Re-Cracking the Nucleosome Positioning Code

Mark R. Segal

Department of Epidemiology and Biostatistics,
Center for Bioinformatics and Molecular Biostatistics,
University of California, San Francisco, CA 94143-0560, USA.
(e-mail: mark@biostat.ucsf.edu).

Abstract

Nucleosomes, the fundamental repeating subunits of all eukaryotic chromatin, are responsible for packaging DNA into chromosomes inside the cell nucleus and controlling gene expression. While it has been well established that nucleosomes exhibit higher affinity for select DNA sequences, until recently it was unclear whether such preferences exerted a significant, genome-wide effect on nucleosome positioning *in vivo*. This question was seemingly and recently resolved in the affirmative: a wide-ranging series of experimental and computational analyses provided extensive evidence that the instructions for wrapping DNA around nucleosomes are contained in the DNA itself. This subsequently labelled *second genetic code* was based on data-driven, structural, and biophysical considerations. It was subjected to an extensive suite of validation procedures, with one conclusion being that intrinsic, genome-encoded, nucleosome organization explains ≈50% of *in vivo* nucleosome positioning. Here, we revisit both the nature of the underlying sequence preferences, and the performance of the proposed code. A series of new analyses, employing spectral envelope (Fourier transform) methods for assessing key sequence periodicities, classification techniques for evaluating predictive performance, and discriminatory motif finding methods for devising alternate models, are applied. The findings from the respective analyses indicate that signature dinucleotide periodicities are absent from the bulk of the high affinity nucleosome-bound sequences, and that the predictive performance of the code is modest. We conclude that further exploration of the role of sequence-based preferences in genome-wide nucleosome positioning is warranted. This work offers a methodologic counterpart to a recent, high resolution determination of nucleosome positioning that also questions the accuracy of the proposed code and, further, provides illustration of techniques useful in assessing sequence periodicity and predictive performance.

1 Introduction

In a recent, widely-celebrated article, Segal *et al.*, [1] performed a diverse series of computational and experimental analyses that purportedly established the existence of a second genetic code. This code pertains to nucleosome positioning along the genome, and asserts that such positioning is determined by the DNA itself. Derivation of the code was based on data-driven, structural, and biophysical considerations. An extensive suite of validation procedures was applied and, based on the attendant results, it was concluded that genomes encode an intrinsic nucleosome organization which explains ≈50% of nucleosome positions *in vivo*. Here, we re-evaluate features and performance of this nucleosome positioning code. Initially, we scrutinize one of the observations showcased in the context of high affinity nucleosome-bound sequence; namely, that such sequences exhibit strong periodicities of select dinucleotides. Subsequently, we revisit some of the approaches used to assess performance of the code. Comparisons are drawn with a competing model derived via motif finding methods.

We note at the outset that there are some important interpretational nuances surrounding relevant aspects of the presentation and methodology of Segal *et al.*, [1]. First is the matter of whether a "code" comprising a rigid set of sequence-based rules for the (deterministic) placement of nucleosomes is being advanced or, alternatively, a "model" is being proposed that detects a signal related to nucleosome positioning. While these are distinct propositions, for our present purposes we do not need to arbitrate between them. Our focus is on the manner whereby the code/model is evaluated, and the key role for discriminatory analyses in this context. This viewpoint also served to frame the analysis conducted by Yuan and Liu [2], described further in the Discussion.

Second is the question of whether the observed select dinucleotide periodicities provide motivation for, or validation of, the proposed code. Again, we remain agnostic with respect to this distinction. As described below, it is clear that dinucleotide frequencies of the obtained high affinity nucleosome-bound sequences are fundamental to the code. That select dinucleotides exhibit apparently strong ≈10bp periodicities for these sequence sets is prominently highlighted, with substantive interpretation as to both the particular dinucleotides displaying such periodicity and the cycle length. Our concern here is in gauging the extent to which (a set of) sequences exhibit periodic behavior, and providing and illustrating techniques that enable such assessment.

It is additionally important to establish what we are not attempting here. We are not seeking to establish a new code, nor do we provide new data. In the Discussion, we comment on some very recent work bearing on these aspects. Rather, our focus is on methodologic issues arising in, but not limited to, the work of Segal et al., [1].

2 Results

2.1 Select dinucleotide periodicities for high affinity nucleosome-bound sequences

The starting point for deriving the nucleosome positioning code (NPC) are sets of yeast DNA sequences that are stably nucleosome bound, these being isolated by an accurate, genome-wide assay. While the assay yielded 518 sequences only the 199 (38%) with lengths within ± 5 basepairs (bp) of the canonical 147bp nucleosome were employed for model development. These 199 sequences were then center aligned and augmented with (i) their reverse complements, reflecting the 2-fold symmetry of the nucleosome structure [3], and (ii) ± 1bp offset sequences, predicated on the notion that such small changes in spacing of key nucleosome DNA sequence motifs incur little free energy cost with respect to histone - DNA interactions. So armed with this aligned set of sequences, probability scores based simply on aggregating position-specific, dinucleotide proportions are computed. These scores provide the foundation for the proposed code. As indicated, the observation that, for select dinucleotides, these proportions exhibit striking periodicities receives considerable attention. In particular, in-phase ≈10bp cycles (coincident with the DNA helical repeat) are evident for AA/TT/TA dinucleotides (see Figure 1), these being out of phase with a ≈10bp GC cycle. Not only are arguments advanced as to the importance of these specific dinucleotides with respect to *bendability* (essential for the sharp bending required for wrapping [4; 5]) and *phase anisotropy* (facilitating necessary positive and negative basepair roll), but also the same periodicities and phase behaviors are exhibited when a set of *in vivo* nucleosome sequences from chicken and three *in vitro* experiments are similarly scored.

Let $S = (S_1, S_2, \ldots, S_{147})$ denote an arbitrary 147bp sequence. Representing the above position-specific, dinucleotide probability as $P_i(S_i|S_{i-1}), i = 2, \ldots, 147$, the nucleosome-DNA model score for S is computed according to a first-order Markov model as

$$P(S) = P_1(S_1) \prod_{i=2}^{147} P_i(S_i|S_{i-1}). \tag{1}$$

A *legal configuration* specifies a *set* of 147bp nucleosomes, scored according to (1), and a start position for each, such that no two nucleosomes overlap and the minimum distance between them is 10bp. The probability of every configuration is then given by the Boltzmann distribution. While the number of configurations is vast, a forward-backward dynamic programming method enables efficient estimation of the probability of placing a nucleosome that starts at each basepair in the genome. This probability (and derivations therefrom) constitutes the second genetic code. Further details can be found in [1] and the supplementary material thereof.

