HHS Public Access

Author manuscript

Stat Appl Genet Mol Biol. Author manuscript; available in PMC 2015 May 12.

Published in final edited form as:

Stat Appl Genet Mol Biol. ; 11(2): . doi:10.2202/1544-6115.1707.

A Generalized Hidden Markov Model for Determining Sequence-based Predictors of Nucleosome Positioning

Carlee Moser and
Boston University

Mayetri Gupta
Boston University

Abstract

Chromatin structure, in terms of positioning of nucleosomes and nucleosome-free regions in the DNA, has been found to have an immense impact on various cell functions and processes, ranging from transcriptional regulation to growth and development. In spite of numerous experimental and computational approaches being developed in the past few years to determine the intrinsic relationship between chromatin structure (nucleosome positioning) and DNA sequence features, there is yet no universally accurate approach to predict nucleosome positioning from the underlying DNA sequence alone. We here propose an alternative approach to predicting nucleosome positioning from sequence, making use of characteristic sequence differences, and inherent dependencies in overlapping sequence features. Our nucleosomal positioning prediction algorithm, based on the idea of generalized hierarchical hidden Markov models (HGHMMs), was used to predict nucleosomal state based on the DNA sequence in yeast chromosome III, and compared with two other existing methods. The HGHMM method performed favorably among the three models in terms of specificity and sensitivity, and provided estimates that were largely consistent with predictions from the method of Yuan and Liu (2008). However, all the methods still give higher than desirable misclassification rates, indicating that sequence-based features may provide only limited information towards understanding positioning of nucleosomes. The method is implemented in the open-source statistical software R, and is freely available from the authors' website.

Keywords

chromatin; nucleosome; DNA sequence; Bayesian modeling

1 Introduction

In complex organisms, DNA is tightly packed into the nucleus of cells, with stretches of DNA about 147 bp in length wrapped around histone proteins (*nucleosomes*) at approximately regular intervals, separated by nucleosome-free *linker* regions (Luger, 2006). Nucleosome-free regions (NFRs) are more susceptible to damage from environmental

agents. For example, mutations in regulatory regions of *oncogenes* can lead to the development of cancerous cells (Hershberg et al., 2005). In studies of chromatin it has been shown that active regulatory regions of the genome have a general tendency towards nucleosome disruption compared to non-regulatory regions (Wallrath et al., 1994). The availability of high-resolution nucleosome positioning data can complement genomic sequence data, leading to increasingly successful methods for discovering transcription factor binding sites (TFBSs) in complex organisms, a field which is typically plagued by high false discovery rates (Narlikar et al., 2007; Bussemaker et al., 2007; Gupta and Ibrahim, 2007).

1.1 Tiling arrays and their analysis

Recently, genome tiling array techniques (Dion et al., 2005; Casolari et al., 2005; Lee et al., 2007) including Formaldehyde Assisted Isolation of Regulatory Elements (Giresi et al., 2007; Hogan et al., 2006), have been used to map genomic positions of nucleosomes. Tiling arrays are a type of microarray chip designed to cover the whole or a major part of a genome through thousands of short fragments (probes) that are usually contiguous (or even overlapping). Tiling arrays are widely used in Chromatin Immunoprecipitation (ChIP-chip) for detection of TFBSs, determining DNA hypersensitivity sites (DNAse-chip) and array CGH to detect chromosomal copy number aberrations. In this article, we used publicly available tiling-array data from the Saccharomyces cerevisiae (yeast) genome (Yuan et al., 2005) for nucleosome detection. Yuan et al. (2005) used a procedure of shearing chromatin by micrococcal nuclease (MNase) digestion to locate nucleosomal positions for a set of regions covering about 270 Kbp and sixteen chromosomes in yeast. In this procedure, nucleosomal DNA was isolated by MNase treatment, labeled with Cy3 fluorescent dye, mixed with Cy5-labeled whole-genomic DNA, and hybridized to a microarray. The DNA probes used were 50 bp in length and overlapped with their neighbors by 30 bp. The final data contained about 25,000 short overlapping DNA segments, corresponding to the microarray probes; for each probe, the available information consisted of intensity measurements from the nucleosome-enriched and reference sample.

The spatially dependent structure of the tiling array suggests that models explicitly incorporating this dependence are likely to be more powerful in detecting true protein-DNA interactions. Hidden Markov models, or HMMs (Juang and Rabiner, 1991), are often used in such contexts; HMMs consist of a doubly stochastic process where a latent Markov process is inferred through observations from another set of stochastic processes. HMMs are not directly appropriate for assessing length-constrained features such as nucleosomes, as they induce exponentially decaying state length distributions. Recently, we introduced a generalized Bayesian framework (Gupta, 2007) for statistical inference from genome tiling arrays, developing a hierarchical model robust to various sources of probe variability and measurement error and an explicit state duration model.

1.2 Determining nucleosome positioning from sequence

The effect of DNA sequence on nucleosome positioning is known to be important, but is still not completely understood (Ioshikhes et al., 2011; Trifonov, 2011; Segal et al., 2006; Ercan and Lieb, 2006; Giresi et al., 2006). Nucleosome positioning (NP) is known to be

influenced by poly-nucleotide and periodic repeats (Ioshikhes et al., 2011; Thastrom et al., 2004; Wang and Widom, 2005), as well as homopolymer stretches (Yuan et al., 2005). Some studies predicted up to 50% of nucleosome positions using DNA sequence (Ioshikhes et al., 2006; Segal et al., 2006); recent evidence shows that rather than a few discrete sequences influencing NP, cumulative effects over long DNA stretches are likely to be important (Ercan and Lieb, 2006; Trifonov, 2011). Segal et al. (2006) used a dinucleotide-based frequency model to differentiate potential nucleosomal and non-nucleosomal regions from a well-positioned subset; and Ioshikhes et al. (2006) used the propensity of periodically distributed AA and TT dinucleotides to define a NP sequence. Recent studies, such as Yuan and Liu (2008), have developed algorithms to differentiate nucleosomal regions, and to calculate overall nucleosome occupancy likelihoods. Yuan and Liu (2008) applied wavelet analysis to determine signals, which were used to model the probability that DNA sequence is part of a nucleosome via logistic regression. The predicted logits, known as N-scores, were used to classify each sequence as a linker or nucleosome. Yuan and Liu (2008) compared their method to other nucleosome classification approaches using an ROC-score, the area under an ROC curve. The methods proposed by Ioshikhes et al. (2006) and Segal et al. (2006) use a non-discriminative approach, and only consider nucleosome sequence data. In addition, these approaches focus only on the nucleotide and dinucleotide level counts.

