone modifications

5 Conclusion and Discussion

Nucleosome dynamics has been implicated in various important cellular processes, such as transcription, DNA repair and DNA replication. *In vivo* it can be influenced by many factors, such as DNA sequence, covalent modifications of histone proteins, chromatin remodeling complexes, etc. However, the contribution of each factors to the regulation of nucleosome dynamics remains elusive. In this work, we propose a computational approach based on Multiple Kernel Learning to quantitatively assess the effects of two important regulatory factors, i.e., genomic sequence and PTMs, on nucleosome dynamics. Our results on *S.cerevisiae* show that, though both factors partially contribute to the regulation of nucleosome dynamics, the effect of PTMs seems to dominate that of genomic sequence. By further analyzing of PTM pairs, we provide evidence that they also cooperate to influence nucleosome dynamics. This suggests that, in general, PTMs may serve as the substrates for recruiting other regulatory factors, such as non-histone proteins, to chromatin. By that way, they can express their regulatory effects on cellular processes.

In this paper, we mainly aimed at comparing the effects of genomic and epigenetic factors, thus formulating an MKL problem of two kernels. However, looking more closely into individual factor of these two types is more meaningful biologically. In that case, the number of features, hence kernels, may be very large. From our investigation, *infinite kernel learning*(IKL) (Özögür Akyüz and Weber 2008), a model developed from MKL idea but allowing infinite kernel combinations. Its theoretical base in term of learning algorithm is further improved by the same authors (2009; 2010). From practical perspective, Gehler and Nowozin (2008) have carried out experiments to compare SMV, MKL, and IKL on various benchmark datasets and concluded that regarding accuracy, for some combining kernel is of no benefit compared to SMV, for the some others IKL outperforms SMV/MKL. It is promising; however, we focus more on the interpretability of the coefficients learnt and the large number of kernels may in fact cause many of them to be assigned almost equally small number. Therefore, we plan to examine biological problem further with MKL and IKL in the future.

Acknowledgment

The authors would like to thank Dr. Guo-Cheng Yuan and Dr. Chih Long Liu for sharing experiment data. The first and second authors have been supported by Japanese Government Scholarship (Monbukagakusho) to study in Japan.

References

Özögür Akyüz S, Weber G (2008) Learning with infinitely many kernels via semi-infinite programming. In: In Proceedings of Continuous Optimization and Knowledge Based Technologies, 20th EURO Mini conference, pp 342–348

Özögür Akyüz S, Weber G (2009) Modelling of kernel machines by infinite and semi-infinite programming.

In: In Proceedings of the Second Global Conference on Power Control and Optimization, pp 306–313

Özögür Akyüz S, Weber G (2010) On numerical optimization theory of infinite kernel learning. J Global Optimization 48(2):215–239

Anderson J, Lowary P, Widom J (2001) Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites. J Mol Biol 307:977–985

Bach F, Lanckriet G, Jordan M (2004) Multiple kernel learning, conic duality, and the smo algorithm. In: Proc 21th Int Conf on Machine Learning (ICML)

Breiman L (2001) Randome forests. Machine Learning 45(1):5–32

Bryan K, Brennan L, Cunningham P (2001) MetaFIND: A feature analysis tool for metabolomics data. BMC Bioinformatics 9(470)

Gehler PV, Nowozin S (2008) Infinite kernel learning. Tech. rep., Max Planck Institute for Biological Cybernetics

Gupta S, Dennis J, Thurman R, Kingston R, Stamatoyannopoulos J, Noble W (2008) Predicting human nucleosome occupancy from primary sequence. PLoS Comput Biol 4(8)

Henikoff S (2008) Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet 9:15–26

Hou Y, Zhou X, Liu J, Yuan J, Cheng H, Zhou R (2010) Nuclear factor-y (NF-Y) regulates transcription of mouse dmrt7 gene by binding to tandem CCAAT boxes in its proximal promoter. Int J Biol Sci 6(7):655–664

Jansen A, Verstrepen K (2011) Nucleosome positioning in saccharomyces cerevisiae. Microbiol Mol Biol Rev 75(2):310–20

Kaplan N, Moore I, Fondufe-Mittendorf Y, Gossett A, Tillo D, Field Y, LeProust E, Hughes T, JD JL, Widom J, Segal E (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458(7236):362–366

Kurdistani S, Grunstein M (2003) Histone acetylation and deacetylation in yeast. Nat Rev Mol Cell Biol 4:276–284

Lanckriet G, Deng M, Cristianini N, Jordan M, Noble W (2004) Kernel-based data fusion and its application to protein function prediction in yeast. In: Proc Pacific Symposium on Biocomputing, pp 300–311