So, given that neighboring nucleosomes are separated from one another by stretches – of length 10 to 50bp – of unwrapped linker DNA, overall some 75% - 90% of genomic DNA is nucleosome bound; a recent estimate for yeast based on a high resolution, tiling array approach being 81% [6]. Accordingly, in view of the genesis of the code, we would anticipate seeing numerous instances of ≈10bp periodicities throughout the genome. Informal (purely visual) assessments of genome-wide

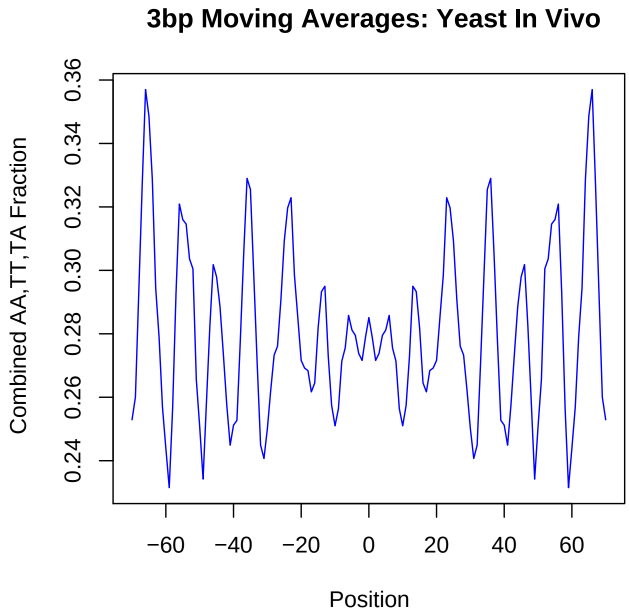

Figure 1: Position specific fractions of combined AA, TT, and TA dinucleotides aggregating over the 199 nucleosome bound sequences.

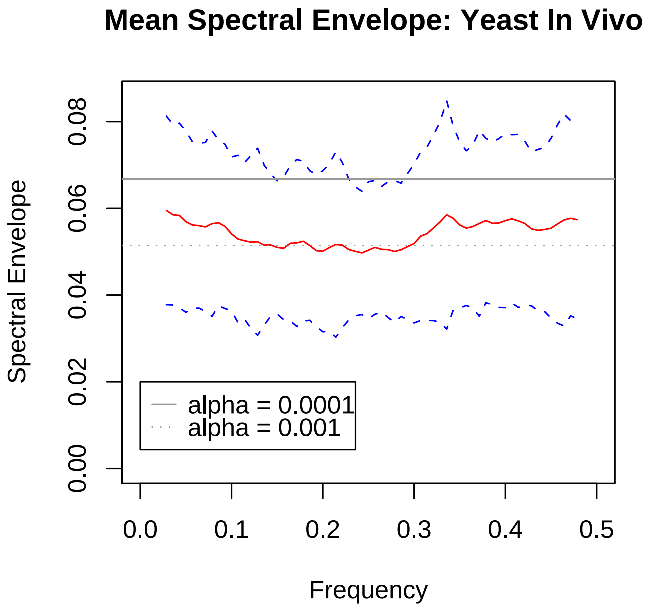

Figure 2: Average (over individual sequence) spectral envelope (red) with standard errors (blue) and analytical (see Methods) critical values (grey).

dinucleotide periodicities have been performed, revealing ≈8bp cycles that have been linked to indications (based on molecular models of a DNA - Dnmt3a (DNA methyltransferase 3a) dimer) that Dnmt3a methylates DNA in a periodic fashion [7; 8]. Formal techniques have also been applied. Notably, the (Fourier-based) *spectral envelope*, built on work of Stoffer and colleagues [9; 10], and briefly outlined in Methods, has been employed. Accordingly, the following quote from Rosen and Stoffer [11], is of interest (emphasis added):

> The spectral envelope picks up a signal at one cycle every three bp, which occurs often in coding sequences we have analyzed. There is another peak in the spectral envelope indicating a signal at one cycle every 10bp. This signal is particularly interesting because, while the double helix makes one turn about every 10 base-pairs, the 10bp signal is *rarely seen* and the importance of this twisting is not clear.

We now take up the task of reconciling this apparent discrepancy: on the one hand, according to the basis of the nucleosome positioning code, 10bp periodicities ought be common while, on the other hand as asserted in [11], they are seldom seen. The contention that an attribute is “rarely seen” is clearly dependent on where and how the search was conducted. As it is problematic to address the “where”, our focus here is on the “how” – the spectral envelope methodology used for eliciting periodicities. In order to assess whether the spectral envelope is a sufficiently refined technique for identifying periodicities in short sequences we apply it to the five nucleosome sequence sets (yeast, chicken *in vivo*; yeast, mouse, synthetic *in vitro*) where, according to Figure 1 and [1] and references therein, such signals are manifest.

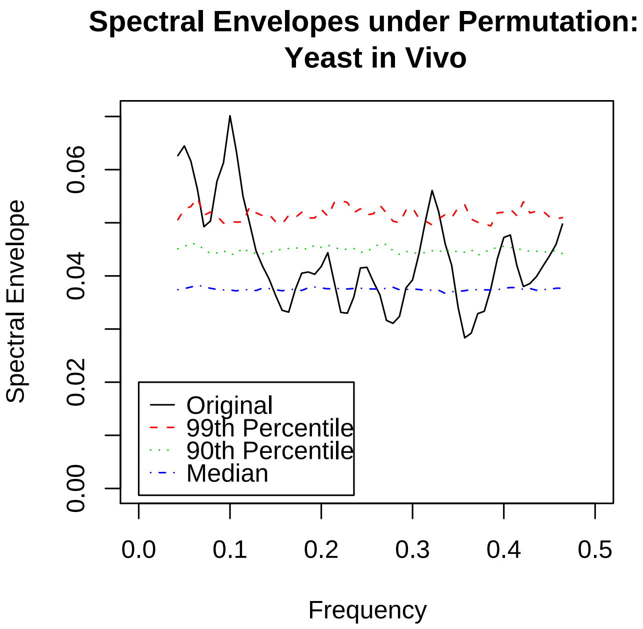

Figure 3: Spectral envelope (black) obtained by aggregating sequences with critical values obtained by permutation.