There is unlikely to be enough signal in the sequence surrounding any one nucleosome to know which individual bases are important for positioning; our goal is to use data from many genome-wide positioning events to derive rules to build meaningful models. Our proposed approach incorporates statistical methods that simultaneously learn sequence properties related to (i) polynucleotide frequencies and (ii) spatial correlations in sequence data that influence NP, leading to a more complete characterization of an NP sequence. We have developed an efficient Bayesian statistical model and methodology, based on a hierarchical generalized hidden Markov model framework, to determine nucleosomal positioning locations by using ChIP-chip tiling array data, that accounts for spatial dependence between probes (Gupta, 2007). In this article, we propose a novel segmentation-based probabilistic model for predicting chromatin structure on the basis of underlying sequence. Sequence features can be tested for predictive ability with the goal of predicting nucleosome positioning and TF binding propensity from sequence factors alone. The model can be estimated through a classical likelihood based approach or a Bayesian approach. We prefer the Bayesian approach for two main reasons: (i) it provides a framework for hierarchically modeling dependence and for dealing with nuisance parameters (such as probe-specific biases) without leading to overwhelming analytical complexity and (ii) it allows a principled way of building prior distributions based on partially known information (such as TFBS patterns) and hence improve estimation of novel features.

2 Methods

We develop a probe-specific model for tiling array data for analyzing nucleosome positioning experiments. The spatial dependence between probes, along with the varying state length assumption, is addressed through a generalized hierarchical hidden Markov model (HGHMM) approach (Gupta, 2007). To allow for flexible modeling of the distribution of latent states, we use a non-homogeneous HMM approach. A two-state model

is developed, with the nucleosomal and nucleosome-free regions corresponding to the hidden states. In the Bayesian approach, we can hierarchically model probes and efficiently pool data to obtain robust parameter estimates. The new approach further allows state-specific transition distributions which we incorporate in two ways. First, the length of sequence generated from an underlying state is allowed to depend on the state identity. Next, the emission densities are allowed to depend on location-specific covariates, which lets us take into account the effect of local sequence composition on the observed binding propensity of a region. Fitting this more complex model in the Bayesian set-up is computationally expensive, especially if using standard Markov chain Monte Carlo fitting methods such as Gibbs sampling (Gelfand and Smith, 1990). For efficient computation, we make use of a recursive data augmentation (Tanner and Wong, 1987) technique which has previously been developed for segmentation-type models (Gupta and Liu, 2005; Gupta, 2007).

We here develop a two stage approach for determining sequence-based characteristics that predict nucleosome positioning. Nucleosome positioning sequence signals have previously been studied in terms of short nucleotide repeats (Ioshikhes et al., 2006; Segal et al., 2006) but these signals are generally too weak to give meaningful predictions in genome-wide analysis. We therefore adopt a reverse approach. Instead of testing for significance of particular sequence signals in predicting nucleosome positioning, we develop a two-state hierarchical HMM, where at the coarsest level, different segment types may potentially have different nucleotide compositions. Sequence-specific characteristics are incorporated into the model as covariates, and the increase in predictive power is tested by comparing to nucleosome positioning data where the true states are known with some accuracy.

2.1 Model for determining sequence determinants of occupancy state

2.1.1 Model description—For notational simplicity, let us represent the ChIP-chip data as a single sequence of observations Y_i, $i = 1, \ldots, N$. Y_i represents the logarithm of the intensity ratio between the enriched and reference sample for probe i of the microarray. Corresponding to each observation, let us assume an unobserved state C_i, ($i = 1, \ldots, N$), where $C_i = 1(0)$ represents a nucleosome-rich (nucleosome-free) state. Also, let $X = (X_1, \ldots, X_N)$ denote measurements for a p–dimensional set of "a priori" sequence-based predictors, where $X_i = (X_{i1}, \ldots, X_{ip})$ for probe i. The p predictors could typically represent sequence-specific scores, such as 1-mers, 2-mers, motif-based scores, or motif-cluster-specific scores.

Our aim is to predict the best set (or combination) of predictors that can predict the class states *C a priori*, after training our model on a set of experiments to determine nucleosome positioning. We use a flexible hidden Markov model-type approach to incorporate (i) possible dependence in measurements of neighboring probes and (ii) linking the covariates (sequence-based characteristics) to the response of interest (nucleosomal state). Adapting the approach from Gupta (2007) the other components of the model are:

1. The initial distribution of states, characterized by the probability vector $\pi = (\pi_0, \pi_1)$. A Dirichlet prior is used for π.

2. The probability of spending time d in state k, given by the distribution $p_k(d|\varphi)$, $d \in D_k$ ($0 \le k \le 1$), characterized by the parameter $\varphi = (\varphi_0, \varphi_1)$. Here we let D_1 (length of a nucleosomal state) vary in the range {6, …, 30} to allow for well-positioned nucleosomes (covered by 6 to 8 probes) as well as temporally varying unstable nucleosomes (between 9 to 30 probes), as suggested by prior biological data. D_0 is unrestricted and can take any positive integer value. $p_k(d)$ is chosen to be a truncated negative binomial distribution, between the range specified by each D_k. More precisely,

$$p_k(d)=c_k \begin{pmatrix} d-1 \\ r_k-1 \end{pmatrix} (1-\phi_k)^{d-r_k}\phi_k^{r_k}, d \in D_k=\{r_k, r_k+1, \ldots, s_k\} \quad (1)$$

where the normalizing constant $c_k=\left[\sum_{d=r_k}^{s_k} \begin{pmatrix} d-1 \\ r_k-1 \end{pmatrix} (1-\phi_k)^{d-r_k}\phi_k^{r_k}\right]^{-1}$. A conjugate Beta($\gamma_k$, δ_k) prior is assumed for φ_k.

