Nucleic Acids Research, 2005, Vol. 33, No. 21 ***6743–6755***
doi:10.1093/nar/gki977

Improved alignment of nucleosome DNA sequences using a mixture model

Ji-Ping Z. Wang* and Jonathan Widom[1]

Department of Statistics, 2006 Sheridan Road and [1]Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, IL 60208, USA

Received September 6, 2005; Revised October 22, 2005; Accepted November 7, 2005

ABSTRACT

DNA sequences that are present in nucleosomes have a preferential ~10 bp periodicity of certain dinucleotide signals (1,2), but the overall sequence similarity of the nucleosomal DNA is weak, and traditional multiple sequence alignment tools fail to yield meaningful alignments. We develop a mixture model that characterizes the known dinucleotide periodicity probabilistically to improve the alignment of nucleosomal DNAs. We assume that a periodic dinucleotide signal of any type emits according to a probability distribution around a series of 'hot spots' that are equally spaced along nucleosomal DNA with 10 bp period, but with a 1 bp phase shift across the middle of the nucleosome. We model the three statistically most significant dinucleotide signals, AA/TT, GC and TA, simultaneously, while allowing phase shifts between the signals. The alignment is obtained by maximizing the likelihood of both Watson and Crick strands simultaneously. The resulting alignment of 177 chicken nucleosomal DNA sequences revealed that all 10 distinct dinucleotides are periodic, however, with only two distinct phases and varying intensity. By Fourier analysis, we show that our new alignment has enhanced periodicity and sequence identity compared with center alignment. The significance of the nucleosomal DNA sequence alignment is evaluated by comparing it with that obtained using the same model on non-nucleosomal sequences.

INTRODUCTION

The genomic DNA of all eukaryotes exists not as naked DNA, but rather as a protein–DNA complex known as chromatin, in which the DNA is locally folded and compacted through a hierarchical series of levels by interaction with proteins known as histones (3). In the first level of compaction, a short stretch of DNA, 147 bp in length, is wrapped in ~1 3/4 superhelical turns about a small disk-shaped octamer of histone proteins, yielding a structure known as the nucleosome core particle, henceforth simply 'nucleosome'. This architectural motif is repeated at intervals, separated by short stretches of unwrapped linker DNA, along the full length of each chromosomal DNA molecule. The structure of the nucleosome has been determined at atomic resolution by X-ray crystallography (4), and steric constraints governing the separation of nucleosomes along the chromosome have been defined (5). Subsequent levels of the chromatin folding hierarchy are less well characterized (6,7).

The steric consequences of wrapping DNA in nucleosomes creates both obstacles and opportunities for protein–DNA interaction, and links the detailed nucleosomal organization of the genomic DNA closely with chromosome function (7–10). Many factors could, in principle, be responsible for governing where nucleosomes are positioned along the genome; but a growing body of evidence demonstrates that the genomic DNA sequence itself is among the dominant determinants of nucleosome positioning *in vivo* (11–20). The DNA sequence features that are most important for nucleosome positioning are ~10 bp periodic recurrences of certain dinucleotides. These dinucleotides, reiterated in phase with the DNA helical repeat, help overcome the natural inflexibility of random sequence DNA, thereby facilitating the DNA's ability to wrap tightly around the histone core (21,22). Taken together, these disparate observations demonstrate that eukaryotic genomes are evolved and constrained to facilitate their own organization into chromatin. For these reasons there is much interest in developing methods to predict DNA sequence-directed nucleosome positioning, genome-wide.

This prediction problem is difficult and has not yet been solved. However, it is closely related to, and could benefit greatly from the solution of, a potentially simpler problem: alignment of DNA sequences that were present in actual nucleosomes. Many earlier studies have attempted to align nucleosomal DNA sequences directly [(1,23–26) and references therein]. Existing multiple sequence alignment methods, including PILEUP (http://www.gcg.com), Clustalw (27),

*To whom correspondence should be addressed. Tel: +1 847 467 6896; Fax: +1 847 491 4939; Email: jzwang@northwestern.edu
Correspondence may also be addressed to Jonathan Widom. Tel: +1 847 467 1887; Fax: +1 847 467 6489; Email: j-widom@northwestern.edu

Gibbs motif sampler (28,29), and hidden Markov models (30–34) consistently fail to yield meaningful alignments on natural nucleosomal DNA sequences. In an alternative approach, nucleosomal DNA sequences were encoded for particular statistically significant features, and then cross-correlation approaches were used to align the encoded sequences. This approach successfully aligned a subset of selected non-natural nucleosomal DNAs (25,26), but it has not succeeded in producing meaningful alignments of natural nucleosomal DNAs (24) (data not shown).

Another alternative approach took advantage of the micrococcal nuclease (MNase) digestion procedure that is used to biochemically isolate individual nucleosomes from chromatin (1). As the nuclease digestion proceeds, individual nucleosomes are liberated from the chromatin filament, then the remaining stretches of linker DNA are nibbled away until only the fully wrapped DNA (147 bp) remains. In practice, the protection afforded by the nucleosome against digestion is incomplete, and one is left with a mixture of nucleosomes containing DNAs that vary in length around ~147 bp. Travers and colleagues (1) sequenced 177 such DNAs, which varied in length from 142 to 149 bp, and aligned the resulting sequences about their centers by assuming that the MNase would digest the linker DNA stretches at each end with approximately equal efficiency. The resulting alignment is referred to as the 'center-alignment' here. However, a phase disturbance between positions 52 and 72 for the AA/TT signal in this alignment predicted a local maximum of probability for AA/TT at the nucleosome dyad axis (where the minor groove faces 'out', away from the histone octamer). This prediction disagrees with existing notions on the sequence-dependent anisotropic flexibility of AA/TT steps (1); moreover, no such phase disturbance is seen in the alignments computed from the selected non-natural nucleosome sequences (26) or in an alignment of natural chromatosomal sequences (the chromatosome includes the nucleosome core particle plus histone H1 and an additional 20 bp of DNA) (35). In fact, because of the known sequence preferences inherent to MNase, it is not expected that the enzyme digestions would proceed with identical rates at the two ends of every nucleosome. Taken together, these findings suggest that this center-alignment strategy is unlikely to yield the best possible alignment.

In this paper, we propose a new Gaussian mixture model approach to nucleosome sequence alignment. Our approach models the periodicity of multiple dinucleotide signals simultaneously, while allowing for variable phase shifts between them. Alignment of nucleosomal DNAs is obtained by maximum likelihood estimation given the model. Using this model we compute a new alignment of the collection of 177 chicken nucleosome sequences. The new alignment is superior to that obtained previously for these same sequences using the center-alignment strategy, and it recapitulates and enhances key findings from the alignments of selected non-natural nucleosome sequences.

MATERIALS AND METHODS

Definition of nucleosome core DNA alignment

DNA or protein sequence alignments are conventionally constructed by maximizing the column-wise similarity of sequences in the aligned region(s) (36). Nucleosomal DNA sequence alignment is, however, defined on a DNA structural or mechanical basis, rather than on direct column-wise similarity (21,22). Our approach replaces the unsolved problem of globally aligning nucleosomal DNA sequences with the related but distinct problem of aligning a set of such DNA sequences onto a nucleosome. We index positions within the nucleosome as $1, 2, \ldots, 147$ from the $5'$ end to the $3'$ end. Suppose we have a set of n DNA sequences $\mathbf{S} = \{S_i : i = 1, \ldots, n\}$, such as the collection of chicken nucleosome sequences, each of which is derived from, but imperfectly covered by, a nucleosome. We then seek to determine the positioning of each actual sequence with reference to the nucleosome. If all of the nucleosomes were perfectly digested to exactly the 147 bp of wrapped DNA, then they would be automatically aligned, even if the base composition at certain positions were completely random. However, the existing sequence collection is imperfectly digested, leaving us with the problem of determining the shift needed to align each sequence onto the nucleosome.

To facilitate formulation of the alignment algorithm, we define the shift parameter δ_i as the signed distance from the first nucleotide of S_i ($5'$ end) to the first position of the nucleosome (see Figure 1). A positive or negative sign of δ_i means that the nucleosome starts within S_i or upstream of S_i, respectively. A δ_i equal to 0 means that the nucleosome starts exactly at the first nucleotide of S_i. With these definitions, an observed position x in sequence S_i thus corresponds to a true aligned position $x - \delta_i$ with reference to the nucleosome position. Aligning the sequences in $\mathbf{S}$ is equivalent to determining the shifts δ_i for $i = 1, \ldots, n$.