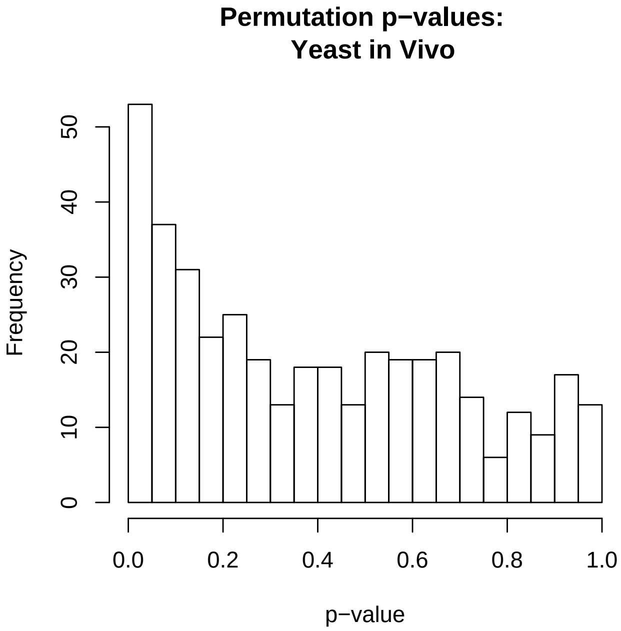

Figure 4: Distribution of individual permutation p-values for the maximal spectral peak in a neighborhood of $\omega = 0.1$ over the 398 (= 199 × 2) nucleosome bound sequences.

Focussing initially on the yeast *in vivo* sequences, we compute the spectral envelope for each sequence and display the resultant *mean* envelope (averaged over the ensemble), along with analytically defined critical values (see Methods), in Figure 2. Contrary to expectation, there is no evidence of 10bp periodicity, corresponding to $\omega = 0.1$. [We note that here, and subsequently, we have been flexible in making evaluations in a neighborhood of $\omega = 0.1$.] However, this process of averaging individual envelopes does not reflect the manner whereby the striking periodicities of Figure 1 were exposed. To produce a corresponding spectral envelope requires averaging the individual sequences (rather than averaging spectral envelopes), this conforming to the computation of select dinucleotide *fractions* as displayed. Doing so produces the spectral envelope presented in Figure 3 where now critical values are (preferably – see Methods) obtained via permutation. Now the 10bp cycle is clearly detected, there being a significant peak at $\omega = 0.1$. Nonetheless, the failure of the former process of averaging individual envelopes begs the question as to the extensiveness of the characterizing periodicity in the (individual) select, nucleosome bound sequences. Figure 4 shows that while more than 50 sequences(out of $398 = 199 \times 2$, reverse complement being included, albeit duplicative, as a concordance check) have a significant ($p < 0.05$) peak at $\omega = 0.1$, the bulk of the sequences do not exhibit this signature. Furthermore, the same applies to the other systems studied.

These findings call into the question the various interpretations as to the consequence of the select dinucleotide periodicities. Whether invoked from either a motivating or validating standpoint, the absence of the prescribed periodicities from $> 80\%$ of the high affinity nucleosome-bound sequences suggests that they are at best a weak component of an *in vivo* preference signature. However, it is important to note that these periodicities were not directly incorporated into the code as defined by the spatial probabilities (1) and subsequent steric hindrance models. So, their paucity does not impact the validity of the code. Segal *et al.*, [1] proffer a mound of evidence in order to demonstrate legitimacy of the positioning code. We next scrutinize what we believe to be the critical components thereof.

2.2 Predictive performance of the positioning code

Many figures analogous to Figure 2 of Segal *et al.*, [1] are presented in their companion supplementary information. Using genomic coordinates these figures display tracks corresponding to stably positioned nucleosomes according to the literature, and stably positioned nucleosomes as predicted by their code. Showcasing the seemingly close alignment between these tracks is presumably intended to contribute to affirming the validity of the code. However, while numerous, these figures are but qualitative snapshots, and so do not provide quantification of the code's genome-wide accuracy. More importantly, they do not afford a *discriminatory* perspective which, we believe, is essential for any hypothesized prediction scheme such as the proposed code. To illustrate: as was cited above, [6], 81% of yeast genomic DNA is nucleosome bound. So, a trivial rule that declared *every* basepair to be nucleosome occupied would enjoy 81% accuracy, greatly exceeding the $\approx 50\%$ claimed for Segal *et al's* code. Where the "highly non-informative and useless" [12] trivial rule breaks down, of course, is in predicting nucleosome *free* or *depleted* genomic regions. A few relevant data sets are presented in [1] (Supplementary Figures 23-25), two of which are recreated here in Figures 5 and 6. [We do not consider the data of Supplementary Figure 23 since the depleted set is sparse and it is the sole open reading frame example, precluding validation; see below.] Indeed, while the analysis of these data sets contributes to the overall $\approx 50\%$ accuracy estimate, they are not accompanied by formal or direct assessments of misclassification performance. Next, we review the methods used to analyze this pivotal data, as well as applying alternate approaches that provide such formalism.

Consider Figure 5. Depicted are two empirical cumulative distribution functions (ecdf(s)) corresponding to two sets of sequences: yeast intergenic regions that are respectively nucleosome depleted (blue, 294 sequences) and occupied (red, 5387 sequences) where the partitioning into depleted / occupied categories conforms to that defined by Bernstein *et al.*, [13] who conducted the original Chip-ChIP experiments. For Figure 6, which utilizes data from Lee *et al.*, [14], the set of depleted regions was obtained by stringent thresholding (30% under-enrichment). The ecdfs are for *average nucleosome occupancy probabilities* (ANOP(s)) which are defined as the average, across all basepairs in a designated sequence, of the probability that a basepair is covered by any nucleosome.

The latter quantity is obtained by summing the probabilities of placing a nucleosome at a given basepair over the preceding 146bp which, in turn, are derived from (1) and the forward-backward dynamic model. As such, these quantities *embody* the nucleosome positioning code.

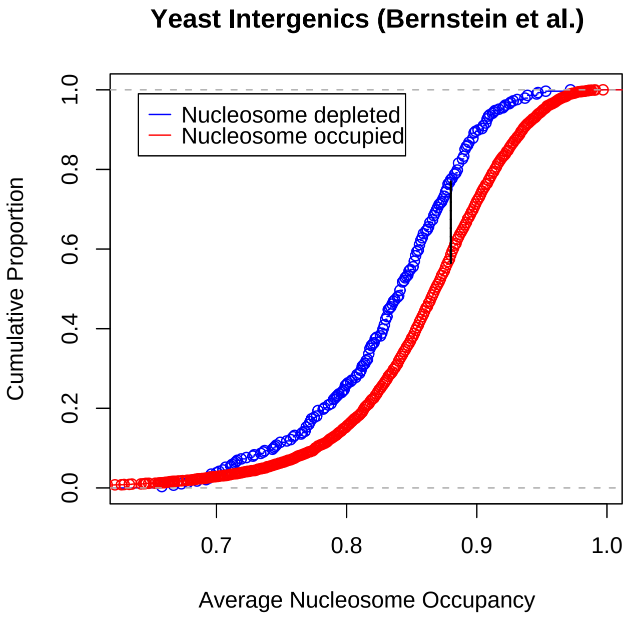

Figure 5: Empirical distribution functions of ANOPs by class: Bernstein *et al.*, data (Supplementary Figure 24). The vertical black bar designates the Kolmogorov-Smirnov test statistic; see text.