3. Emission model. If C_i's are independent (which they are not), a natural way of relating C_i's to sequence specific predictors would be through a logistic link function:

$$g(C_i)=\frac{\exp(X_i'\beta-\mu)}{1+\exp(X_i'\beta-\mu)}$$

so that for a new X^*, we could predict states using

$$P(C_i^*=1|X_i^*)=\frac{\exp(X_i^{*'}\beta-\mu)}{1+\exp(X_i^{*'}\beta-\mu)}, \quad (2)$$

where $\beta = (\beta_1, \ldots, \beta_p)$ is a p-dimensional regression coefficient vector, and μ a scalar intercept term. To incorporate the dependent nature of adjacent probes, within the framework of the HMM, we define $Z_i = X'_i\beta$ and note that the right side of (2) is equivalent to the probability distribution function of a logistic distribution, that is, for every i, $P(C_i = 1|X_i)$ is equivalent to $P(Z_i > 0)$, where Z_i can be interpreted as a measurement on a latent variable. This formulation can be thus considered equivalent to using a logistic emission distribution on Z_i within the HMM, i.e.

$$f(z_i|c_i)=\frac{\exp[-(z_i-\mu_{c_i})]}{[1+e^{-(z_i-\mu_{c_i})}]^2}, -\infty<z_i<\infty.$$

where μ_c denotes the probe mean for state c ($c \in \{0, 1\}$).

4. Transition model. The transition probabilities between the states $\tau_{jk} = P(C_i = k|C_{i-1} = j)$, are given by the matrix $\tau = (\tau_{jk})$, $(0 \le j,k \le 1)$. Assume a Dirichlet prior for state transition probabilities, i.e. $\tau_{k0}, \tau_{k,1} \sim Dirichlet(\eta)$, where $\eta = (\eta_0, \eta_1)$.

Hyperparameters for the Dirichlet and Beta prior densities are chosen to be non-informative. Our model is fitted using a cross-validation algorithm, trained on a gold standard data set, and then applied to a test set. Below we detail the two sets of steps that comprise the algorithm.

2.1.2 Model training

Step 1. Determine the nucleosomal state, *C*, for each probe in the training data. This may be done using a profile HMM (Yuan et al., 2005) or a Bayesian data augmentation algorithm under an HGHMM model (Gupta, 2007).

Step 2. Train model with predictors *X* (sequence-based covariates: word counts, or principal components derived from word count matrix, discussed later) that can predict *C* in the training data set. In this step, we assume the states are known (from Step 1), and we estimate the parameters β_c and μ_c for each state $c = \{0, 1\}$ in the training data using standard likelihood-based approaches.

2.1.3 Model testing and prediction

Step 1. With all sequence-based covariates *X** in the test data set (corresponding to *X* in the training data set), we fit a new generalized HMM, the model which is detailed in Section 2.1.1. Here, we use the notation C_j^* to denote the fitted state of probe *j* in the test data set. In this step, we iteratively do the following:

- Determine latent nucleosomal states C_j^*, ($j = 1, \ldots, N$) for the *N* probes in the test set using a recursive data augmentation procedure that simultaneously estimates states and state durations. The details of this step, adapted from Gupta (2007), are given in the Appendices. In contrast to Gupta (2007), a logistic distribution is used in place of a hierarchical Gaussian model.
- Estimate transition probabilities τ_{kl} ($0 \le k, l \le 1$) by sampling from their posterior (Dirichlet) distributions. Although this could be potentially done during model training, estimating these instead in the testing stage allows greater flexibility in adapting to the nucleosomal landscape that may vary across different regions of the DNA.
- Estimate initial state distribution parameters π and state duration distribution parameter φ_k ($k = 0, 1$) from their posterior distributions. Sampling π is straightforward due to its conjugate prior distribution; for sampling φ_k efficiently, an adaptive rejection Metropolis (ARMS) algorithm is used, similarly as in Gupta (2007).

Step 2. After fitting the HGHMM, we estimate posterior probabilities $P(C_i^*=1)$ for any subset/subsequences of interest, based on the posterior samples from the MCMC algorithm. Alternatively, a Viterbi algorithm could be used to predict the states after estimating the emission and transition parameters. However, since the full MCMC-based sampling gives estimates from the joint distribution of parameters, rather than the conditional distribution (as in Viterbi), we prefer to use this approach when feasible. In

typical runs of our algorithm, it converged within a few iterations, hence was not especially computationally intensive.

As discussed in more detail in the following section, we applied this method on the Yuan et al. (2005) data set, through ten-fold cross-validation.

3 Results and Empirical Studies

3.1 HGHMM Analysis

Tiling-array data for the Saccharomyces cerevisiae genome (Yuan et al., 2005) were used to assess the performance of the proposed generalized hidden Markov model. For each of about 25,000 DNA probes, the following information was available: DNA sequence start and end coordinates, chromosome of occupancy, and nucleosomal state predicted by Yuan et al. (2005). The nucleosomal states indicated whether a given 50 bp segment of DNA was a linker or nucleosome-free region (NFR), a nucleosome, or a fuzzy nucleosome.

We first used a logistic regression model, which requires a dichotomous outcome; therefore fuzzy nucleosomes were also specified as nucleosomes. The largest chromosomal region for our data is chromosome 3, which represents 57% of the total set of probes. This region, which has the fewest number of sequence gaps, was used for our analysis. Two additional exclusion criteria were also used, which further reduced the data size to 12261 probes. Probes with missing nucleosomal or other information were not included in the analysis; also, nucleosomal regions that were composed of less than 5 probes were excluded. Five contiguous overlapping probes were equivalent to 130 bp of DNA sequence, and nucleosomes are ~ 147 bp in length. This final probe distribution was used for analysis.

The DNA sequence was used to predict the nucleosomal states of each of the probes. To relate the nucleotides of the DNA sequence to the states, model covariates were obtained from DNA words. DNA words are smaller sub-segments of the sequences of varying length. There exist 340 possible one, two, three, and four letter word combinations, formed from the four nucleotides (A, C, G, and T) which compose DNA. For each of the 12261 overlapping probes, the count of each word was calculated. The word counts were then transformed using principal components analysis to account for the correlation due to the overlapping nature of the probes and the words. Orthogonal covariates, based on the principal components (PCs) were computed– they consisted of different parts of the 340 word counts and were ordered by the percentage of variability they explained. The first ten PCs, which explained 67% of the variability in the data, were selected as the covariates for the models. A larger set of PCs (26) which explained about 80% of the variability, was also considered for the analysis, but similar results were seen– hence we chose the smaller set for computational efficiency.