A mixture model: capturing 'hot spots' while allowing variability

Although the significance of the ~10 bp periodicity of some key dinucleotides (e.g. AA/TT, GC and TA) in nucleosomal DNAs is well established, only rarely does any one of these motifs recur with 10 bp spacing in any natural 147 bp-long DNA. This suggests that the periodicity of a particular dinucleotide signal exists as an average feature across nucleosomes in the genome, in the sense that the distance between two neighboring signals tends to be ~10 bp in expectation, while the actual distance randomly deviates from 10 bp according to some distribution. This assumption is natural if the special dinucleotides act by locally increasing DNA

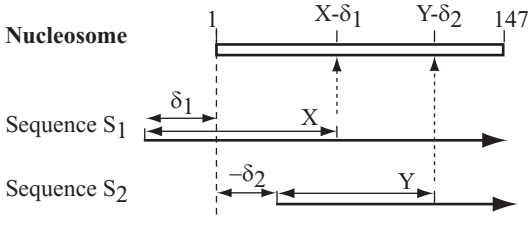

Figure 1. A diagram of nucleosomal DNA sequence alignment. The positions along a nucleosome are indexed as $1, 2, \ldots, 147$ from the $5'$ end to the $3'$ end. The alignment shift δ_i is defined as the signed distance from the first nucleotide of sequence S_i to the first position of the nucleosome core. Aligning the nucleosomal DNA sequences in a set $\mathbf{S} = \{S_i : i = 1, \ldots, n\}$ is equivalent to determining the shift parameter δ_i for each i. A position x in an unaligned sequence S_i corresponds to the position $x - \delta_i$ with reference to the nucleosome position.

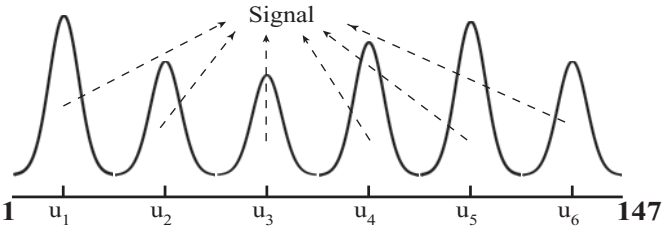

Figure 2. The mixture model captures 'hot spots' while allowing variability. This model hypothesizes that there are a series of hot spots in the nucleosome core region for a particular dinucleotide signal of interest. The probability of observing a dinucleotide signal of this type decays with distance from the hot spot.

flexibility (21,22); and, more importantly, it is supported by experimental data (1).

To account for the averaged periodicity of dinucleotides while allowing for variability in their detailed locations, we propose a location mixture model. We suppose that, for a particular type of dinucleotide signal (e.g. AA/TT), there are a series of 'hot spots' positioned sequentially along the nucleosome at u_m for $m = 1, \ldots, M$ (see Figure 2). Signals emit around the mth hot spot according to a distribution f indexed by the hot spot location, i.e. $f(;u_m)$. We suppose further that, given an observed signal, the probability that it emits from the mth hot spot is p_m, subject to $\sum_{m=1}^{M} P_m = 1$. Hence, the magnitude of p_m indicates the degree of 'hotness' of the mth spot. With these assumptions, the probability of observing a dinucleotide signal at position x can be expressed as a mixture distribution as follows:

$$f(x;\, \mathbf{u}, \mathbf{p}) = \sum_{m=1}^{M} f(x;u_m)p_m. \qquad \mathbf{1}$$

In statistics nomenclature, the density $f(x;\, u_m)$ is often called the mth component distribution in the mixture and p_m its weight (37–39). In the following, we reserve, without specification, an f with vector parameters ($\mathbf{u}$, $\mathbf{p}$) for the mixture distribution, and an f with only one scalar parameter e.g. $f(x;\, u_m)$, for a single component distribution.

We consider K different dinucleotide signals simultaneously. Let $\mathbf{u}_k = (u_{k1}, \ldots, u_{kM_k})$ be the hot spot locations, $\mathbf{p}_k = (p_{k1}, \ldots, p_{kM_k})$ be the component weights and M_k (which may vary with k) be the total number of hot spots for the kth dinucleotide signal. Let x_{ijk} be the observed location of the jth dinucleotide of the kth type from the ith sequence (unaligned) for $i = 1, \ldots, n$, $j = 1, \ldots, J_{ik}$ and $k = 1, \ldots, K$. After alignment, the true position of this signal at x_{ijk} becomes $x_{ijk} - \delta_i$ (see definition of δ_i above). Suppose the kth type dinucleotide signal follows a mixture distribution as

$$f(x_{ijk};\, \mathbf{u}_k, \mathbf{p}_k, \delta_i) = \sum_{m=1}^{M_k} f(x_{ijk} - \delta_i;u_{km})p_{km}. \qquad \mathbf{2}$$

Let $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_K)$, $\mathbf{p} = (\mathbf{p}_1, \ldots, \mathbf{p}_K)$ and $\mathbf{d} = (\delta_1, \ldots, \delta_n)$. If we assume the emissions of these signals are independent, then the log likelihood can be written as

$$\ell(\mathbf{U}, \mathbf{P}, \mathbf{d}; \mathbf{x}) = \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{J_{ik}} \log[f(x_{ijk} - \delta_i;\mathbf{u}_k, \mathbf{p}_k)]. \qquad \mathbf{3}$$

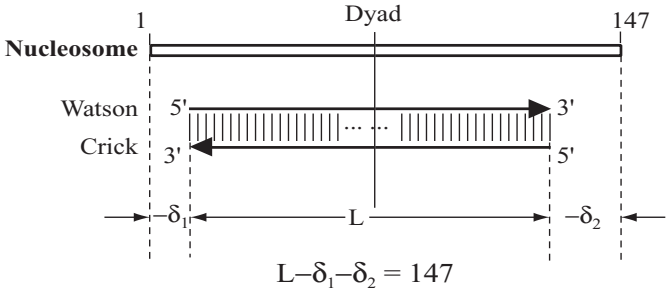

Figure 3. Palindromic symmetry and alignment constraint. For a pair of Watson and Crick strands S_1 and S_2 of length L, we require that the alignment shift parameters δ_1, δ_2 satisfy the constraint $L - \delta_1 - \delta_2 = 147$. Palindromic symmetry is imposed by demanding that the shifts for each strand of a given sequence be optimized simultaneously, subject to this constraint.

In this article, we first model only the three dinucleotide signals for which the ~10 bp periodicity has been proven to be statistically the most significant: AA/TT, GC and TA (1,2). These will correspond to $k = 1, 2, 3$, respectively. (Note: by writing AA/TT, we are treating AA and TT signals as equivalent. This assumption is tested and confirmed, below). Even for this reduced model with $K = 3$, there are too many parameters, and the problem is further complicated by the inherent weakness of the signals in any given nucleosomal DNA. Maximization of the likelihood over this entire parameter space is an intimidating problem. We therefore introduced several further simplifications into the model to reduce the number of parameters, and to take advantage of independent information that is present in each of the two DNA strands.

Model simplification

Our major simplification utilizes the ~10 bp periodicity of key dinucleotides: we suppose that, for each periodic dinucleotide signal, neighboring hot spots are spaced by ~10 bp. Therefore, if the first position of hot spot $\mathbf{u}_k$ is known, then the positioning of the remaining hot spots for that signal is automatically determined:

$$u_{km} = u_{k1} + (m - 1) \times 10, m = 2, \ldots, M_k, k = 1, 2, 3.$$

Since nucleosomal DNA is 147 bp long, M_k could be either 14 or 15 for different signals, depending on how close to one end of the nucleosome the first hot spot is located. One remarkable feature of the proposed method is its capability of detecting M_k even if the initial value for M_k is mis-specified (see Results). The parameterization of $\mathbf{U}$ now is simplified as (u_{11}, u_{21}, u_{31}). The phase shifts between signals can then be calculated based on (u_{11}, u_{21}, u_{31}).

Alignment constraint and center symmetry

Figure 3 illustrates a fragment of double-stranded DNA of length L, with the two strands ('Watson' and 'Crick') labeled as S_1 and S_2, respectively. δ_1 and δ_2 are the signed distances of the 5′ ends of S_1 and S_2, respectively, from the corresponding edges of the nucleosome (positive if the corresponding DNA 5′ end extends beyond the end of the nucleosome, or negative otherwise). By definition of the nucleosome alignment (Figure 1), δ_1 and δ_2 are simply the alignment shifts for S_1 and S_2. Mathematically, this implies that δ_1 and δ_2 satisfy the

following constraint:

$$L - \delta_1 - \delta_2 = 147. \quad \mathbf{4}$$

This constraint means that we do not allow the two strands to slide past each other in the alignment. By averaging together the limited information present on the two strands of natural nucleosomal DNA, we can improve the signal-to-noise ratio of each DNA sequence. Because we are aligning both Watson and Crick strands simultaneously under the constraint (4), a AA/TT, GC or TA signal present at position x (after alignment) on one strand implies the existence of another AA/TT or GC or TA signal, respectively, on the other strand at position $147 - x$. Therefore, the hot spot locations specified in $\mathbf{u}_k$ is pre-determined to be center symmetric about position 73.5, reflecting the 2-fold (dyad) rotational symmetry of the nucleosome (4). The center is shifted leftward by one half base relative to the position of the nucleosome dyad axis (at position 74) because the right-most start position of a dinucleotide signal inside the nucleosome is position 146, not 147. It turns out that there are only two possible center-symmetric positionings of $\mathbf{u}_k$ under the strict 10 bp spacing, i.e. either with a central hot spot at position 73.5 or with two hot spots located ±5 bp about position 73.5. This pre-determined structure of $\mathbf{u}_k$ captures the well-known results from the center alignment and Fourier analysis of natural nucleosome sequences (1,2), and of the cross-correlation alignment of the selected non-natural nucleosome sequences (26), but here follows analytically from the model. The alignment algorithm will automatically detect this structure at convergence, and thereafter we will make an adjustment to the hot spot spacing such that $\mathbf{u}_k$ takes integer values, while complying with this constraint.