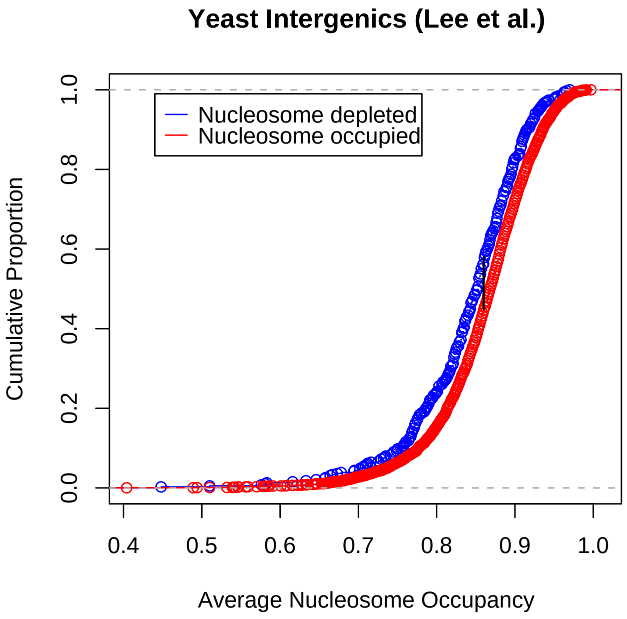

Figure 6: As for Figure 5: Lee *et al.*, data (Supplementary Figure 25).

As depicted in Figures 5 and 6, differences between the depleted and occupied sequences are assessed via the Kolmogorov-Smirnov (KS) test, which attains seemingly impressive p-values: 10^{-9} and 10^{-5} respectively, although we are dealing with large sample sizes. Now, the KS test is an established means for assessing differences between cumulative distribution functions. But, it does not address predictive (classification) performance of the code. Of course, many such measures are available, including misclassification rate, sensitivity, specificity, and positive and negative predictive value. It is straightforward to show (see Methods) that the optimization implicit in the KS test corresponds to maximizing the *sum* of sensitivity plus specificity. While this is not a bad criteria, it is not necessarily a good one, and can mislead especially in imbalanced (class size) settings as here. A more established means of assessing a classifier's performance is via cross-validated (CV) receiver operator characteristic (ROC) curves (sensitivity vs 1- specificity) and the attendant area-under-the-curve (AUC).

Results from such an approach are presented in Figures 7 and 8. Here we are using ANOPs to discriminate between the two classes. The classifier employed was gradient boosting [15; 16], overkill for this single feature setting, but used for comparability with subsequent multi (two) feature inputs. While results with a single feature are, by definition, invariant to classifier choice, similar invariance was observed with two features. An ROC curve coincident with the diagonal, with associated AUC = 0.5, corresponds to random classification (e.g., based on a coin toss) and indicates that the feature(s) have no discriminatory power. That the cross-validated ROC curves in Figure 8 are, in fact, virtually coincident with the diagonal, with average AUC = 0.54, shows that here ANOPs have little predictive content. Results for Figure 7 are marginally better, with an average AUC, over the 5 CV folds, of 0.60.

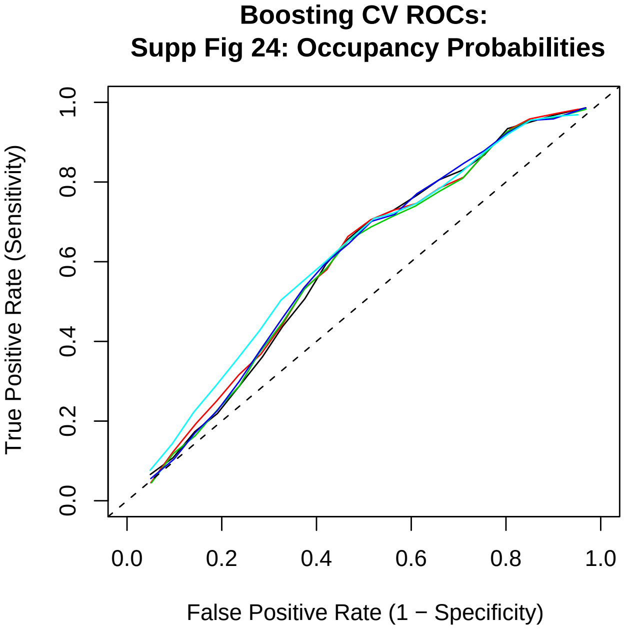

Figure 7: ROC curves: Bernstein *et al.*, data. The five individual traces correspond to the each of the five cross-validation folds.

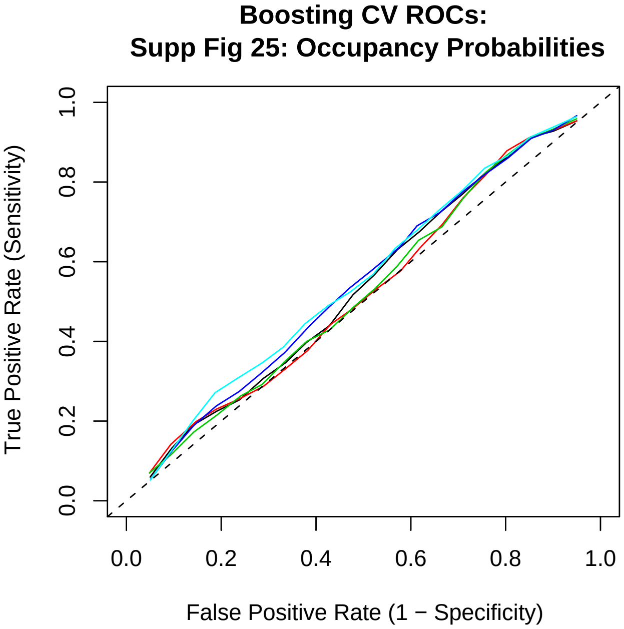

Figure 8: As for Figure 7: Lee *et al.*, data.