Our initial analysis used the 10 covariates as predictors in a multiple logistic regression analysis. The outcome of interest was nucleosomal state, and we modeled the probability of a probe being a nucleosome free region (NFR), across chromosome 3. This modeling strategy ignores the underlying correlation structure of the probes. To assess the predictive accuracy of the logistic model, cross validation strategies were used. The data were stratified

into 10 groups of equal size, 1226 probes per group. For each of the 10 subgroups, the nucleosomal state of each probe was predicted based on the combined data from the remaining groups. The algorithm predicted the average percentage of nucleosomal regions to be 84%, across all ten sets. The results of the cross validation are presented in Table 1; all values are calculated using a 0.5 cutoff for the posterior probability.

The overall misclassification rate of the method is a combination of the false positive and false negative predictions. The evaluated performance characteristics of the prediction algorithm are as follows:

Sensitivity = P(Predicted NFR | NFR)

Specificity = P(Predicted nucleosome | nucleosome)

False Negative = P(Predicted nucleosome | NFR)

False Positive = P(Predicted NFR | nucleosome)

The logistic analysis had low sensitivity on average, yet high specificity. This indicates that the model was not good at detecting NFRs, but was more successful at detecting nucleosome regions. (Changing the cutoff from 0.5 may bring these values closer to the HGHMM predictions, but the misclassification rates are still substantially higher.) Next, the HGHMM model was applied to the data using a logistic emission distribution, and also a normal approximation to the emission distribution. The data were again divided into 10 different test sets of size 1226 probes.

Prediction and cross-validation analysis was conducted for each test set. For each test set, the HGHMM model fit was run for 1000 iterations of the MCMC samples and the predictions with the largest posterior probability were selected. Misclassification rates and other prediction assessment rates were calculated - as well as receiver operator curves. The output for the HGHMM logistic and HGHMM normal approximation for the emission distributions are also shown in Table 1. The predictions in the table are classified with a 0.5 cutoff for the posterior probability. The two HGHMM-based methods yielded similar results with sensitivity levels of around 0.54, and specificity levels of around 0.68. The methods seem to classify nucleosome-rich regions with greater accuracy than the NFRs, however, even the NFR classification was improved compared to the crude logistic model-based estimation.

3.2 Comparison of HGHMM with other methods

The HGHMM method showed promising results, compared to the non-HGHMM logistic method. Additional comparisons were done between the HGHMM method and other approaches to modeling nucleosome positioning with tiling-array data– Yuan and Liu (2008) and Segal et al. (2006). The *S. cerevisiae* tiling-array data were analyzed with both these methods and compared to the HGHMM results.

Yuan and Liu (2008) developed an algorithm to predict nucleosome positioning by modeling covariates using wavelet analysis. This method models the probability that a sequence segment is part of a nucleosome, and is expressed as the predicted log-odds,

known as the N-score. The N-score computation algorithm requires sequences to be 131 bp in length, and thus our tiling-array data were recombined into overlapping segments of 131 bp DNA sequences. Five consecutive tiling-array probes, which cover 130 bp of sequence, were combined along with one additional base-pair to create the 131 bp sequences. Each of the following 131 bp sequences were found by shifting the previous by 20 bp, equivalent to a probe shift. Gaps in the sequence were defined as any region with more than one missing probe. The 131 bp sequences were generated as above until a gap was encountered. Because of the probe overlap, the sequence remained continuous if only one probe was missing, otherwise a gap was recorded. In total, 10727 overlapping sequences of 131 bp length were created. These sequences were analyzed using the Yuan and Liu (2008) N-score method in Matlab. All methods and software were obtained from the Yuan and Liu (2008) website. An N-score for each 131 bp sequence represented the log-odds of a nucleosome being positioned along the given sequence. N-score values smaller than zero were identified as NFRs, and N-score values larger than zero were identified as nucleosomes. In order to compare the predictions from the N-score method to the true nucleosomal states, the true states were also transformed to a 131 bp resolution. For each 131 bp sequence, the proportion of true NFR probes was calculated and was used to classify the new 131 bp sequences as NFR or nucleosomal. The proportion was calculated based on non-missing values, such that if one probe was missing, only the true states of the non-missing probes were included in the calculation of the NFR percentage. If the proportion of NFR probes was greater than 0.5 then the new sequences were classified as NFR, and if the proportion was less than 0.5, then the new sequences were classified to be nucleosomal.

The second method, the Segal et al. (2006) prediction algorithm, models the probability that a basepair is located within a nucleosome region. The tiling-array data was combined to form non-overlapping segments of continuous DNA. In total, there were 442 gaps, which results in 443 continuous sequences of varying length. Each continuous sequence was analyzed with Segal et al. (2006) method, and the probability that each base pair was part of a nucleosome was calculated. The resulting bp resolution probabilities, corresponding to each of the original probes, were averaged to obtain an overall estimate for the nucleosomal probability for each probe. The bp probabilities from Segal et al. (2006) were averaged over the 50 bp length to compare to the HGHMM at the individual probe level. Differing cutoffs were used to classify the findings and are seen in Figure 1, which compares the classification of the NFR regions for the HGHMM and Segal et al. (2006) analyses with a receiver operator curve (ROC). The ROC shows that the HGHMM results, which were combined across test sets, outperform those of the Segal et al. (2006) approach, when comparing at the probe level.

The classifications for the three approaches compared to the true states are displayed in Table 2. To compare all the results simultaneously, the Segal and HGHMM results were also transformed to 131 bp resolution (rows D and E in Table 2). The NFR classification for the Segal and HGHMM results, at the 131 bp resolution, were done with the same method as the true state classification. The Segal method produces a probability for every basepair that it is part of a nucleosomal region. We took each segment out of the 443 segments (produced by gaps in the data) and divided it into non-overlapping segments of 131 bp (excluding any basepairs left over at the ends). Next, we calculated the average nucleosomal probability of

the basepairs within each sequence to assign one value to each 131 bp sequence. If this probability exceeded 0.5, the 131 bp segment was assigned to have been predicted a nucleosome, otherwise it was considered nucleosome-free. For the HGHMM, we recombined the overlapping probes into non-overlapping 131 bp sequences, and then averaged the posterior probabilities of being predicted a nucleosome within that segment. Each segment was assigned to have been predicted a nucleosome or NFR based on whether this probability was greater or less than a cutoff of 0.5. Finally, for the N-score method of Yuan and Liu (2008), each 131 bp segment was assumed to be predicted a nucleosome or NFR based on whether the N-score (which is on the logit scale) was greater or less than zero. The HGHMM logistic model slightly under-predicted the nucleosomal state percentages (in this data this is estimated as 62.9%). The Yuan and Liu (2008) method appeared to strongly under-predict nucleosomes, whereas the Segal et al. (2006) approach over-predicted the values.