Gaussian mixture

We use a Gaussian distribution for f for its simplicity and effectiveness in this problem (see Discussion). A Gaussian distribution carries two parameters, the mean and variance, which are often referred as the location and scale parameters. The density is uni-modal and symmetric about the mean. The Gaussian mixture model defined here has mean parameters specified in $\mathbf{u}_k$ for $k = 1, 2, 3$, but with a pre-specified common variance $\sigma^2 = 2.5^2$. We chose $\sigma = 2.5$ because a Gaussian distribution with mean u_{km} and variance 2.5^2 has ~95% coverage in the interval $u_{km} \pm 2 \times 2.5$, which spans nearly a full DNA helical turn. We found that the alignment results are essentially independent of the choice of σ for values ranging from 1.5 to 5. For this reason, σ was treated as a known constant in the alignment rather than an unknown parameter. The alignment shifts δ_i are obtained by maximum likelihood estimation using the Expectation–Maximization (EM) algorithm. The details of the algorithm are available in supporting material.

RESULTS

Simultaneous alignment using AA/TT, GC and TA signals

We apply our method to the set of 177 chicken nucleosome DNA sequences obtained and analyzed previously by Travers and colleagues (1). The sequences ranged in length from 142 to 149 bp, but those longer than 146 were truncated to 146 bp (A. Travers, personal communication). We first consider alignment using only the dinucleotide signals AA/TT, GC and TA, which have been shown in earlier studies to be the most statistically significant periodic features of nucleosomal DNA (1,25,26).

We tentatively initialize the parameters as follows:

$$\mathbf{u}_1^{(0)} = \mathbf{u}_2^{(0)} = \mathbf{u}_3^{(0)} = (1, 11, \ \ldots, 141),$$
$$p_{mk}^{(0)} = 1/M_k = 1/15, k = 1, 2, 3, \quad m = 1, \ \ldots, 15,$$
$$\delta_i^{(0)} = 0, \ i = 1, \ \ldots, n.$$

Note that $\mathbf{u}_1$ here has 15 hot spots, as do $\mathbf{u}_2$ and $\mathbf{u}_3$. As we discussed above, M_k could be either 14 or 15 if strict 10 bp periodicity holds. In this tentative run, we will first determine M_k for each k. Then, based on these results, we will adjust the parameters accordingly. We initialized $\mathbf{u}_k$ with the same values for $k = 1, 2, 3$ purposely to allow these signals to compete on an equal footing for relative influence on the alignment.

The results at convergence of this initial alignment run are illustrated in Figure 4A. For each of the three dinucleotide signals modeled (AA/TT, GC and TA), we plot their frequency of occurrence as a function of position in the nucleosome. The results are shown as a 3 bp moving average, to eliminate the 3 bp periodicity due to codons and allow for direct comparison with the earlier 'center-alignment' of these same sequences (1) (see also Introduction).

Figure 4A reveals important features of the M_k and phase shifts. The mixture model detected 14 true peaks for the AA/TT and TA signals, despite the incorrect initial specification of 15 hot spots for these signals. The extra hot spot was placed outside the nucleosome, at position −1.5, and had weight ≈0. In addition, the AA/TT and TA signals were in phase, with a half-period phase shift (5 bp) relative to the GC signal.

At convergence, the $\mathbf{U}$ estimate as shown in Figure 4A was:

$$\hat{\mathbf{u}}_1 = (-1.5, 8.5, \ \ldots, 138.5),$$
$$\hat{\mathbf{u}}_2 = (3.5, \ \ldots, 143.5),$$
$$\hat{\mathbf{u}}_3 = (-1.5, \ \ldots, 138.5).$$

Since the center-symmetry point is at position 73.5, our imposition of a strict 10 bp spacing required that all of the hot spots take positions at half bases. For example, for the AA/TT and TA signals, if the nucleosome actually has two true hot spots at positions 68 and 79 that are also symmetric about position 73.5, the exact 10 bp periodicity imposed on $\mathbf{u}_1$ forces the algorithm to place the hot spots at positions 68.5 and 78.5. Similarly, for the GC signal, a true hot spot at position 73 must be paired with one at position 74, but the algorithm will form only one hot spot at position 73.5 instead. To make these hot spots start at integer positions, while maintaining the required center-symmetry about position 73.5, we adjusted the hot spot locations as follows:

$$\mathbf{u}_1 = (8, \ \ldots, 68, 79, \ \ldots, 139),$$
$$\mathbf{u}_2 = (3, \ \ldots, 73, 74, \ \ldots, 144),$$
$$\mathbf{u}_3 = \mathbf{u}_1.$$

The two central hot spots for AA/TT and TA are now spaced 11 bp apart; the central hot spot for GC at position 73.5 is split into a pair of hot spots at positions 73 and 74. Because of the

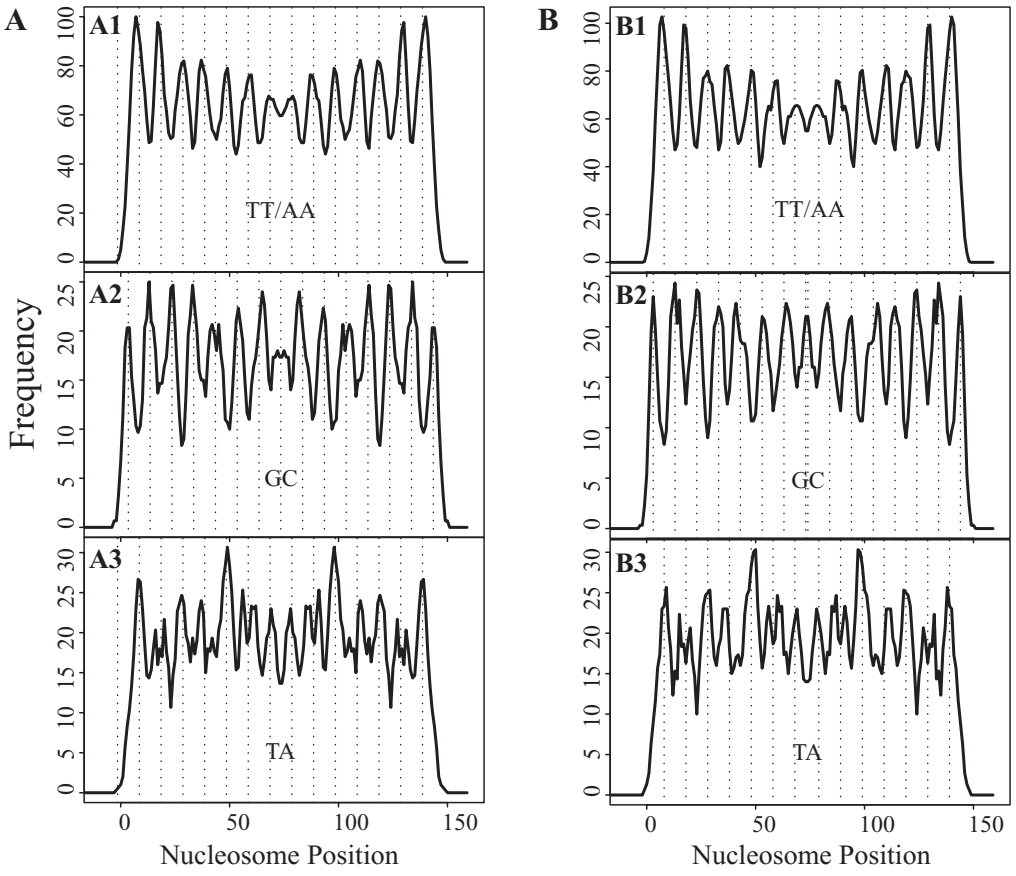

Figure 4. Plots of dinucleotide frequency averaged over a 3 bp window for alignments under strict 10 bp periodicity with initial setting $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}_3 = (1, 11, \ldots, 141)$ (A) and the adjusted setting with $\mathbf{u}_1 = \mathbf{u}_3 = (8, \ldots, 68, 79, 89, \ldots, 139)$ and $\mathbf{u}_2 = (3, 13, \ldots, 73, 74, 84, \ldots, 144)$ (B).

strict center-symmetry of initial values of **U**, the update of it remains the same in each iteration (see algorithm in supporting material).

The results at convergence of this new alignment run are illustrated in Figure 4B. The dinucleotide frequency plots for AA/TT, GC and TA resemble those of Figure 4A but differ in detail. The positions of the peaks in these new frequency plots agree better with the parameters **U** (particularly for the TT/AA signal), implying that this adjusted model better represents the signals in the actual sequences. This conclusion is further supported by our Fourier analysis results, where the normalized amplitude [fractional variation in occurrence (FVO), see definition below] at the optimal periodicity around 10 bp was uniformly improved in the new alignment: the FVOs for AA/TT, GC and TA in the alignment of Figure 4B versus A were 0.29 versus 0.27, 0.34 versus 0.32 and 0.22 versus 0.20, respectively.

The relative locations of the three dinucleotide signals in this alignment agree with those determined from the alignment of selected non-natural nucleosomal DNAs (26), which was computed using unrelated methods, providing further evidence that this new alignment is a good one. As anticipated (26), whereas the selected sequences show strongest alignment only for the central ~71 bp over which selective pressure was exerted, this new alignment of the natural nucleosomal DNAs extends over the full length of the nucleosome.