However, before accepting these results, and accordingly dismissing the genome-wide discriminatory content of the proposed nucleosome code, we thought it prudent to confirm the input ANOPs. These had ben derived using the script provided by Segal *et al.*, see `http://genie.weizmann.ac.il/pubs/nucleosomes06/segal06_exe.html` As a simple check we re-estimated the respective ecdfs for each of the occupancy / depleted datasets. The results, as presented in Figures 9 and 10, reveal dramatic differences with the ecdfs obtained using the original ANOPs, given in Figures 7 and 8. Moreover, the nature of these differences – reduced separation between classes – would clearly serve to attenuate discriminatory signal. The reason for the differences is handling of boundary effects. These were naïvely ignored in our initial computation of ANOPs, despite a recommendation to employ at least 5000bp of flanking sequence around the target sequence of interest (`http://genie.weizmann.ac.il/pubs/nucleosomes06/segal06_prediction.html`). In actuality, the original ANOPs did not use flanking sequence but, rather, were derived from a whole genome solution (on a per chromosome basis) with mapping to the appropriate region. It has been reasonably asserted (Eran Segal, personal communication) that this is the appropriate course since (i) each region is part of the entire chromosome and neighboring sequences influence its nucleosome organization, and (ii) boundary effects that will otherwise dominate especially for short sequences. Nonetheless, in view of (a) the dramatic differences obtained from the global *vs* local approaches, and (b) the highly dynamic nature of nucleosome remodeling, the question of the *extent* to which distant sequence ought influence derived ANOPs warrants additional investigation. Further, irrespective of the manner in which boundary effects are handled, the averaging used in obtaining sequence- / region- level ANOP summaries can conceal disparate position level occupancy probabilities within a sequence (not shown).

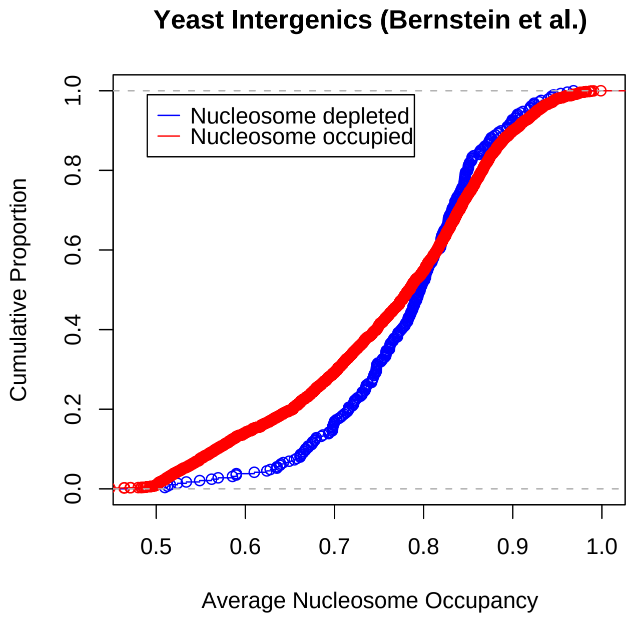

Figure 9: ANOP empirical cumulative distribution functions (ecdfs) without flanking data: Bernstein *et al.*, data; *cf* Figure 5.

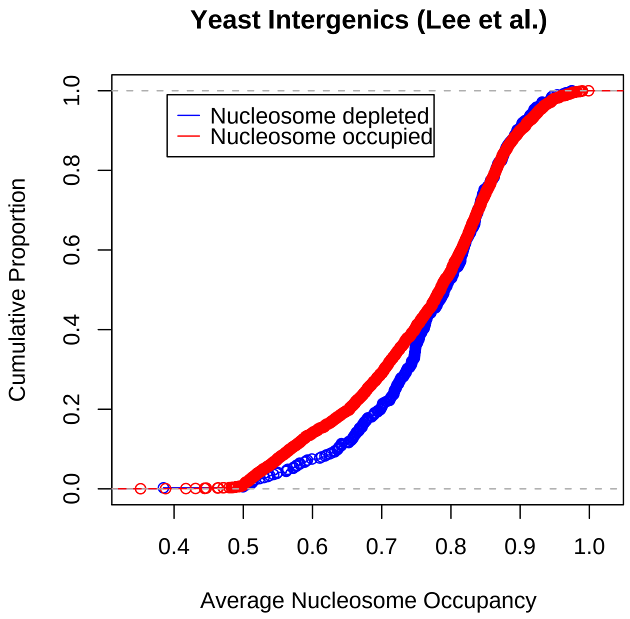

Figure 10: As for Figure 9: Lee *et al.*, data; *cf* Figure 6.

We revisited cross-validated classification {depleted, occupied} performance using the chromosome-level, boundary-corrected ANOPs [1], again employing boosting. Results are depicted in Figures 11 and 12. The improvements over uncorrected ANOPs are slight, with the new AUCs for the Bernstein *et al.*, and Lee *et al.*, datasets being 0.62 and 0.57 respectively. Still, modest though these AUCs are, perhaps they are sufficient to contend that the proposed nucleosome positioning code has useful predictive content. Since it is problematic to make such judgements solely on *absolute* AUCs, we next turn to alternate means for pursuing sequence-based classification.

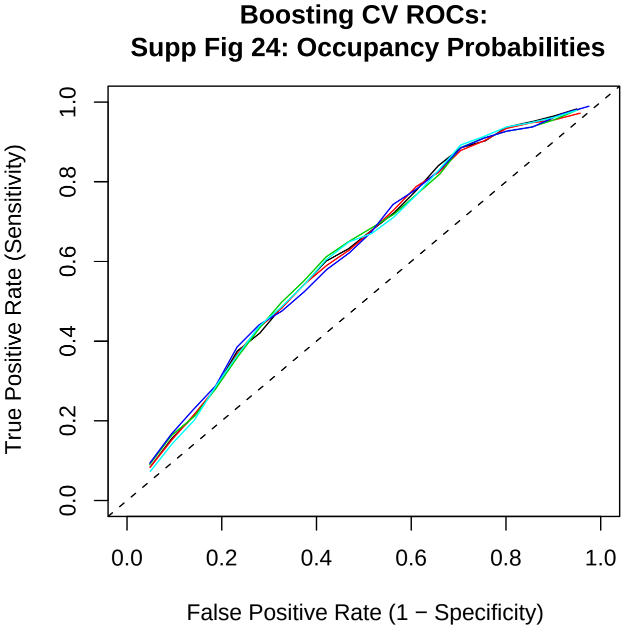

Figure 11: ROC curves using boundary corrected ANOPs: Bernstein *et al.*, data.

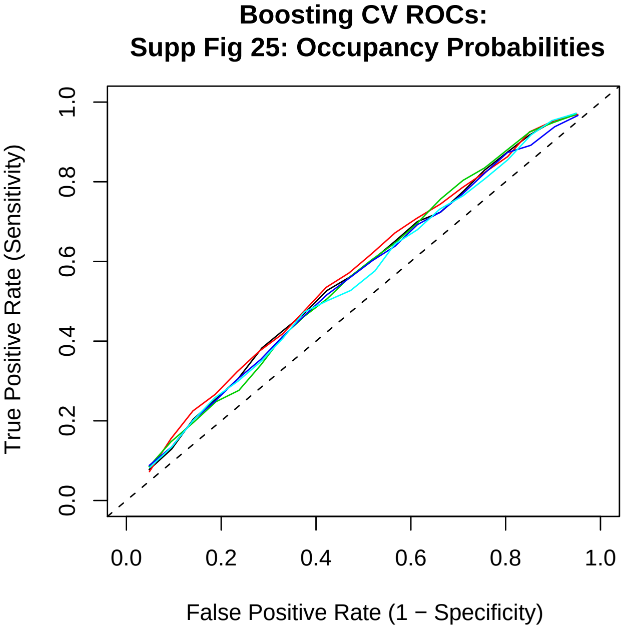

Figure 12: ROC curves using boundary corrected ANOPs: Lee *et al.*, data.