The misclassification rates are shown, along with the predicted percentages, in Table 3. The Yuan and Liu (2008) approach and the HGHMM had comparable results for sensitivity and specificity, the HGHMM having a 3.2% lower misclassification rate overall. The Segal et al. (2006) approach had poor NFR prediction, but was stronger at predicting the nucleosomal regions. When examining the Segal results at the probe level, the percentage of nucleosomes decreases slightly; the results are displayed in Table 2.

The nucleosome predictions from the three approaches were also compared to determine the amount of overlap between the methods, in terms of predicted states. The least amount of mismatch was between the Yuan and HGHMM methods, with 28.4% of the predictions being concordant for nucleosomes, and 32.9% concordant for NFRs. The comparisons between the Segal method and the others more agree for the nucleosome regions than the NFRs. Examining the three-way comparison of the methods, it is clear that the methods are not necessarily consistent in their predictions. The Segal method does not have strong predictive ability when examining sequences at the 131 bp resolution. NFR predictions were not similar across methods; however, the nucleosomal predictions were comparable. Details of these comparisons are in Tables A1 and A2 in the Appendices.

The Area Under the Curve (AUC) for the Segal analysis and for the HGHMM logistic analysis, both at the probe level, from Figure 1 were calculated to be 38.63% and 62.25%, respectively. A similar calculation was done for the three methods at the 131 bp resolution. The AUCs for the Yuan, Segal, and HGHMM methods are 57.41%, 43.63%, and 58.93%, respectively. This summary measure, AUC, is similar to the ROC score assigned to the different approaches outlined in Yuan and Liu (2008). The initial ROC scores reported in Yuan are based on a different data set composed of 199 nucleosome sequences and 292 linker sequences, which is used to train for cross-validation and prediction. The second set of ROC scores, based on genome-wide predictions, were obtained using nucleosome-enriched probes (highest 10% of log-ratios) and NFR-enriched probes (lowest 10% of log-ratios). Hence our AUC values appear slightly lower than seen in Yuan and Liu (2008), being more conservatively estimated.

4 Discussion

Our nucleosomal state prediction approach incorporates a large amount of flexibility through segment-specific transition distributions and hierarchical modeling. For fitting and testing models, while limiting the computational cost, we made use of efficient Monte Carlo procedures such as recursive data augmentation. For larger data sets, it may be possible to use numerical and analytical approximations at various stages that will speed the computation by orders of magnitude without compromising the predictive power of the model.

Results were similar across the HGHMM Logistic and HGHMM Normal model; and both HGHMM-based approaches have higher AUC values than the crude logistic model. The non-HGHMM logistic model over-predicts nucleosome occupancy with a predicted percentage of 84%. The true percentage of nucleosome occupancy is 62.9%. The estimated percentages for each of the two HGHMM methods with different emission distributions are about 60%. The HGHMM logistic and the approximated approach with the HGHMM normal both give similar results. Contrasting the the HGHMM model with Segal et al. (2006) and Yuan and Liu (2008), the HGHMM model appears to be most consistent with the Yuan and Liu (2008) approach. The Yuan method under-predicts the nucleosomal percentages, while the Segal method over-predicts nucleosomal occupancy both at the probe and the 131 bp levels. The main reason for the discrepancy of the Segal method is potentially the lack of a training set for nucleosome-free regions, only concentrating on a known set of nucleosomal regions.

One ultimate goal of this approach is to get a sense of how each sequence feature contributes (or is unrelated) to nucleosome positioning. Table 4 shows that a number of sequence features were strongly related to nucleosome positioning, including A/T-containing dimers that have already been implicated in nucleosome positioning in other studies. In particular, we observed that A/T-containing dimers and trimers were the top contributors to the 1st, 4th, 5th and 7th PC. Also, the 3rd PC which was most strongly correlated with nucleosome positioning, was seen to heavily depend on C- and G- containing k-mers, which suggests that there may be mechanisms at work other than the rigidity of the DNA alone in positioning nucleosomes.

In summary, the HGHMM approach appears to give overall lower misclassification rates compared to other methods, and has great flexibility in use and interpretation, as it can be used directly on any data without the need for any specific subdivision into pre-specified windows (such as the 131-bp windows necessary in Yuan and Liu (2008)) which can constitute a problem, for example, in data containing gaps and missing probes. In addition, this method provides a direct interpretability in terms of how different sequence features contribute to nucleosome positioning; by looking at the weight age of each sequence feature within the principal components used to fit the model, we can directly see how much each k-mer is related to positioning of nucleosomes or NFRs. In conclusion, sequence factors appear to be generally indicative of differences between nucleosomal and nucleosome-free regions, but may have limited predictive power. Other chromatin measurements such as crystalline structure of the DNA (Greenbaum et al., 2007) may need to be integrated into

nucleosomal positioning models for maximal predictive efficiency. Moving on from a two-step discriminative approach to develop a unified framework to estimate the regression parameters simultaneously with fitting the HMM would be ideal. However, given the small proportion of variability seemingly explained by sequence-based characteristics alone, this approach would probably only be successful if further relevant biological data could be incorporated into such a model.

Acknowledgments

The authors would like to thank Jason Lieb and Greg Hogan for making the yeast nucleosomal array data available and for numerous discussions and insights. This research was supported in part by the NIH/NHGRI award HG004946.