Compared with the center alignment of these same sequences (1), 26 of the 177 sequences maintained the identical shifts in the mixture model alignment, and 47 moved by only 0.5 bp, while a majority (104 sequences, ~59%) moved by 1 bp or more. The difference of the two alignments in terms of shifts δ_i is summarized in a histogram that is available in supporting material.

Our new alignment systematically improves both the periodicity and amplitudes of the peaks across the full nucleosome length, compared with the center alignment. Our alignment differs strikingly from the center alignment over the middle of the nucleosome. The center alignment revealed a phase reversal of the AA/TT and GC signals near the center of the nucleosomal DNA, such that there is a local maximum of probability of AA/TT dinucleotides and a minimum of probability of GC dinucleotides at the dyad axis. In contrast, our new alignment resembles our alignment of non-natural nucleosomal DNAs and maintains the phases of these dinucleotides across the full nucleosome length with only a 1 bp jump across the middle (which is needed to satisfy the nucleosome symmetry constraints). These near-constant phases in our new alignment result in a minimum of probability of AA/TT dinucleotides at the nucleosome dyad symmetry axis, and a maximum probability of GC dinucleotides there, exactly opposite the results from the center alignment. The alignments are compared further using Fourier analysis, below.

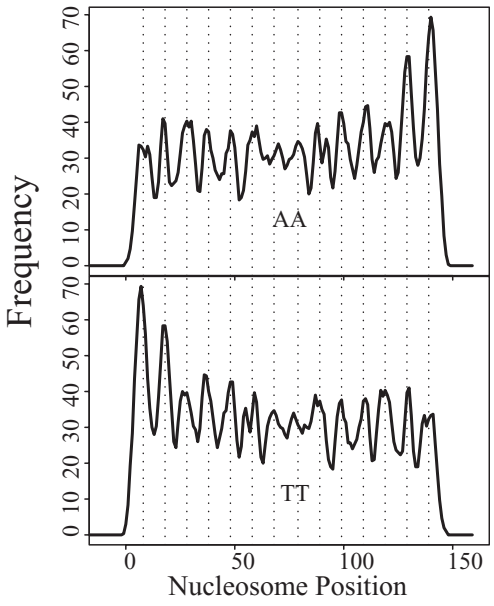

Figure 5. Frequency plot of TT and AA signals in the alignment presented in Figure 4B.

The alignment in Figure 4B reveals a progressive decrease in the AA/TT signal from either end inward toward the center, suggesting that placing AA/TT dinucleotides nearby hot spots that are located near the nucleosome ends benefits histone–DNA interactions more than do AA/TT dinucleotides placed nearby more-centrally located hot spots. Individual plots for the AA and TT signals (Figure 5) reveal pairs of particularly strong peaks. For the TT dinucleotide, these occur at positions 8 and 18, close to the 5′ end of the nucleosome; AA dinucleotides reveal an equivalent pair of peaks, at the symmetry-related locations near the 3′ end. GC and TA dinucleotide signals did not reveal comparable systematic decays (increases) over the nucleosome length (Figure 4B). For the GC signal, the first and last peaks were lower than neighboring peaks, but this is probably a consequence of the set of DNA sequences available to us for the alignment, rather than an inherent property of nucleosomal DNA. The DNAs are all shorter than 147 bp, and thus are truncated prior to at least one end of the nucleosome. The TA signal is weaker and appeared to be noisier than the other two. Interestingly, the TA signal, however, reveals a pair of strong peaks at positions 48 and 99, roughly bracketing the central 1/3 of the nucleosome.

The dinucleotide signals for all 10 distinct dinucleotides, resulting from this alignment, are compared in Figure 6, A1–10 (Figure 6, A1–3 are the same as Figure 4, B1–3). Strikingly, most of the dinucleotide signals appear periodic, but with only two distinct phases, either the same as for AA/TT, or as for GC. The TA and AT signals are in phase with AA/TT; all the rest of the clearly periodic signals are in phase with GC, but are weaker, as reflected in their peak-valley ratios across the plots. The CG signal is extremely rare (<10 at any peak position, in the alignment of 354 strands), presumably a reflection of its under-representation in eukaryotic genomes generally. Nevertheless, despite its rarity, the periodicity of the CG dinucleotide appears to be strikingly significant. The amplitudes and phases of each signal are analyzed objectively, using Fourier methods, in a subsequent section.

Simultaneous alignment using all dinucleotide signals

The dominance of the AA/TT, GC and TA signals after alignment could be an artificial consequence of these signals being the only ones on which the alignment is based. Since most dinucleotides appear to be periodic in this alignment, we repeated the alignment procedure using all 10 of the dinucleotide signals. The resulting alignment was essentially the same as that obtained using only AA/TT, GC and TA signals. For example, ~86% of the sequences had δ change $\leq$1 bp. The frequency plots for the 10 signals are available in supporting material.

Test for equivalence of the AA and TT dinucleotide signals

A persistent question in the literature is whether or not AA and TT dinucleotide signals are equivalent in nucleosomal DNAs. The question of the equivalence of TT and AA dinucleotides has been discussed in (40). Here, we provide additional evidence based on our alignment that supports the claim that these two dinucleotides are equivalent. We repeated the alignment procedure outlined above, except we considered four dinucleotide signals: AA, GC, TA and TT, allowing for variable phase shifts between all of them. To minimize the number of parameters, instead of allowing the relative phases of AA and TT to vary freely, we carried out 10 independent calculations in which the relative phases were fixed but incremented in steps of 1 bp from 0 through 9 bp. In each case, the model converged. We assessed the quality of the resulting alignment by looking at the relative heights of peaks and the peak-valley ratios for all four signals. The alignment computed with relative phase shift equal to 0 proved superior to the other nine alignments (data not shown). We conclude that AA and TT signals are used interchangeably in nucleosomal DNAs, and for the subsequent work we consider them together as a single AA/TT signal.

Column-wise base composition frequency

We purposely modeled dinucleotide frequencies rather than single-nucleotide frequencies, such that, were a sequential-dependent structure of neighboring bases to exist, it could be directly accounted for. A natural question is: does such a dependent structure exist? And if so, does it exist throughout the entire nucleosomal DNA? Or does it exist only in particular regions? If there are regions (or the entire sequence) where there is no significant dependent structure, then it might suffice to model only single-nucleotide frequencies instead of dinucleotide frequencies. Let p_x^i, p_y^{i+1} be the true frequencies of base x at position i and of base y at position $i+1$, and let $p_{xy}^{i,i+1}$ be the dinucleotide frequency xy at position i. Based on the observed sequence data after alignment, we would like to test the independence hypothesis as follows:

$$H_0 : p_{xy}^{i,i+1} = p_x^i p_y^{i+1},$$
versus
$$H_a : p_{xy}^{i,i+1} \neq p_x^i p_y^{i+1}.$$

We used the χ^2 test with the alignment of (Figure 6A) to test the independence hypothesis in the 4×4 contingency table for

