Multi-alphabet consensus algorithm for identification of low specificity protein–DNA interactions

Anatoly V. Ulyanov* and Gary D. Stormo

Department of Molecular, Cellular and Development Biology, University of Colorado at Boulder, Campus Box 347, Colorado, Boulder, CO 80309-0347, USA

Received September 16, 1994; Revised and Accepted March 10, 1995

ABSTRACT

A method for the identification and characterization of protein–DNA interactions is presented. We have developed an approach for finding unknown multiple patterns that occur imperfectly in a set of several sequences. The pattern may contain letters from the nucleotide alphabet (A, C, G and T) including ambiguous characters (A/C, A/G, A/T; A/C/G, etc.). This method reveals weak DNA signals on an unaligned set of DNA fragments known to be functionally related and assumes no prior information on the sequences’ alignment. It determines the locations of the signals from only the information intrinsic to the sequences themselves. We have applied this method to analyze the binding sites of cAMP receptor protein (CRP). The consensus based on these data are discussed and a comparison of the consensus with the crystal structure of CAP–DNA complex is presented. We further show that in a mixture of DNA sequences, containing binding sites for two different proteins, both classes of binding sites can be discovered simultaneously by this method. The DNA sequences of nucleosome cores from chicken erythrocyte and a set of the other known nucleosomal sequences show existence of symmetrical features in nucleosome-binding DNA sequences. We also show multi-alphabet patterns that can play a role in the phasing signal on the nucleosome DNA molecule and have compared the results with existing models of nucleosome positioning.

INTRODUCTION

The current consensus methods often rely on preliminary approximate alignment of the input data set, although this requirement may be relaxed by allowing the ‘window’ to include the entire sequence (1,2). These methods allow for approximately matched consensus sequences by including in the analysis both the words that actually occur in the sequences and also words in the ‘neighborhood’ of those words, where neighboring words are similar, within some defined distance, of the words that occur. In this work we present an alternative approach to the problem of analyzing unaligned sets of DNA sequences to find consensus sequences within them. The general statistical consensus search method is formulated as a mathematical procedure that calculates the occurrences of all possible degenerate patterns in a given data set of nucleotide sequences and estimates their statistical significances. The main problem arising in such an approach is the enormous number of degenerate patterns possible for even a modest size of consensus sequence. The number of all possible patterns is equal to 15^Length and for a reasonable pattern length of 10 bp, we would need about a 600 billion element array or 1074 Gbytes of memory to keep all of them. We present a statistical method that exactly solves the problem for a 10 letter degenerate alphabet, and an effective, approximate way to solve the problem for the full 15 letter alphabet. The approach uses the results of the exact solution to estimate the probability of every 15 letter degenerate pattern and allows us to reduce markedly the volume of the calculation by estimation of which degenerate patterns have an occurrence significantly different from random.

To compare patterns with different levels of degeneracy, we use a binomial statistic to evaluate a pattern’s occurrence among a set of nucleotide sequences. The assessment of the pattern’s significance by large deviations of the binomial distribution provides accurate information about the distribution of patterns among a data set. The other advantages of this method are that it is not sensitive to a ‘mixture’ of signals and that it can extract the different patterns of different signals even if they are located on the same data set. To demonstrate this capability we mixed two data set. The first is CAP-binding sites and the second set consists of 11 DNA sequences containing promoters regulated by the E.coli LexA protein (14). Both patterns have been found by our method and are detected as the best in two groups of available alternative patterns. This method also is fairly insensitive to signal/noise ratio—e.g. how many random nucleotides surround the signal fragments.

To show the sensitivity we also apply the method to nucleosome data, where experiments have shown that preference of nucleosome positioning exists at least around regulatory regions [see reviews (4) and (5)], but the origin of this preference is still not clearly understood. Moreover, because of the symmetry of the nucleosome particle it was expected that the DNA data set also should have some kind of symmetry in the nucleotide distribution around the dyad axis, but until now it has not been shown although the influence of the short repeated sequences on the energy of interaction have been found (6,7). The problem of the nucleosome positioning still remains unsolved. In general, five different