Appendices

A1.1 Bayesian data augmentation algorithm for HGHMM state prediction

This is adapted from Gupta (2007). For notational simplicity, assume a single long sequence of length N, $Y = \{y_1, \ldots, y_N\}$, with r replicate observations for each $y_i = (y_{i1}, \ldots y_{ir})'$. If there are gaps, each separated segment of the sequence should be taken separately, and the same procedure repeated for each segment. Let the set of all parameters be generically denoted by $\theta = (\mu, \tau, \varphi, \pi)$, and let the latent variables $C = (C_1, \ldots, C_N)$ and $L = (L_1, \ldots, L_N)$ denote the state identity and state lengths, where $L_i = l$ is a non-zero number denoting the state length if it is a point where a run of states ends. Then,

$$L_i = \begin{cases} l & \text{if} C_{i+1} \neq C_i = C_{i-1} = \cdots = C_{i-l+1} = k \neq C_{i-l} \text{for some} k \in \{1, \ldots K\}, \\ 0 & \text{otherwise.} \end{cases}$$

The observed data likelihood then may be written as:

$$L(\theta; Y) = \sum_C \sum_L p(Y|C, L, \theta) P(L|C, \theta) P(C|\theta) \quad (3)$$

Recursive data augmentation

In the data augmentation algorithm, the key is to update the states and state length durations in an recursive manner, after calculating the required probability expressions through a *forward summation* step. Let an indicator variable I_t take the value 1 if a segment boundary is present at position t of the sequence, meaning that a state run ends at t ($I_t = 1, \Leftrightarrow L_t \neq 0$). In the following, the notation $y_{[1:t]}$ is used to denote the vector $\{y_1, y_2, \ldots y_t\}$. Define the partial likelihood of the first t probes, with the state $C_t = k$ ending at t after a state run length of $L_t = l$, by the "forward" probability:

$$\alpha_t(k, l) = P(C_t = k, L_t = l, I_t = 1, y_{[1:t]}).$$

Also, let the state probability marginalized over all state lengths be given by

$$\beta_t(k)=\sum_{l=r_k}^{s_k}\alpha_t(k,l) \quad (4)$$

Let $d_{(1)} = \min\{D_1, \ldots, D_K\}$ and $d_{(K)} = \max\{D_1, \ldots, D_K\}$. Then, assuming that the length spent in a state and the transition to that state are independent, i.e. $P(l,k|l', k') = P(L_t = l|C_t = k)\tau_{k'k} = p_k(l)\tau_{k'k}$, we have

$$\alpha_t(k,l)=\sum_{k'\neq k}\sum_{l'\in D_{k'}}\alpha_{t-l}(k',l')P(l,k|l',k')P(y_{[t-l+1:t]}|C_t=k) =P(y_{[t-l+1:t]}|C_t=k)p_k(l)\sum_{k'\neq k}\tau_{k'k}\beta_{t-l}(k'), \quad (5)$$

for $2 \le t \le N$; $1 \le k \le K$; $l \in \{d_{(1)}, d_{(1)} + 1, \ldots, \min[d_{(K)}, t]\}$. To complete the calculation, the boundary conditions needed are: $\alpha_t(k, l) = 0$ for $t < l < d_{(1)}$, and $\alpha_l(k, l) = \pi_k P(y_{[1:l]}|C_l = k)p_k(l)$ for $d_{(1)} \le l \le d_{(K)}$, $k = 1, \ldots, K$. $p_k(\cdot)$ denotes the k-th truncated negative binomial distribution given in (1).

The states and state duration lengths (C_t, L_t) $(1 \le t \le N)$ can now be updated, for current values of the parameters $\theta = (\mu, \tau, \varphi, \pi)$, using a *backward sampling*-based imputation step.

Algorithm

1. Set $i = N$. Update $C_N|y, \theta$ using

$$P(C_N=k|y,\theta)=\frac{\beta_N(k)}{\Sigma_k\beta_N(k)}.$$

2. Next, update $L_N|C_N = k, y, \theta$ using

$$P(L_N=l|C_N=k,y,\theta)=\frac{P(L_N=l,C_N=k|y,\theta)}{P(C_N=k|y,\theta)}=\frac{\alpha_N(k,l)}{\beta_N(k)}.$$

3. Next, set $i = i-L_N$, and let $LS(i) = L_N$. Let $D_{(2)}$ be the second smallest value in the set $\{\min D_1, \ldots, \min D_K\}$. While $i > D_{(2)}$, repeat the following three steps:

- Draw $C_i|y, \theta, C_{i+LS(i)}, L_{i+LS(i)}$ using

$$P(C_i=k|y,\theta,C_{i+LS(i)},L_{i+LS(i)})=\frac{P(C_i,C_{i+LS(i)}|L_{i+LS(i)},y,\theta)}{P(C_{i+LS(i)},L_{i+LS(i)},y,\theta)} =\frac{\beta_i(k)\tau_{kC_{i+LS(i)}}}{\Sigma_k\beta_i(k)\tau_{kC_{i+LS(i)}}},$$

  where $k \in \{1, \ldots, K\} \setminus C_{i+LS(i)}$, the simplification resulting from the assumption that the duration in the previous state and the next state transition are independent events.

- Draw $L_i|C_i, y, \theta$ using

$$P(L_i=l|C_i,y,\theta)=\frac{\alpha_i(C_i,l)}{\beta_i(C_i)}.$$

- Set $LS(i-L_i) = L_i$, $i = i-L_i$.

Note that the proposed sampling algorithm is generally applicable to any length restricted HMM and not limited to the forms of the state-specific distributions used here. Once the states and state duration lengths (C_i, L_i) ($1 \le i \le N$) have been updated, updating the parameters from their posterior distributions is straightforward.

A1.2 Additional tables

Table A1

Two-way Classification for Yuan, Segal, and HGHMM methods at 131 bp level.

Yuan vs. Segal			
	Segal NUC	Segal NFR	Overall Mismatch
Yuan NUC	3673	504	0.6005
Yuan NFR	5938	612	
Yuan vs. HGHMM			
	HGHMM NUC	HGHMM NFR	Overall Mismatch
Yuan NUC	3049	1128	0.3863
Yuan NFR	3016	3534	
Yuan vs. Segal			
	HGHMM NUC	HGHMM NFR	Overall Mismatch
Segal NUC	5154	4457	0.5005
Segal NFR	911	205	

NUC: nucleosomal region; NFR: nucleosome-free region.

Table A2

Three-way Classification for Methods at 131 bp level.