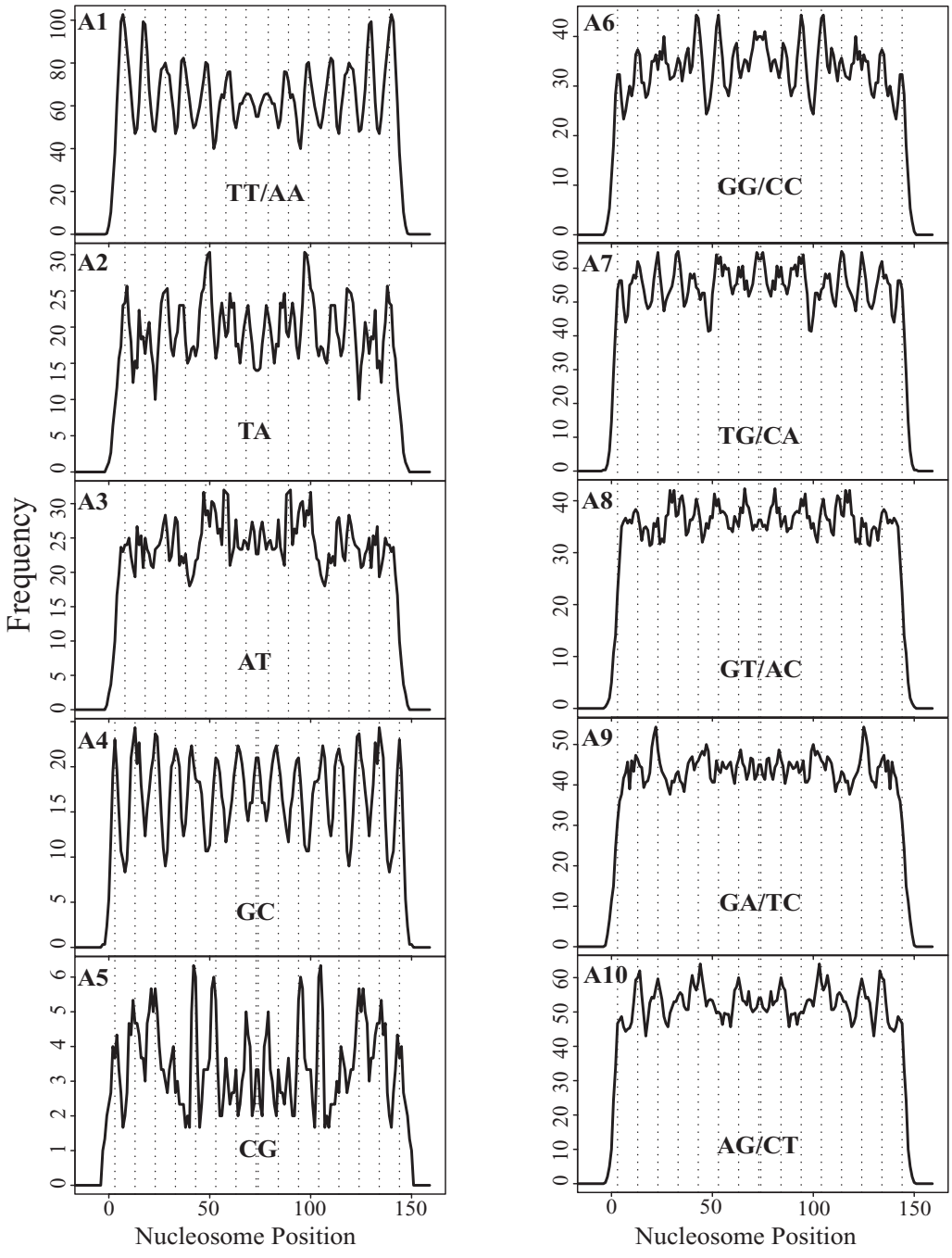

Figure 6. Comparison of dinucleotide frequency plots from the mixture alignment using AA/TT, GC and TA signals.

each $i = 1, \ldots, 146$, where the entry xy in the ith table is the count of dinucleotide xy at position i, for $x, y = A, C, G, T$. Among 146 tests, 144 were significant at level 0.05, and 74 remained significant under the conservative Bonferroni adjustment, i.e. P-value $< 0.05/146$. This justifies our assumption that a dinucleotide xy at position i is not formed simply as an independent combination of two sequential nucleotides x and y at positions i and $i + 1$. It is for this reason that the dinucleotide frequency model is more effective than an independence model that only accounts for the base composition at hot spot sites.

The base composition for A, C, G, T over the entire nucleosome length is plotted in Figure 7A–D, and the frequency of pyrimidine $T + C$ versus purine $A + G$ is plotted in Figure 7E and F. In general, the frequencies of the four bases are ranked as $T > A > C > G$ from the 5′ end inward toward the dyad axis, and $A > T > G > C$ from the dyad axis outward toward the 3′ end. Pyrimidines predominate in the 5′ end half, while purines predominate in the 3′ end half. The sharp increase in pyrimidine frequency at the extreme 5′ end, and of purine frequency at the extreme 3′ end were noted in the original analysis of these sequences, and were attributed to directional

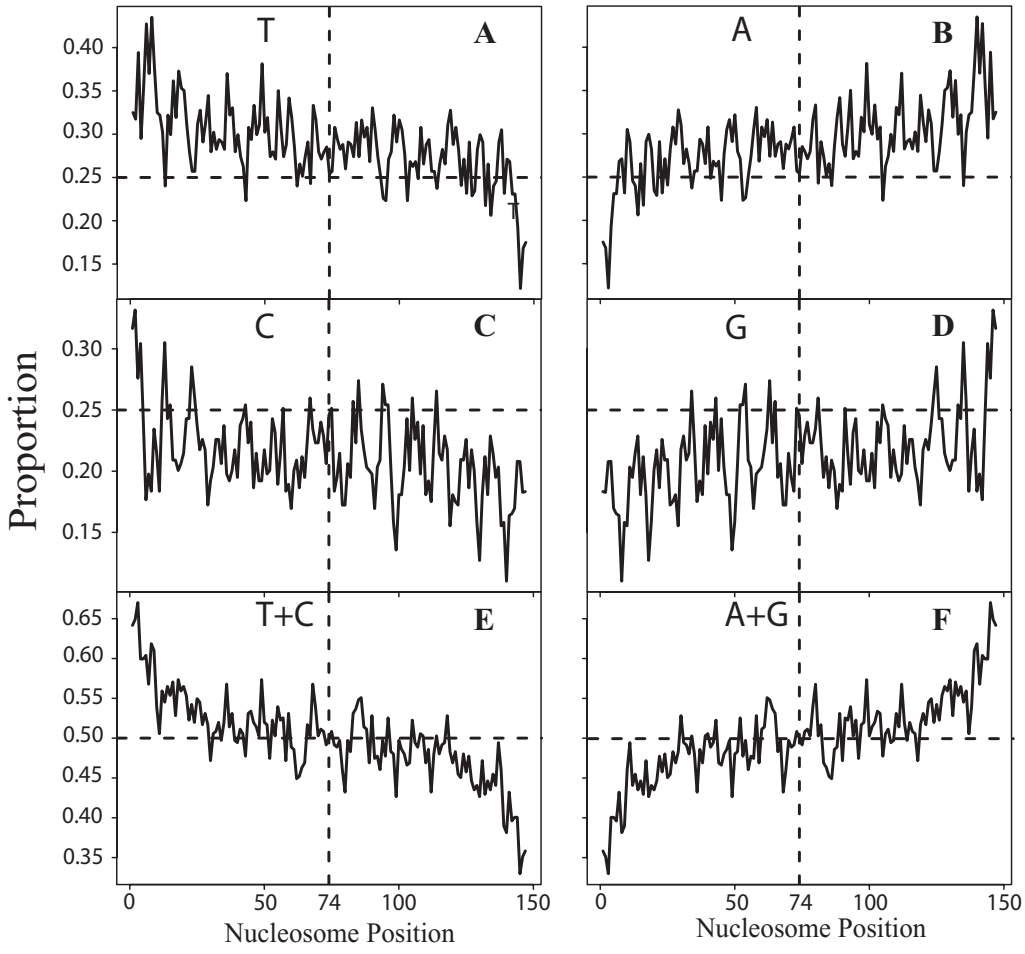

Figure 7. Base composition plot in core region of mixture alignment. (**A**–**D**) Frequencies of T, A, C and G, respectively, in the aligned sequences, plotted as a function of position along the length of the nucleosome. (**E** and **F**) Frequencies of pyrimidines (T+C) or purines (A+G), respectively.

biases occurring during ligation of blunt-ended nucleosomal DNA fragments into the SmaI restriction site of the vector (1). While such effects might influence sequence preferences at the extreme 5′ and 3′ ends, it is difficult to imagine that such effects could influence the base composition at locations far inside the cloned DNA fragments. We conclude instead that at least some of the marked directionality in base composition is an inherent feature of these nucleosomal DNAs.

Fourier analysis

We used the Fourier transform to evaluate the significance of each dinucleotide signal's periodicity. Transforms were evaluated using the raw dinucleotide frequencies in the mixture alignment based on TT/AA, GC and TA signals (referred to as alignment A_1, below). For comparison with the earlier center-alignment analysis of these sequences (1), we included only the 177 'Watson' strands of each sequence, and we also calculated a center alignment as follows: the sequences of length 143, 145 were aligned with $\delta = -2$ [$=-(147–143)/2$] and -1, respectively, and those of length 142, 144 and 146 with $\delta = -2, -1, 0$. This approach follows that used earlier except it uses the newer value of 147 bp, rather than 146, for the true length of nucleosomal DNA (i.e. the center position is at bp 74, instead of at 73.5). We refer to this center alignment as A_2, below. The Fourier transforms were calculated in Matlab, based on the following formula:

$$F(k) = \sum_{j=3}^{144} [f(j) - \bar{f}]\mathrm{e}^{-2\pi jk/N}, \quad k = 0, 1, \ldots, N-1, \qquad \mathbf{5}$$

where N was chosen to be 2000 [for comparison with (1)], $f(j)$ is the raw frequency of a dinucleotide signal observed at position j in each alignment, and $\bar{f} = \frac{1}{142}\sum_{j=3}^{144} f(j)$. To avoid end effects due to sequence truncation we only used $f(j)$ for $j = 3, \ldots, 144$ in $F(k)$ [following the reasoning in (1)]. In Table 1, we report, for each dinucleotide signal, the period $\mathcal{T}^*$ that attained the maximum amplitude, denoted as F^*, over the window between $k = 179$ and 209, corresponding to a range of periods from 9.57 (=2000/209) to 11.17 (=2000/179) bases. To measure the amplitude of a periodicity at $\mathcal{T}^*$ while accounting for substantial differences in numbers of occurrences of differing signals, we followed (1) and defined the FVO as follows:

$$\mathrm{FVO} = \frac{F^*}{\bar{f} \times 5 \times 142/\mathcal{T}^*}, \qquad \mathbf{6}$$

where $142/\mathcal{T}^*$ is the number of periods covered over the window of $j = 3$–144. Because the period $\mathcal{T}^*$ is $\sim$ 10 bp

Table 1. Fourier analysis of variations in the occurrence of dinucleotides

Signal (°)	Alignment	Period (bp)	Amplitude	FVO	Phase (°)
AA/TT	A_1	10.10	672	0.29	−162
	A_2	10.20	389	0.17	−127
GC	A_1	10.05	206	0.34	−2
	A_2	10.10	152	0.25	24
TA	A_1	10.05	151	0.22	−157
	A_2	10.15	81	0.12	−142
TG/CA	A_1	10.15	208	0.11	30
	A_2	10.26	139	0.07	64
GG/CC	A_1	10.15	155	0.13	28
	A_2	10.10	108	0.09	27
GT/AC	A_1	10.58	58	0.05	148
	A_2	10.47	97	0.08	118
AG/CT	A_1	10.10	137	0.07	26
	A_2	—	—	—	—
GA/TC	A_1	—	—	—	—
	A_2	10.64	33	0.02	57
CG	A_1	10.20	30	0.25	55
	A_2	—	—	—	—
AT	A_1	10.10	52	0.06	−151
	A_2	—	—	—	—

Periods that attained largest amplitude within the range of 9.57–11.17 bases are reported [(see also (1)].
FVO: fractional variation in occurrence is the occurrence frequency-normalized amplitude, defined in the text; A_1 refers to the mixture alignment that utilized the AA/TT, GC and TA signals only; and A_2 refers to the center alignment (see text). A '—' means that the signal does not have an amplitude peak within the periodicity range considered (9.57–11.17 bp).

for all signals, FVO is essentially an occurrence-normalized amplitude.