* To whom correspondence should be addressed at present address: EMBL, Postfach 10.2209, Meyerhofstrasse 1, D-6900 Heidelberg, Germany
mechanisms of nucleosome positioning are proposed. The origin of replication, histone–DNA interaction, other positioning proteins, flanking structures and chromatin folding may be responsible for nucleosome positioning. In the first experiments studying chromatin organization of SV40 and distribution of nucleosomes on single-copy sequences in rat liver and Drosophila nuclei have shown that nucleosome location was random. Meanwhile there are some genomes, such as Tetrahymena thermophila and S. cerevisiae, that may be organized in positional nucleosomes [see review (5)]. In the case of histone–DNA interaction two different signals, 'transitional' and 'phasing', might play a role in the nucleosome positioning. The first model of phasing signal suggests that bending of DNA can be induced by periodical repeats AAmnTTnnn (30). More recent analysis of the chicken chromatin suggested that alternating (A or T)- and (C or G)-rich regions tended to occur with a period of DNA super-helix and may play a role of phasing signal (20, 25). The question of the existence of a translational signal seems much more difficult. Short A-tracks were proposed to be important in translational positioning of nucleosome DNA (20). However, in the experiments with artificial DNA fragments specially constructed with a wide variety of mutations in the central regions, no evidence for translational preference of A-tracks has been reported (7).

We compare two sets of nucleosome DNA sequences. The first set represents nucleosome sequences randomly extracted from total chicken chromatin. The second set is a collection of sequences for mapped nucleosomes published in 1993 (21).

**METHODS AND DATA**

Most DNA binding proteins can bind to multiple sequences. To detect ill-defined sequence patterns in a set of unaligned DNA fragments, we have developed a statistical method that is called multi-alphabet consensus analysis (MACA). This method adopts two main principles. First, different kinds of protein–DNA interactions along DNA sequences may involve different 'alphabets'. For example, interaction of the protein groups with the chemical groups of the bases within the major groove leads to a conservative sequence of nucleotides. But in the case of interactions within the minor groove, the degeneracy of A–T or C–G may arise since positions of the chemical groups in the minor groove are practically unchanged when an A–T pair is replaced by T–A. We previously used this approach for the analysis of aligned sets of signal sequences and it allowed us to make some new conclusions about DNA–protein interactions (8). Secondly, we ascertain the significance of the patterns from a binomial distribution. Given an expected probability for the occurrence of any pattern, we identify those patterns for which the total number of observed occurrences have the lowest probability; i.e. are least likely to have occurred by chance. In general the calculation of the expected probability could be done in a variety of ways. For this paper we only use a simple model where all patterns are assumed to occur with a probability based on their composition, without regard to any dinucleotide, or higher order, biases. We also assume the statistical–mechanical, position-independent model of protein–DNA interaction (9–10) which has been shown experimentally to be reasonable, at least in some cases (12–13). Within the scope of this model, the energy of DNA–protein interactions is considered an additive function (i.e. every position within the interaction site contributes a local energy which are added together to get the total energy of interaction). This model is used in most of the ‘weight matrix’ methods for identifying binding sites in unaligned sequences (14,19,32,33).

**Multi-Alphabet Consensus Analysis (MACA)**

The method described here has some similarities to other algorithms by various authors (1,2,15). The main advantages of our approach are the use of degenerate nucleotide alphabets and determining statistical significance of the patterns that are found. A pattern is an L-letter word, containing at each position one letter of the alphabet. The general 10 letter alphabet A consists of four particular subalphabets as follows: nucleotide alphabet A, C, G, T and 2-fold degenerate alphabet A/G, C/T (purine, pyrimidine alphabet), A/T, C/G (degeneracy available due to minor groove interactions) and A/C, G/T (degeneracy available due to major groove interaction) (8). This uses all available 2-fold degeneracies. A 15 letter alphabet is also used that includes all possible degeneracies of the 4 nt.

The input data set contains nucleotide sequences not necessarily of the same length. Each is known to contain one or more binding sites of one protein. The main computation in the method can be described as follows. First we calculate the occurrences of the patterns composed of only nucleotides (i.e. A, C, G and T) using the fast 'suffix arrays' technique (16). Next, the occurrences of complex patterns are calculated by summing the corresponding occurrences of simple patterns. To estimate the significance of every pattern, we use a binomial probability distribution. Let $N_{se}$ be the number of sequences and $N_{tot}$ be the total number of nucleotides in the input data set. Then if the length of a pattern is $L$ the total number of all available positions of the pattern is calculated as:

$$N = N_{tot} - (L - 1) \times N_{se}$$

(1)

If the vector $f = \{f_1, f_2, \ldots, f_{10}\}$ is the set of expected frequencies of every letter of the alphabet $A$, then the expected frequency, $f_{pat}$, to find the given pattern assuming that all positions in the pattern are independent is calculated as $f_{pat} = \prod b_i$, where $b_i$ is the frequency of the correspondent letter $b_i \in A$, in the position $i$ of the pattern. Let us suggest that some pattern has been observed $K$ times in the data set. Thus, the probability to find $K$ or more occurrences of the pattern in the set is calculated by summation of the binomial probabilities:

$$P = \sum_{i=K}^{N} p_i = \sum_{i=K}^{N} C_N^i \times p^i \times (1 - p)^{N-i}$$