2.3 Discriminatory motif finding

Since the ability to distinguish nucleosome depleted sequences from nucleosome occupied sequences is fundamental to positioning code evaluation, we now turn to methods that tackle such objectives directly. Given a set of (provisionally) related sequences there is an extensive literature on finding patterns common to the set. Such patterns are called *motifs* and techniques for their elicitation are termed motif finding methods. Broadly, there are three classes of motif finding approaches: statistical (model-based) [17; 18; 19; 20], enumerative [21; 22], and dictionary-based [23; 24]. Additionally, within each of these families there are *generative* and *discriminative* approaches, as well as hybrids thereof. Given our classification emphasis, and the availability of both nucleosome depleted and occupied sequence sets, we focus on discriminative methods, as has been recently advocated [25; 26; 27]. Interestingly, this approach is also promoted by the lead author of the positioning code; see Segal et al., [28].

From an algorithmic standpoint the status of discriminatory enumerative [29] or hybrid dictionary - statistical [30] methods is considerably more advanced than pure statistical methods [31]. This is, in large part, a consequence of effective use of hash tables coupled to a hidden Markov model (HMM) for motif representation. Accordingly, we employ *Wordspy* (`http://cic.cs.wustl.edu/wordspy/`) which has been shown to outperform [30] a suite of competing motif finding methods on benchmark datasets [32].

We first applied Wordspy to the sequence sets of Supplementary Figure 24 (Bernstein *et al.*, [13]);

see Methods for specifications. Among the leading motifs extracted of length $\geq$ 10 nucleotides were poly(dA.dT) stretches and variants on the Rap1 binding motif (CACCCATACAT). These, or subsequences thereof, have previously been implicated in nucleosome positioning [33; 13; 6]. Next, we used the frequency of occurrence (counts per sequence / sequence length) of these two motifs as features in classifying the (nominally – see below) unseen data of Supplementary Figure 25 (Lee *et al.*, [14]). The results obtained, again using boosting, are presented in Figure 13. It is immediately apparent that this simple model, based on just two motifs, provides greater discriminatory power than the nucleosome positioning code, the respective AUCs being 0.76 (two motifs) and 0.57 (ANOPs, Figure 12).

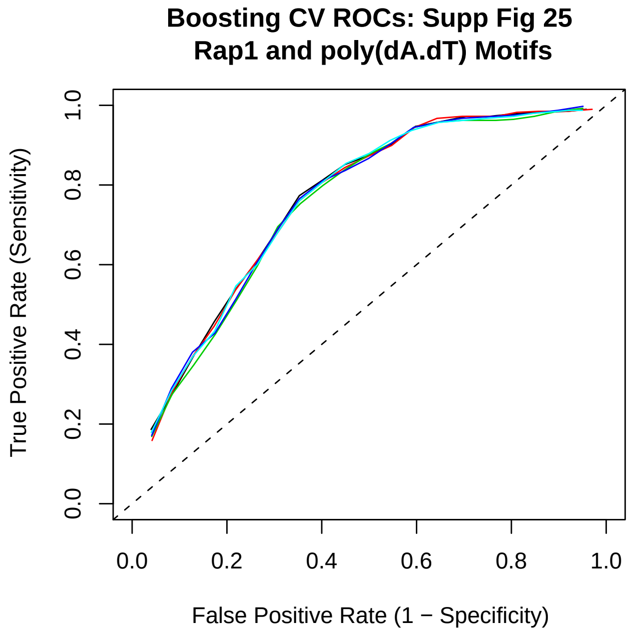

Figure 13: Discrimination between depleted and occupied sequences using poly(dA.dT) and Rap1 motifs; *cf* Figure 12.

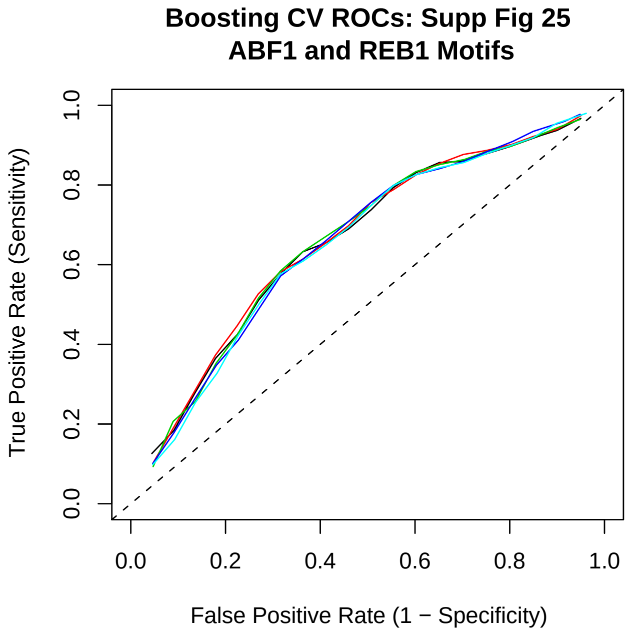

Figure 14: Discrimination between depleted and occupied sequences using ABF1 and REB1 motifs; *cf* Figure 12.

A few comments are in order. Firstly, while it is the case that there are significant differences in sequence *length* between the depleted and occupied sets, these are not driving the discriminatory performance depicted in Figure 13. Even when we do not normalize for sequence length, but just use raw motif counts as features, we obtain comparable classification performance.

Secondly, our strategy of using one dataset for eliciting motifs and the other for evaluating classifier accuracy, might suggest that the two sets are *independent*. However, this is far from the case. Since both constitute whole genome categorizations of nucleosome depletion / occupancy for all intergenic reasons, they differ (essentially) only in the manner whereby partitioning into these classes is effected. Thus, it may not surprise that motifs found using the Supplementary Figure 24 partition provide discriminatory power when applied to Supplementary Figure 25, and further incur concerns about data reuse and associated optimistic performance assessments [34]. It was in order to mitigate this concern that we utilized cross-validation based measures of classifier accuracy. However, the fact that the overlap between the depleted sets is meager – only one third of the depleted sequences in Supplementary Figure 24 are so identified in Supplementary Figure 25 – calls into question Segal *et al's* analysis and interpretation of these data. To paraphrase: if nucleosome depletion (occupancy) is such a moving target how can it be captured by a static code?