The period and phase angle of certain signals from the center alignment (A_2), reported in Table 1, differed slightly from those in (1). There are two possible reasons for these small differences: (i) we take the center reference position 74 instead of 73.5; (ii) the DNA sequences we used (1) were provided to us in truncated form, where sequences longer than 146 bp were truncated to 146. Regardless of these slight differences, the phase shifts between the most significant signals, including AA/TT, GC, TA, TG/CA and GG/CC, were quite consistent. For example, if we measure phases relative to that of the AA/TT signal, the phase shifts for each of the other four from A_2 compared with (1) differed by $\leqslant$1.5 bp ($\approx$54°). In addition, the FVO for these signals from A_2 are comparable with those reported in (1). Fourier analysis of alignment A_2 yielded no peaks in amplitude near $\sim$10 bp periodicity (between 9.57 and 11.17 bp periodicity) for AG/CT, CG and AT signals, which were also reported as 'not significant' in (1).

The Fourier analysis from alignment A_1 confirmed our impressions from the frequency plots in Figure 6. AA/TT, GC and TA were the three most significant signals in terms of FVO, with $\mathcal{T}^*$ ranging from 10.05 to 10.10 bp. These periodicities are lower than sometimes reported for nucleosomal DNA, but are in the range of previous observations (see Discussion). The CG signal was rare but significant with $\mathcal{T}^* \approx 10.20$ bp. Both the amplitude and the normalized amplitude (FVO) were greater from the alignment A_1 than from the center alignment A_2 for essentially every dinucleotide signal having a significant periodicity at $\approx$10 bases, including AA/TT, GC, TA, TG/CA and GG/CC (GT/AC is the one exception, while GA/TC, AG/CT, CG and AT were not compared because they did not have an amplitude peak in the range of 9.57 and 11.17 in either alignment).

Significance of the alignment

One might think to evaluate the significance of an alignment of the nucleosome sequences by comparing the alignment with that obtained from equivalent computations on non-nucleosome sequences, such as shuffled sequences (random sequences that maintain the nucleotide frequency of the real nucleosome sequences) or natural sequences chosen randomly from the chicken genome (while having the same lengths as our real nucleosome sequences). Such approaches are problematic for several reasons. Of these two sets of non-nucleosome sequences, the randomly chosen real sequences might be the more appropriate and stringent test, but these sequences are problematic because they will in fact often partially overlap with real nucleosomes. Since the average nucleosome repeat length in chicken red blood cells (the cell type from which the real nucleosome sequences derive) is $\approx$208 bp, it follows that any given stretch of the genome has an $\approx$70% (147/208) probability of actually coming from a nucleosome. Moreover, any DNA sequence will incorporate into nucleosomes. Therefore, given even just a random sequence, our algorithm will identify the position along that sequence that best-matches the mixture model's characteristics, and, given a set of such sequences, our algorithm will optimally align them.

We carried out such calculations anyway, using both the shuffled sequences and the randomly chosen natural sequences, and found that the algorithm does align them, as expected. However, significant overall differences were observed in the quality of the resulting alignments compared with the alignments of the natural nucleosome sequences. The relative positioning of AA/TT, GC and TA signals in the hypothesized nucleosome region is sensitive to the initial values of $\mathbf{U}$ used for the alignments of both non-nucleosome sequences, while it is not sensitive for the real nucleosome sequences. For example, when we initialized $\mathbf{u}_k = (1, \ldots 141)$, $k = 1,2,3$, alignment of the shuffled sequences yielded AA/TT and TA signals at $\mathbf{u}_1 = \mathbf{u}_3 = (3.5 \ldots 143.5)$ and GC signals at $\mathbf{u}_2 = (8.5 \ldots 138.5)$ at convergence, contrary to the established locations of these signals in real nucleosomes; and alignment of the randomly chosen chicken sequences yielded AA/TT and GC signals at $(3.5 \ldots 143.5)$ and TA at $(8.5 \ldots 138.5)$, again contrary (but in a different way) to the established locations of these signals in real nucleosomes. When we forced $\mathbf{U}$ to be the same as obtained in the alignments of the real sequences, the resulting alignments of the non-nucleosome sequences were significantly poorer: for example, the FVO of the key signals from Fourier analysis was significantly lower (data not shown).

In Figure 8, we further compare the frequency plots of AA/TT signals resulting from the center and mixture alignments of the nucleosome sequences and the non-nucleosome chicken genomic sequences. Center alignment of the randomly chosen sequences yielded no significant signal, while center alignment of the real nucleosomal sequences yielded a robust signal. This confirms that the experimentally obtained nucleosome DNAs do contain significant information content that uniquely reflects their nucleosomal origin. As expected, the

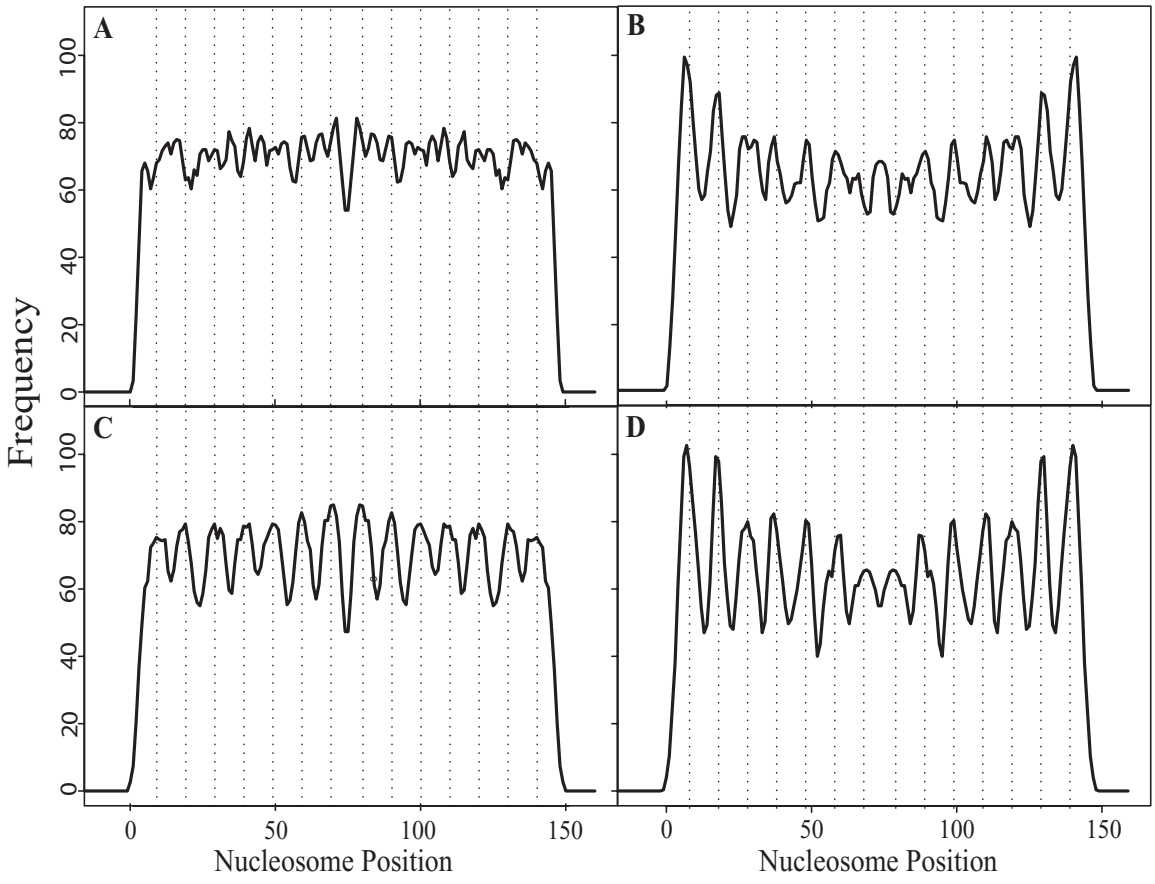

Figure 8. Mixture model alignment compared to center alignment, for chicken nucleosome sequences and randomly chosen chicken genomic sequences. Results for AA/TT signals are shown (A, C, randomly chosen genomic sequences under center and mixture model alignments, respectively; B, D real nucleosome sequences under center and mixture model alignments, respectively); other signals yield comparable results with those shown here for AA/TT.

mixture model successfully aligned the randomly chosen sequences, but the resulting alignment is less good even than the center alignment of the real nucleosome sequences. Finally, as reported above, the mixture model alignment of the natural sequences improves over the center alignment. Quantitative analysis of these alignments by Fourier transformation confirms these visual impressions (Table 1 and data not shown). Based on these findings, we conclude that the alignment computed using the mixture model on the natural nucleosome sequences has significant information content that is uniquely attributable to nucleosomes.