		HGHMM NUC	HGHMM NFR
Yuan NUC	Segal NUC	2586	1087
	Segal NFR	463	41
Yuan NFR	Segal NUC	2568	3307
	Segal NFR	448	164
Overall NFR Match		0.0153	
Overall NUC Match		0.2411	
Overall Mismatch		0.7436	

References

Bussemaker HJ, Foat BC, Ward LD. Predictive modeling of genome-wide mRNA expression: from modules to molecules. Annu Rev Biophys Biomol Struct. 2007; 36:329–347. [PubMed: 17311525]

Casolari J, Brown C, Drubin D, Rando O, Silver P. Developmentally induced changes in transcriptional program alter spatial organization across chromosomes. Genes Dev. 2005; 19:1188–1198. [PubMed: 15905407]

Dion M, Altschuler S, Wu L, Rando O. Genomic characterization reveals a simple histone h4 acetylation code. Proc. Natl. Acad. Sci. USA. 2005; 102:5501–5506. [PubMed: 15795371]

Ercan S, Lieb JD. New evidence that DNA encodes its packaging. Nat Genet. 2006; 38:1104–1105. comment. [PubMed: 17006463]

Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 1990; 85:398–409.

Giresi PG, Gupta M, Lieb JD. Regulation of nucleosome stability as a mediator of chromatin function. Curr. Opin. Genet. Dev. 2006:16. in press.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007; 17:877–885. [PubMed: 17179217]

Greenbaum JA, Pang B, Tullius TD. Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res. 2007; 17:947–953. [PubMed: 17568010]

Gupta M. Generalized hierarchical Markov models for the discovery of length-constrained sequence features from genome tiling arrays. Biometrics. 2007; 63:797–805. [PubMed: 17825011]

Gupta M, Ibrahim JG. Variable selection in regression mixture modeling for the discovery of gene regulatory networks. J. Am. Stat. Assoc. 2007; 102:867–880.

Gupta M, Liu JS. De-novo cis-regulatory module elicitation for eukaryotic genomes. Proc. Nat. Acad. Sci. USA. 2005; 102:7079–7084. [PubMed: 15883375]

Hershberg R, Yeger-Lotem E, Margalit H. Chromosomal organization is shaped by the transcription regulatory network. Trends Genet. 2005; 21:138–142. [PubMed: 15734572]

Hogan GJ, Lee C-K, Lieb JD. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. PLoS Genet. 2006; 2:e158. [PubMed: 17002501]

Ioshikhes I, Hosid S, Pugh F. Variety of genomic DNA patterns for nucleosome positioning. Genome Res. 2011

Ioshikhes IP, Albert I, Zanton SJ, Pugh BF. Nucleosome positions predicted through comparative genomics. Nat Genet. 2006; 38:1210–1215. [PubMed: 16964265]

Juang B-H, Rabiner LR. Hidden Markov models for speech recognition. Technometrics. 1991; 33:251–272.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet. 2007; 39:1235–1244. [PubMed: 17873876]

Luger K. Dynamic nucleosomes. Chromosome Res. 2006; 14:5–16. [PubMed: 16506092]

Narlikar L, Gordan R, Hartemink EJ. A.: Nucleosome occupancy information improves de novo motif discovery. RECOMB 2007. LNCS (LNBI, Springer. 2007:107–121.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang J-PZ, Widom J. A genomic code for nucleosome positioning. Nature. 2006; 442:772–778. [PubMed: 16862119]

Tanner M, Wong WH. The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. 1987; 82:528–550.

Thastrom A, Bingham LM, Widom J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J Mol Biol. 2004; 338:695–709. [PubMed: 15099738]

Trifonov EN. Thirty years of multiple sequence codes. Genomics Proteomics Bioinformatics. 2011; 9:1–6. [PubMed: 21641556]

Wallrath LL, Lu Q, Granok H, Elgin SC. Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. Bioessays. 1994; 16:165–170. [PubMed: 8166669]

Wang J-PZ, Widom J. Improved alignment of nucleosome DNA sequences using a mixture model. Nucleic Acids Res. 2005; 33:6743–6755. evaluation Studies. [PubMed: 16339114]

Yuan GC, Liu JS. Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput. Biol. 2008; 4:e13. [PubMed: 18225943]

Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. Genome-scale identification of nucleosome positions in S. cerevisiae. Science. 2005; 309:626–630. [PubMed: 15961632]