Thirdly, that the proposed positioning code was experimentally verified, presumably bolstered claims as to its legitimacy. Yet, experimental confirmation of the importance of the two motifs identified above also exists [33; 13]. For example, nucleosomes are depleted in the vicinity of Rap1 consensus sites and this depletion can be reversed by the small molecule rapamycin or by removing Rap1 binding sites. We are not advancing an alternate code based on these two motifs. Rather, we are questioning whether claims for a genome-wide code can be affirmed by select *in vitro* experimentation, which does not preclude alternate explanations, and especially when genome level accuracy is modest.

3 Discussion

The purpose of this paper was to revisit both attributes and performance of the recently proposed nucleosome positioning code (NPC) [1]. Using spectral envelope techniques, we first showed that the signature periodicities deemed to characterize sequences with high nucleosome affinity are absent from a sizable number of such sequences. Subsequently, using motif finding methods and classification techniques, we developed an alternative, experimentally justified, “code” which provided superior discriminatory performance to that of the original NPC.

A recent study by (William) Lee *et al.*, [6], of nucleosome positioning that achieved the highest basepair resolution to date through use of tiling arrays, also reached the conclusion that other structural features and motifs are more predictive than NPC occupancy probabilities. The Rap1 and poly(dA.dT) motifs figure very prominently in the lists of predictive DNA properties they elicited [6, Figures 5a, 6c]. Two other features, the ABF1 and REB1 transcription factor binding sites, also were forefront on these lists. To further assess their importance on unseen data we applied the same cross-validated, gradient boosting classification approach to the Bernstein *et al.*, data (Supplementary Figure 24) and (Cheol-Koo) Lee *et al.*, data (Supplementary Figure 25), using ABF1 and REB1 frequencies as inputs. Although AUCs were smaller (0.72 and 0.66 respectively) than those attained with Rap1 and poly(dA.dT) (0.82, 0.76) they were still significantly greater those obtained using the NPC (0.62 and 0.57). Figure 14 displays the cross-validated ROC curves for the (Cheol-Koo) Lee *et al.*, data. Thus, these motifs exhibit both consistency across datasets and superior predictive performance to the NPC.

The high resolution approach of [6] enables the authors to declare, on purely empiric grounds, that the 199 high nucleosome affinity sequences upon which Segal *et al's* code is based have a “nearly random distribution of occupancy ratios and do not correspond to well-positioned nucleosomes.” They go on to speculate that global, genome-wide positioning is governed by exclusion signals, whereas local positioning is influenced by select periodicities. This conforms with our view that genome scale positioning needs to be tackled as a discrimination / prediction problem. Yuan and Liu [2] pursued this approach by complementing the 199 nucleosome-bound yeast sequences with 296 nucleosome depleted linker sequences. They used stepwise logistic regression to perform classification based on wavelet-derived features. Not only did they achieve appreciably greater accuracy than that obtained using the NPC of Segal *et al.*, but they demonstrate how the by adoption of a discriminatory approach, with negative instances, can improve the NPC. It is our view that a wide range of genomics problems can benefit from such discriminatory approaches, as opposed to positive-sequence-only generative methods.

Methods

Spectral envelope

For real-valued data use of Fourier methods and spectral analysis is fundamental to assessing periodicity. The spectral envelope construct was developed for frequency domain analysis of *categorical* time series by the *scaling* thereof [9; 10]. It has obvious relevance to detecting periodicities in DNA or protein [35] sequence.

The spectral density, or *periodogram*, of a sampled, mean-centered, real-valued time series, $X_t, t = 1, \dots, n$ at frequency ω is

$$I_n(\omega) = \left| n^{-1/2} \sum_{t=1}^{n} X_t \exp(-2\pi i \omega t) \right|^2 = \left| n^{-1} \sum_{t=1}^{n} X_t \cos(-2\pi \omega t) \right|^2 + \left| n^{-1} \sum_{t=1}^{n} X_t \sin(-2\pi \omega t) \right|^2 \quad (2)$$

The spectral density, $f(\omega)$, of a stationary time series is the limit ($n \to \infty$) of $E[I_n(\omega)]$. We have $\text{var}(X_t) = 2 \int_0^{1/2} f(\omega) d\omega = \sigma^2$. These constructs generalize to a k-dimensional multivariate time-series, $\mathbf{Y_t}$, with now the spectral density $f_Y(\omega)$ being a $k \times k$ complex-valued Hermitian matrix.

Consider a categorical time series, $X_t, t = 0, \pm 1, \pm 2, \dots$ with finite state space $C = \{c_1, c_2, \dots, c_k\}$. The c_j's designate (all) possible categories: $X_t = c_j$ when the time series is in state c_j at time t. We recast the one-dimensional categorical time series as a multivariate k-dimensional time series, $\mathbf{Y_t}$, via unit vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$ such that $\mathbf{Y_t} = \mathbf{e}_j$ when $X_t = c_j$. The scaling process uses a scaling vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ to convert the multivariate time series to a univariate, real valued series, $X_t(\boldsymbol{\beta})$, by assigning category c_j numeric value β_j. Thus, $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}' \mathbf{Y_t}$. Then we have $f_Y(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}' f_Y^{re}(\omega) \boldsymbol{\beta}$, where $f_Y^{re}(\omega)$ denotes the real part of $f_Y(\omega)$. The key question is then how to select the scalings $\boldsymbol{\beta}$. Stoffer et al., [10] advocate choosing $\boldsymbol{\beta}$ to maximize the power (variance) at each frequency $\omega \in (-1/2, 1/2]$ relative to the total power $\sigma^2 = \sigma^2(\boldsymbol{\beta})$. This corresponds to the optimization criterion

$$\lambda(\omega) = \sup_{\boldsymbol{\beta}} \left\{ \frac{\boldsymbol{\beta}' f_Y^{re}(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\} \quad (3)$$

where V is the covariance matrix of $\mathbf{Y_t}$. The value $\lambda(\omega)$ is called the *spectral envelope* since it envelopes the (standardized) spectrum for any scaling. Importantly, the spectral envelope is readily computed via the discrete Fourier transform [10]. Application to sequence data comes via mapping the relevant categories (e.g., 4 letter nucleotide or 20 letter amino acid alphabets) to a multivariate time series. In our applications here, following [1], we use the *di*-nucleotide alphabet.

Assessment of peak significance is a central concern. In Figure 2 we have used the analytic approximations advanced by Stoffer *et al.*, [10]: the α-level critical value for the estimated spectral envelope $\hat{\lambda}(\omega)$ is here $(2/(n-1))\exp(z_\alpha/\nu_n)$, where n is the sequence length, z_α is the $1-\alpha$ critical value of the standard normal distribution, and ν_n is a function of weights depending on the smoothing employed in estimating the underlying spectral density $f_Y(\omega)$. However, these critical values are accompanied by the somewhat ambiguous advice [10]: "From our experience thresholding at very small values of α relative to sample size works well." Accordingly, in Figure 3, thresholds are obtained by permutation.