Robustness of the alignment algorithm

The alignment in Figure 6 has been obtained based on the 'profile' trained from all the 354 chicken nucleosome core particle sequences (177 sequences, both strands). One might wonder about the robustness of our approach, for example, how variation of the sequence profile affects the prediction of the alignment shifts, or how sensitive are the predicted alignment shifts δ_is to the size of the training dataset.

We adopted a resampling approach to evaluating the robustness of our alignment method as follows. We repeatedly sampled a fraction of γ of Watson–Crick pairs (without replacement) from the sequence set **S**, 300 times. Each random sample $\mathbf{S}_b$ for $b = 1, \ldots, 300$ was treated as an independent training dataset. Alignment shifts for sequences within $\mathbf{S}_b$ were obtained as described above (using the AA/TT, GC and TA signals only). Our goal was to investigate the consistency of δ_i predicted in these subsamples.

Table 2. Consistency evaluation of the alignment method

γ	$\bar{\beta}$	$\tilde{\beta}$
0.9	0.907	0.963
0.5	0.765	0.996
0.3	0.705	0.996

Let $\mathbf{d}_i$ be the set of predicted shift values for sequence S_i from these subsamples for $i = 1, \ldots, n$; and let δ_i^* be the most frequently obtained shift for S_i in $\mathbf{d}_i$. One simple measure of consistency would be the fraction of the shift values in $\mathbf{d}_i$ that are equal to δ_i^*, for each i. The average of this fraction across i, denoted as $\bar{\beta}$, represents a stringent measure of the consistency of our predictions. We also consider a slightly less-stringent measure of consistency, in which we calculate the fraction of the shift values for S_i that are constant within $\delta_i^* \pm 1$; we denote the corresponding average of this fraction across i as $\tilde{\beta}$. The results of these calculations are reported in Table 2. At $\gamma = 0.9$, 90.7% of the shifts for a particular sequence predicted in the subsamples were identical, and 96.3% were within ±1 bp of the most frequently obtained prediction for each

sequence. As γ decreased to 0.3, the increased variation between subsamples caused a drop of 20% in the consistency measure $\overline{\beta}$. Nevertheless, almost all of the shifts (99.6%) remained within ±1 bp about the mode. These results imply that the mixture 'profiles' for the dinucleotide signals are reasonably consistent across different subsamples of the training data, and they lend further confidence to the overall quality of the alignments.

DISCUSSION

We have developed a novel methodology that improves the alignment of experimentally obtained nucleosome core DNA sequences over what has previously been possible. Our new alignment exhibits enhanced sequence identity and periodicity; it accords with independently computed alignments on independent nucleosome DNA sequences, and it accords with current ideas concerning the sequence-dependence of DNA bendability.

Our 'hot-spot' model is built upon the well-known periodicity of key dinucleotide signals. We have argued that this periodicity is an 'averaged' property in the sense that the dinucleotide signals are positioned with variability around the fixed hot spots that are strictly periodic (with a 1 bp offset across the dyad axis). The hot spot locations (**U**) and their weights (**P**) constitute the two most important aspects of the 'profile' of the nucleosome, which directly determines the optimal alignment shift δ_i for each sequence. If the true profile is given, then δ_i can be obtained independently for each i using the EM algorithm (available at supporting material web site: http://bioinfo.stats.northwestern.edu/jzwang/suppNucleosome.html).

Based on the alignment results in Figure 6, we conclude that the 10 dinucleotide signals fall into three groups as regards the amplitudes of their periodicity: AA/TT, TA, GC, CG > AT, GG/CC, TG/CA > GT/AC, GA/TC, AG/CT (but the CG signal is extremely rare). The Fourier analysis confirms this apparent ranking (with amplitudes expressed as FVOs); moreover, it shows that these signals occur in only two distinct relative phases.

The periodicities resulting from the mixture alignment for the AA/TT, GC and TA signals (10.05–10.10 bp) are lower than the 10.30 bp overall periodicity reported in the atomic resolution crystal structure (4). This difference is not attributable to different reference frames, as the 10.30 bp periodicity represents the value obtained after conversion to the local reference frame, as appropriate for comparison to our results. We note, however, that other studies have reported a wide range of periodicities for differing nucleosome samples, with a range that greatly exceeds the apparent experimental error. High resolution crystallographic studies on 146 bp-containing nucleosomes yielded periodicities of 10.23 and 10.15 bp (41). Center alignment of the chicken nucleosome collection yields periodicities ranging from 10.15 to 10.26 bp (1), while a genome-wide Fourier analysis by (2) yielded 9.9 to 10.3 bp periodicities. High resolution solution analyses of three different nucleosome sequences have yielded values of 10.3 ± 0.1 bp (42), 10.0 or 9.8 bp (for AA/AT/TA/TT dinucleotides), and 10.0 bp (for CC/CG/GC/GG dinucleotides) (43).

Thus, the periodicities implicit in our alignments fall well inside the range of values obtained by others, yet the origin of this apparent wide range of periodicities itself is unclear. At least two factors appear to contribute significantly. First, solution biochemical studies prove that differing DNA sequences incorporate into nucleosomes with different helical twists (43). Therefore, the crystal structures, which were obtained using just two related sequences (with or without one extra base pair), do not necessarily reflect details of the structures and periodicities of nucleosome DNAs generally. Second, the crystallographic studies show that the 146 bp-containing nucleosomes stretch and over-twist their DNA, making up a 1–2 bp length deficit in a space of just 12 bp, to satisfy crystal packing constraints (41). These crystal structures suggest that the nucleosome specifically stabilizes these over-twisted states, and that diffusive motion of such twist defects may play a role in nucleosome mobility and remodeling (4). Our finding that the improved alignment presented here is accompanied by a slightly reduced periodicity suggests that natural nucleosome DNAs may be evolved to favor under-twisted states, perhaps resembling those present in the 146 bp-containing nucleosome crystals.

Importantly, neither the strength of the periodic signals nor the actual alignments that result from our alignment procedure are sensitive to the exact locations (i.e. periodicities) of the hot spots. We noted that the first and last peaks of TT/AA signal (Figure 6, A1) appear to be offset outward by about 1 bp relative to the nearest hot spot locations. This might suggest that these two hot spots for TT/AA actually occur at positions 7 and 140 instead of 8 and 139. Adjusting the hot spot positions accordingly would result in an alignment that has 14 quasi-periodic signals positioned in a 2 bp wider range of the core region than in the earlier alignment. This would increase the resulting apparent periodicity $\mathcal{T}^*$ by about $2/13 = 0.15$ bp, to $\sim$10.25 bp. To test directly whether the alignment was sensitive to the detailed locations and periodicities of the hot spots, we generated an alternative model in which the hot spot locations were chosen to match the locations of maximal positive and negative base roll angles as seen in the high resolution crystal structure (4). Specifically, we set the hot spot locations as follows: AA/TT (6,16,26,38,48,58,68,79,89,99,109,121, 131,141), GC (2,12,22,32,43,53,63,73,74,84,94,104,115,125, 135,145) and TA (same as AA/TT). The resulting FVO and $\mathcal{T}_*$ (parentheses, in bp) for those three signals were 0.28 (10.26), 0.36 (10.20) and 0.22 (10.20), respectively (the dinucleotide frequency plots are presented in the supporting material). These FVOs are essentially identical to those for the same signals in alignment A_1: 0.29, 0.34 and 0.22, respectively (Table 1). The actual alignment resulting from this alternative set of hot spot locations was essentially identical (data not shown). We therefore focused our analysis on the simpler set of hot spot locations used in alignment A_1.

Two factors determine the importance of a dinucleotide signal in the alignment: the number of occurrences, and the periodicity, of the signal. The likelihood equation is weighted by the number of occurrences of each signal (indexed by j in the likelihood); hence, a rare dinucleotide signal is not influential in determining δ_i even if its periodicity is strong, as is the case for CG. On the other hand, since $\log[f(x_{ijk} - \delta_i; u_{km})]$ in the likelihood function is essentially proportional to the negative of the quadratic distance $(x_{ijk} - \delta_i - u_{km})^2/2\sigma^2$ for our model with Gaussian f, by maximizing the likelihood one actually minimizes the total quadratic distance weighted by

z_{ijkm}. If a signal is strong around a particular hot spot u_{km}, the algorithm will minimize the quadratic distance by finding δ_i such that those signals are tightly positioned around u_{km}. In contrast, a relatively flat (or aperiodic) signal is less influential to the likelihood because change of δ_i does not result in significant change in the quadratic distance. This partially explains why the alignment based only on the AA/TT, TA and GC signals gave very similar results regarding the dinucleotide frequency and alignment shifts δ_i (86% δ_is had zero or ±1 bp change) compared with the alignment using all 10 dinucleotide signals.