Le N, Ho T, Tran D (2009) Characterizing nucleosome dynamics from genomic and epigenetic information using rule inductiong learning. BMC Genomics 10(3:S27)

LeRoy G, BRickards, Flint S (2008) The double bromodomain proteins brd2 and brd3 couple histone acetylation to transcription. BMC Bioinformatics 30(1):51–60

Li B, Carey M, Workman J (2007) The role of chromatin during transcription. Cell 128(4):707–719

Liu C, Kaplan T, Kim M, Buratowski S, Schreiber S, Friedman N, Rando O (2005) Single-nucleosome mapping of histone modifications in s.cerevisiae. PLoS Biol 3(10):1753–1769

Luger K, Mäder A, Richmond R, Sargent D, Richmond T (1997) Crystal structure of the nucleosome core particle at 2.8 a resolution. Nature 389:251–260

Oliver S, Denu J (2011) Dynamic interplay between histone h3 modifications and protein interpreters: emerging evidence for a "histone language". Chembiochem 12(2):299–307

Peckham H, Thurman R, Y YF, Stamatoyannopoulos J, Noble W, Struhl K, Weng Z (2007) Nucleosome positioning signals in genomic DNA. Genome Res 17(8):1170–1177

Polach K, Lowary P, Widom J (2000) Effects of core histone tail domains on the equilibrium constants for dynamic DNA site accessibility in nucleosomes. J Mol Bio 298:211–223

Prost A, Dunleavy E, Almouzni G (2009) Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol 10:192–206

Ratsch G, Sonnenburg S, Schafer C (2006) Learning interpretable SVMs for biological sequence classification. BMC Bioinformatics 7(1:S9)

Reif D, Motsinger A, McKinney B, Crowe J, Moore J (2006) Feature selection using a random forests classifier for the integrated analysis of multiple data types. In: Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on, pp 1–8

Rippe K, Schrader A, Riede P, Strohner R, Lehmann E, Längst G (2007) DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodelling complexes. Proc National Academy of Sciences of the United States of America 104(40):15,635–15,640

Schnitzler G (2008) Control of nucleosome positions by DNA sequence and remodelling machines. Cell Biochem Biophys 51(2-3):67–80

Schwartz S, Meshorer E, Ast G (2009) Chromatin organization marks exon-intron structure. Nat Struct Mol Biol 16(9):990–995

Segal E, Widom J (2009) What controls nucleosome positions? Trends Genet 25(8):335–43

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore I, Wang J, Widom J (2006) A genomic code for nucleosome positioning. Nature 442(7104):772–778

Sonnenburg S, Ratsch G, Schafer C, Scholkopf B (2006) Large scale multiple kernel learning. J Machine Learning Res 7:1531–1565

Tanaka Y, Yoshimura I, Nakai K (2010) Positional variations among heterogeneous nucleosome maps give dynamical information on chromatin. Chromosoma pp 412–010

Tillo D, Hughes T (2009) G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics 10(442)

Tue N, Yoshioka Y, Yamaguchi M (2011) NF-Y transcriptionally regulates the drosophila p53 gene. Genes 473(1):1–7

Wan J, Lin J, Zack D, Qian J (2009) Regulating periodicity of nucleosome organization and gene regulation. Bioinformatics 25(14):1782–1788

Waterborg J (2002) Dynamics of histone acetylation in vivo. a function for acetylation turnover? Biochem Cell Biol 8:363–378

Widlund H, Vitolo J, Thiriet C, Hayes J (2000) DNA sequence-dependent contributions of core histone tails to nucleosome stability: differential effects of acetylation and proteolytic tail removal. Biochem 39:3835–3641

Windom J (2001) Role of dna sequence in nucleosome stability and dynamics. Q Rev Biophys 34(3):269–324

Wirén M, Silverstein R, Sinha I, Walfridsson J, Lee H, Laurenson P, Pillus L, Robyr D, Grunstein M, Ekwall K (2005) Genomewide analysis of nucleosome density histone acetylation and HDAC function in fission yeast. EMBO J 24(16):2906–2918

Yang Z, Zheng C, Hayes J (2007) The core histone tail domains contribute to sequence-dependent nucleosome positioning. J Biol Chem 282(11):7930–8

Yuan G, Liu Y, Dion M, Slack M, LF LW, Altschuler S, Rando O (2005) Genome-scale identification of nucleosome positions in S.cerevisiae. Science 309:626–630

Zhao J, Herrera-Diaz J, Gross D (2005) Domain-wide displacement of histones by activated heat shock factor occurs independently of swi/snf and is not correlated with RNA polymerase II density. Mol Cell Biol 25(20):8985–8999