Kolmogorov-Smirnov statistic reformulation

Let $r_i(x); 0 \leq r_i(x) \leq 1$ be the score (here ANOP) for case (here sequence) x in group i; $i = 1\,(\text{depeleted}), 2\,(\text{occupied})$. Let f_i be the cumulative distribution function (cdf) for group i: $f_i(t) = \int_0^t r_i(u)du$. The associated empirical cdf (ecdf) is $\hat{f}_i(t) = \sum_{j=1}^{n_i} I\{r_i(x_j) \leq t\}/n_i$ where n_i is the number of sequences in group i, and I is the indicator function. The Kolmogorov-Smirnov (KS) statistic is $\text{KS} = \sup_t |f_1(t) - f_2(t)|$ and is estimated from the corresponding ecdfs. Without loss of generality let $\hat{f}_1(t^*) > \hat{f}_2(t^*)$ at t^*, the value at which the KS statistic attains it's maxima. If we consider a classification rule g based on thresholding the r_i values at t^* then it is immediate that $\text{sensitivity}(g) = \hat{f}_1(t^*)$ and $\text{specificity}(g) = 1 - \hat{f}_2(t^*)$. Thus $\text{KS} = \text{sensitivity}(g) + \text{specificity}(g) - 1$.

WordSpy motif finding inputs

The following settings were used in applying discriminatory motif finding using WordSpy to the Bernstein *et al.*, data of Supplementary Figure 24. A maximum word (motif) length of 12 was declared. Degeneracy and subtle motifs were allowed and no repeat filtering was imposed. The maximum number of motifs examined for each word length was restricted to 100. Both strands were searched. All other inputs were retained at default values.

Acknowledgments

Eran Segal graciously provided data and responded to queries. Hong Gu and Mark Garey provided software assistance. Guo-Cheng Yuan, Jun Liu, and Mike Eisen provided preprints and helpful comments. Mark Gerstein, Yuval Kluger, Yuanyuan Xiao and Ru-Fang Yeh provided helpful comments.

References

[1] Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J. (2006). A genomic code for nucleosome positioning. *Nature*, 442: 772-778.

[2] Yuan G-C, Liu JS. (2007). Genomic sequence is highly predictive of local nucleosome depletion. PLoS Comput Biol 4(1): e13. doi:10.1371/journal.pcbi.0040013.

[3] Richmond T, Davey CA. (2003). The structure of DNA in the nucleosome core. *Nature*, 423: 145-150.

[4] Widom J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys*, 34: 269-324.

[5] Cloutier TE, Widom J. (2005). DNA twisting flexibility and the formation of sharply looped protein-DNA complexes. *PNAS*, 102: 3645-3650.

[6] Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, 39: 1235-1244.

[7] Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, 449: 248-251.

[8] Ferguson-Smith AC, Greally JM. (2007). Perceptive enzymes. *Nature*, 449: 148-149.

[9] Stoffer DS, Tyler DE, McDougall AJ. (1993). Spectral analysis for categorical time series: scaling and the spectral envelope. *Biometrika*, 80: 611-622.

[10] Stoffer DS, Tyler DE, Wendt DA. (2000). The spectral envelope and its applications. *Stat Sci*, 15: 224-253.

[11] Rosen O, Stoffer DS. (2007). Automatic estimation of multivariate spectra via smoothing splines. (2007). *Biometrika*, 94: 335 345.

[12] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16: 412–424.

[13] Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. (2004). Global nucleosome occupancy in yeast. *Genome Biol*, 5: R62.

[14] Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36: 900-905.

[15] Freund Y, Schapire RE. (1996). Experiments with a new boosting algorithm. *Proc 13th Intl Conf Machine Learning.* San Francisco: Morgan Kauffman, 148-156.

[16] Friedman JH. (2001). Greedy function approximation: A gradient boosting machine. *Ann Stat*, 29: 1189-1232.

[17] Lawrence CE, Altschul SF, Bogouski MS, Liu JS, Neuwald AF, Wootton JC. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262: 208-214.

[18] Bailey TL, Elkan C. (1995). Unsupervised learning of multiple motifs in biopolymers using EM. *Mach Learn*, 21: 51-80.

[19] Hertz GZ, Stormo GD. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15: 563-577.

[20] Hughes JD, Estep PW, Tavazoie S, Church GM. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296: 1205-1214.

[21] van Helden J, Andre B, Collado-Vides J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281: 827-842.

[22] Sinha S, Tompa M. (2003). YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 31: 3586-3588.

[23] Bussemaker H, Li H, Siggia E. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci USA*, 97: 10096-10100.

[24] Gupta M, Liu J. (2003). Discovery of conserved sequence patterns using a stochastic dictionary model. *J Am Stat Assoc*, 98: 55-66.

[25] Sinha S. (2003). Discriminative motifs. *J Comp Bio*, 10: 599-615.

[26] Wang G, Yu T, Zhang W. (2005). WordSpy: identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Res*, 33: W412-6.

[27] Segal MR. (2007). Prediction of RNA Splice Signals. In *Statistical Advances in Biomedical Sciences: State of the Art and Future Directions.* A Biswas, S Datta, J Fine and MR Segal, Editors, In Press.

[28] Segal E, Barash Y, Simon I, Friedman N, Koller D. (2002). From promoter sequence to expression: A probabilistic framework. In *Proc 6th Intl Conf Research in Computational Molecular Biology (RECOMB)*, Washington DC, 263-272.

[29] Ye K, Kosters WA, Izerman AP. (2007). An efficient, versatile and scalable pattern growth approach to mine frequent patterns in unaligned protein sequences. *Bioinformatics*, 23: 687-693.

[30] Wang G, Zhang W (2006). A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol*, 7(6): R49.

[31] Narlikar L, Gordan R, Hartemink AJ. (2007). Nucleosome occupancy information improves *de novo* motif discovery. In *Proc 11th Intl Conf Research in Computational Molecular Biology (RECOMB)*, Oakland, CA.

[32] Tompa M, Li N, Bailey TL, Church GM et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23: 137-44.

[33] Anderson JD, Widom J. (2001). Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol Cell Biol*, 21: 3830-3839.

[34] Hastie TJ, Tibshirani RJ, Friedman JH. (2001). *The Elements of Statistical Learning*. New York: Springer.

[35] Collins K, Gu H, Field C. (2006). Examining protein structure and similarities by spectral analysis technique. *Stat Appl Genet Mol Biol*, 5, Article 23.