We chose the Gaussian mixture for its simplicity in computing, especially because closed-form solutions exist in the EM algorithm. One might question the appropriateness of using the Gaussian distribution in our situation, where the actual sample space for x is an integer lattice rather than a continuous domain $(-\infty, +\infty)$. This, however, is not a concern, for two reasons. First, as we commented above, the essential feature of the Gaussian function for our model is the quadratic distance kernel, by which the distance between a signal and a hot spot u_{km} weighted by z_{ijkm} is penalized. Second, one can regard the Gaussian with mean $= u_{km}$ as a diffusive discrete distribution defined on the lattice $u_{km} \pm 1$, $\pm 2, \ldots$, the probability mass of which is proportional to the Gaussian density, i.e. Prob$(x = u_{km} + j) \propto f(j; 0)$, where $f(j; 0)$ is the density at j of a Gaussian distribution with mean 0 and standard deviation as specified ($\sigma = 2.5$ in this work) for $j = 0, \pm 1, \pm 2, \ldots$. With such a definition, the computing algorithm would be exactly the same as used in this article (see details in the supporting material).

Aligning both strands simultaneously greatly improves the stability of the alignment. This constraint requires that a dinucleotide signal outside of the nucleosome region, e.g. $x_{ijk} < 0$, must be paired with one at $147 - x_{ijk} > 147$ on the other strand. Consequently, the quadratic distance penalty is doubly executed on δ_i compared with if each strand were separately aligned without the constraint. This helps prevent Watson–Crick pairs from shifting too far beyond either end of the nucleosome region. More importantly, this strategy accords with the experimental fact that most of the chicken nucleosome core DNA sequences should roughly span the internal region of the nucleosome, given the way in which they were produced and their length (142–146 bp). Hence, every dinucleotide or pair can be regarded as an emission around one of the hot spots. If the sequence is much longer or much shorter than 147 bp, this algorithm will not apply without modification. Our model is unlikely to make good predictions of the nucleosome positioning when applied to sequences that are significantly (more than 10 bp) longer or shorter than the full nucleosome length. On the other hand, the real goal of our work is to predict nucleosome positioning genome-wide. We anticipate that a nucleosomal DNA profile built from the alignment obtained here will facilitate method development towards this goal.

ACKNOWLEDGEMENTS

This work is supported by the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology grant # 1 R01 GM075313 awarded to J.-P. Wang and J. Widom. The authors thank Andrew Travers for providing the chicken nucleosome core DNA sequence data, for help with the center alignment and Fourier analysis, and for comments on the manuscript; the author thank Kelly Thayer for helpful suggestions and comments on the manuscript, Eran Segal for discussion and providing the randomly chosen chicken genomic sequences, and Bruce Spencer and Wenxin Jiang for helpful discussions. Funding to pay the Open Access publication charges for this article was provided by National Institute of General Medical Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Satchwell,S., Drew,H. and Travers,A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.
2. Widom,J. (1996) Short-range order in two eukaryotic genomes: relation to chromosome structure. *J. Mol. Biol.*, **259**, 579–588.
3. van Holde,K.E. (1989) *Chromatin*. Springer-Verlag, New York.
4. Richmond,T.J. and Davey,C.A. (2003) The structure of DNA in the nucleosome core. *Nature*, **423**, 145–150.
5. Widom,J. (1992) A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc. Natl Acad. Sci. USA*, **89**, 1095–1099.
6. Widom,J. (1989) Toward a unified model of chromatin folding. *Annu. Rev. Biophys. Biophys. Chem.*, **18**, 365–395.
7. Widom,J. (1998) Structure, dynamics, and function of chromatin *in vitro*. *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 285–327.
8. Wolffe,A. (1992) *Chromatin Structure and Function*. Academic Press, London.
9. Felsenfeld,G. (1996) Chromatin unfolds. *Cell*, **86**, 13–19.
10. Kornberg,R.D. and Lorch,Y. (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, **98**, 285–294.
11. Pina,B., Burggemeier,U. and Beato,M. (1990) Nucleosome positioning modulates accessibility of regulatory proteins to the mouse mammary tumor virus promoter. *Cell*, **60**, 719–731.
12. Fragoso,G., John,S., Roberts,M. and Hager,G. (1995) Nucleosome positioning on the MMTV LTR results from the frequency-biased occupancy of multiple frames. *Genes Dev.*, **9**, 1933–1947.
13. Lomvardas,S. and Thanos,D. (2002) Modifying gene expression programs by altering core promoter architecture. *Cell*, **110**, 261–271.
14. Linxweiler,W. and Horz,W. (1985) Reconstitution experiments show that sequence-specific histone–DNA interactions are the basis for nucleosome phasing on mouse satellite DNA. *Cell*, **42**, 281–290.
15. Neubauer,B., Linxweiler,W. and Horz,W. (1986) DNA engineering shows that nucleosome phasing on the African green monkey alpha-satellite is the result of multiple additive histone–DNA interactions. *J. Mol. Biol.*, **190**, 639–645.
16. Shen,C.H., Leblanc,B.P., Alfieri,J.A. and Clark,D.J. (2001) Remodeling of yeast CUP1 chromatin involves activator-dependent repositioning of nucleosomes over the entire gene and flanking sequences. *Mol. Cell Biol.*, **21**, 534–547.
17. Shen,C.H. and Clark,D.J. (2001) DNA sequence plays a major role in determining nucleosome positions in yeast CUP1 chromatin. *J. Biol. Chem.*, **276**, 35209–35216.
18. Mai,X., Chou,S. and Struhl,K. (2000) Preferential accessibility of the yeast his3 promoter is determined by a general property of the DNA sequence, not by specific elements. *Mol. Cell. Biol.*, **20**, 6668–6676.
19. Moreira,J.M., Horz,W. and Holmberg,S. (2002) Neither Reb1p nor poly(dA*T) elements are responsible for the highly specific chromatin organization at the ILV1 promoter. *J. Biol. Chem.*, **277**, 3202–3209.
20. Terrell,A.R., Wongwisansri,S., Pilon,J.L. and Laybourn,P.J. (2002) Reconstitution of nucleosome positioning, remodeling, histone acetylation, and transcriptional activation on the PH05 promoter. *J. Biol. Chem.*, **277**, 31038–31047.
21. Widom,J. (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q. Rev. Biophys.*, **34**, 269–324.
22. Cloutier,T. and Widom,J. (2004) Spontaneous sharp bending of double stranded DNA. *Mol. Cell*, **14**, 355–362.

23. Baldi,P., Brunak,S., Chauvin,Y. and Krogh,A. (1996) Naturally occurring nucleosome positioning signals in human exons and introns. *J. Mol. Biol.*, **263**, 503–510.
24. Ioshikhes,I., Bolshoy,A., Derenshteyn,K., Borodovsky,M. and Trifonov,E.N. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J. Mol. Biol.*, **262**, 129–139.
25. Lowary,P. and Widom,J. (1998) New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.*, **276**, 19–42.
26. Thåström,A., Bingham,L. and Widom,J. (2004) Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.*, **338**, 695–709.
27. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
28. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
29. Liu,J., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *The Journal of American Statistical Association*, **90**, 1156–1170.
30. Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA*, **91**, 1059–1063.
31. Baldi,P. and Chauvin,Y. (1996) Hybrid modeling, HMM/NN architectures, and protein applications. *Neural Comput.*, **8**, 1541–1565.
32. Eddy,S.R. (1998) Profile hidden markov models. *Bioinformatics*, **14**, 755–763.
33. Krogh,A. (1998) *Chapter 4, Computational Methods in Molecular Biology. Elsevier Science*, New York.
34. Liu,J., Neuwald,A.F. and Lawrence,C.E. (1999) Markovian structures in biological sequence alignments. *The Journal of American Statistical Association*, **94**, 1–15.
35. Muyldermans,S. and Travers,A.A. (1994) DNA sequence organization in chromatosomes. *J. Mol. Biol.*, **235**, 855–870.
36. Mount,D.W. (2001) *Bioinformatics–sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, New York.
37. Titterington,D., Smith,A. and Makov,U. (1985) *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.
38. Lindsay,B.G. (1995) *Mixture Models: Theory, Geometry and Applications, Institute of Mathematical Statistics*, Hayward, CA.
39. McLachlan,G. and Peel,D. (2000) *Finite Mixture Models*. John Wiley & Sons, New York.
40. Travers,A.A. and Klug,A. (1987) The bending of DNA in nucleosomes and its wider implications. *Philos. Trans. Soc. Lond. Series B*, **317**, 537–561.
41. Davey,C.A., Sargent,D.F., Luger,K., Maeder,A.W. and Richmond,T.J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 å resolution. *J. Mol. Biol.*, **319**, 1097–1113.
42. Gale,J.M., Missen,K.A. and Smerdon,M.J. (1987) UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. *Proc. Natl Acad. Sci. USA*, **84**, 6644–6648.
43. Flaus,A. and Richmond,T.J. (1998) Positioning and stability of nucleosomes on MMTV 3'LTR sequences. *J. Mol. Biol.*, **275**, 427–441.