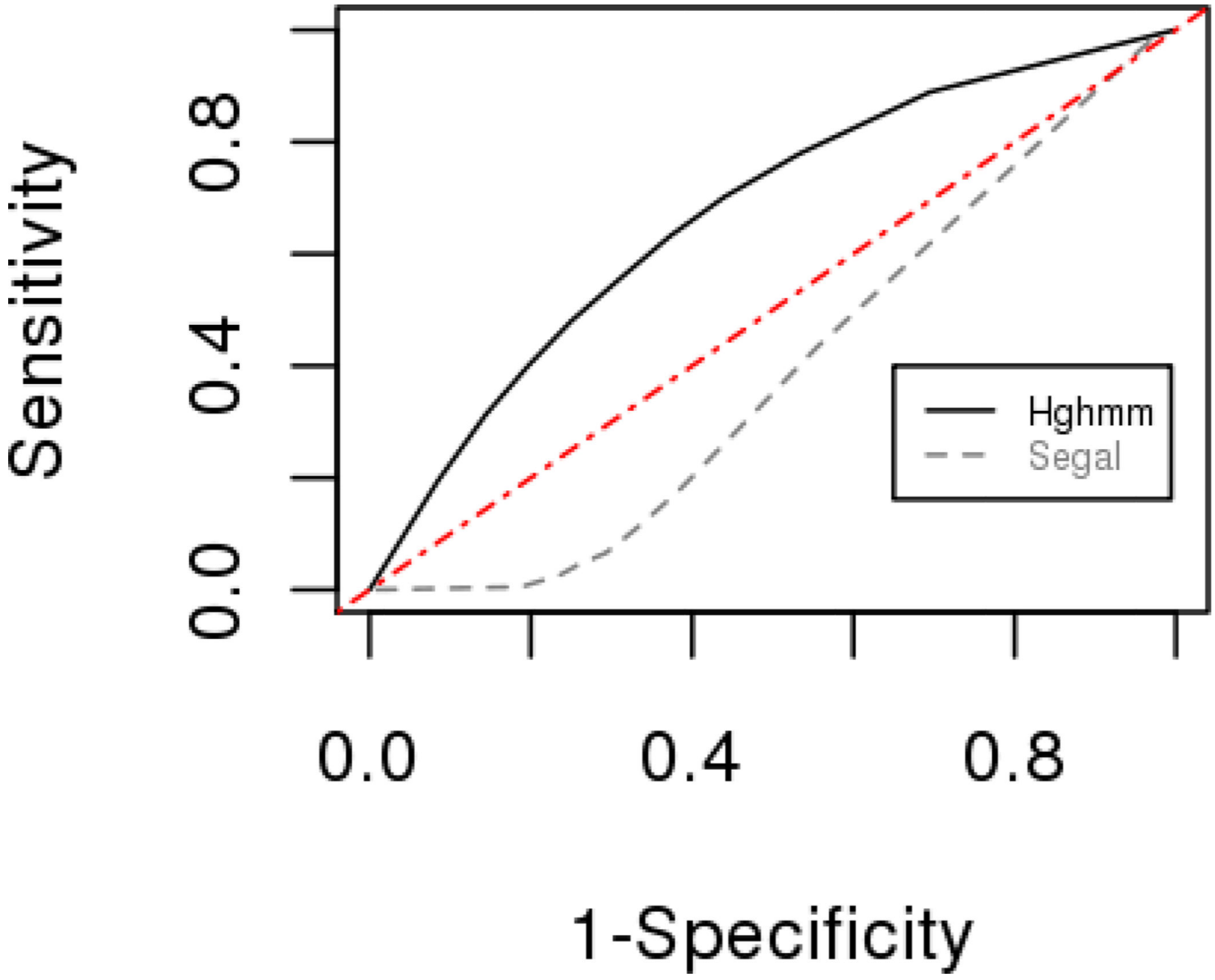

Figure 1.
Receiver Operator Curve for HGHMM Logistic model and Segal model at unit probe level.

Table 1

Measures of performance across test sets with 0.5 cut off.

	SENS	SPEC	FP	FN	NPP
Non-HGHMM Logistic	0.2258	0.8856	0.1144	0.7742	0.8444
HGHMM Logistic	0.5436	0.6851	0.3149	0.4564	0.6001
HGHMM Normal	0.5330	0.6813	0.3187	0.4670	0.6015

Column headers: SENS=Sensitivity, SPEC=Specificity, FP=False Positive, FN=False Negative, NPP=Nucleosome Prediction Percentage.

Table 2

State classification table.

		True NUC (%)	True NFR (%)	Nucleosomes (%)
A	Segal	70.7	6.6	79.1
B	HGHMM	69.1	55.4	60
C	Yuan	45.2	69.6	38.9
D	Segal	84.2	3.0	89.6
E	HGHMM	64.1	53.8	56.5

A: Segal compared to true states at probe level with 0.5 cut off; B: HGHMM logistic compared to true states at probe level with 0.5 cut off; C: Yuan compared to true states at 131 bp level; D: Segal compared to true states at 131 bp level; E: HGHMM compared to true states at 131 bp level. “NUC”: nucleosome; “NFR”: nucleosome-free region.

Table 3

Measures of performance for methods at the 131 bp level with 0.5 cut off.

	SENS	SPEC	FP	FN	NPP
Yuan	0.6965	0.4517	0.5483	0.3035	0.389
Segal	0.0302	0.8424	0.1576	0.9698	0.896
HGHMM	0.5381	0.6405	0.3595	0.4619	0.565

Column Headers: SENS=Sensitivity, SPEC=Specificity, FP=False Positive, FN=False Negative, NPP=Nucleosome Prediction Percentage.

Table 4

K-mer compositions of the top 10 principal components with factor loadings.

PC 1		PC 2		PC 3		PC 4		PC 5		PC 6		PC 7		PC 8		PC 9		PC 10	
	wt		wt		wt		wt		wt		wt		wt		wt		wt		wt
a	0.5270	g	−0.3754	c	−0.5251	ta	0.3942	tt	0.3257	tc	0.3741	ta	−0.3812	gt	0.3212	ac	−0.3368	cc	−0.3889
t	−0.5079	t	0.3745	g	0.5120	at	0.3881	ttt	0.2999	ga	0.3531	at	0.3384	ga	−0.3091	gt	−0.2394	gc	0.3821
aa	0.3409	a	0.3479	ca	−0.2021	aa	−0.3526	ct	−0.2669	ag	0.2548	ca	0.2776	ca	0.2703	cc	0.2358	gt	−0.2847
tt	−0.3230	c	−0.3471	tg	0.1921	tt	−0.3231	aaa	−0.2638	ac	−0.2065	ct	−0.2520	tc	0.2583	tg	−0.2282	ct	0.2563
aaa	0.1759	tt	0.2767	cc	−0.1868	aaa	−0.2735	aa	−0.2628	ct	0.1983	tg	0.2459	tg	−0.2535	gc	0.2182	ca	0.2227
ttt	−0.1628	aa	0.2635	gg	0.1821	ttt	−0.2368	ca	0.2232	aga	0.1929	ag	−0.2442	ag	0.2516	at	0.2181	gca	0.1895
tc	−0.1239	ta	0.1854	tc	−0.1706	ata	0.2233	tg	−0.2053	gt	−0.1915	tc	0.1568	ct	−0.2134	gg	0.2145	ga	−0.1698
ct	−0.1197	at	0.1835	ac	−0.1631	tat	0.2192	tc	−0.2000	tct	0.1877	tca	0.1530	tga	−0.1923	aca	−0.1750	gct	0.1609
ga	0.1179	ttt	0.1614	ga	0.1614	aaaa	−0.1665	ag	0.1990	tg	−0.1800	cat	0.1461	ac	−0.1886	ca	−0.1733	cg	−0.1564
ag	0.1130	aaa	0.1610	gt	0.1558	tttt	−0.1352	tttt	0.1854	ttc	0.1339	aat	0.1437	tca	0.1786	tgt	−0.1579	tgc	0.1434
Z	0.688		3.271		6.120		1.428		0.719		−0.903		0.715		−0.106		−1.128		−0.942
p	0.491		0.001		9.3e–10		0.153		0.472		0.367		0.475		0.916		0.259		0.346

The top 10 k-mers are shown, with their weights (factor loadings) in decreasing order, standardized regression coefficients (Z) and corresponding p-values (p). Negative signs on the regression coefficient (with positive factor loadings) indicate an inverse relationship with nucleosome positions (positively related to NFRs), and vice-versa.