(2)

where $C_N^i$ is the binomial coefficient and $p$ is the probability of the given pattern. In our model we suggest that $p$ is equal to $f_{pat}$. It should be pointed out that more complicated models for $f_{pat}$ may be used. For example it could be calculated based on dinucleotide, or higher order frequencies, or even by Markov models. This will complicate its calculation but may give more reliable results in some cases. The important point is that however $f_{pat}$ is calculated we use that as the value of $p$ to be used in the determination of the binomial probability. For all of the examples in this paper the simple calculation of $f_{pat}$ describe above has been used.

For large $N$ we used a limit approximation of the probability of $K$ or more successes, each with success probability $p$, when specified fraction of success, $a = K/N$, satisfies $0 < p < a < 1$:
that we have used in our calculation is the same as elsewhere.

As soon as a multi-alphabet binding site pattern is chosen we also assume that the letter b has the highest score, would 'represent' position i in the alphabet. The main idea is if some of the patterns composed of only nucleotides (i.e. A, C, G and T) have significant occurrence then with high probability a 15 letter alphabet pattern created from a combination of them also would have significant occurrences.

Any 15 letter alphabet pattern can be represented as a logical combination of 10 letter patterns. For example, the combination of the two 10 letter patterns (A/T)CG and CG(A/T) produce the 15 letter pattern G(C/G)C (here G is A/C/T). We assume that the most significant 15 letter patterns should consist of the most significant 10 letter patterns. In this assumption the algorithm is composed of two steps. The first step is to find the N best 10 letter patterns e.g. patterns that have maximal values of the probability that some domain of insignificant 15 letter patterns. In this assumption the algorithm is independent (28).

Combinatorial 15 letter pattern generator

Using the methods that have been described in the previous section allows us to find the best patterns based on the 15 letter alphabet. The main idea is if some of the patterns composed of only nucleotides (i.e. A, C, G and T) have significant occurrence then with high probability a 15 letter alphabet pattern created from a combination of them also would have significant occurrences.

Any 15 letter alphabet pattern can be represented as a logical combination of 10 letter patterns. For example, the combination of the two 10 letter patterns (A/T)CG and CG(A/T) produce the 15 letter pattern G(C/G)C (here G is A/C/T). We assume that the most significant 15 letter patterns should consist of the most significant 10 letter patterns. In this assumption the algorithm is composed of two steps. The first step is to find the N best 10 letter patterns e.g. patterns that have maximal values of the probability that some domain of insignificant 15 letter patterns. In this assumption the algorithm is independent (28).

Multi-alphabet weight matrix representation of a pattern

As soon as a multi-alphabet binding site pattern is chosen we also can use a more general approach for representation of the binding site—e.g. weight matrix. A set of aligned binding sites that match to the pattern can be extracted from the original data set and a weight matrix may be calculated. The weight matrix algorithm that we have used in our calculation is the same as elsewhere [see for example (18)]. The multi-alphabet weight matrix used here also recognizes the idea that different types of protein-DNA interactions along DNA sequences may be represented by different 'alphabets'.

The weight matrix contains the values −Ln(Pb) of the bases at each position l, where Pb was established by formula:

\[ P_b = \sum_{i=1}^{N} p_{b_i} = \sum_{i=1}^{N} C_k \times f_b \times (1 - f_b)^{(N-k)} \]

where N and K have the same meaning as in formula (2).

In this notation f_b is the observed frequency of the letter b (b ∈ A). The subalphabet (i.e. a particular kind of the degeneracy) in which the letter b has the highest score, would 'represent' position i in the matrix. For example, if in a particular position the highest score is observed for the nucleotide A, than the scores in the matrix are calculated for every nucleotide (P_A, P_G, P_C and P_T), but if the occurrence of purine is most significant, only two scores P_AG and P_CT would be placed in the matrix: P_A = PA*GC_P_G = PC*G_P_C = P_CT.

The total score of a window in some position on a test sequence is calculated by summation of −Ln(Pb) along the window. This total score gives us an estimation of the probability of finding a particular pattern on the test sequence by chance.

CAP and lexA binding sites

The sequences of CAP-binding sites are the same as in (19). The data set consisted of 23 experimentally determined CAP-binding sites distributed over 18 fragments each 105 bp long. Each of the 18 fragments contains one or two CAP-binding sites. Recently, it was shown (11) that locus Eco MalK contains an additional CAP-binding site which was not included in the original data set (19,33). The sequences of lexA binding sites are the same as those in (14). This data set contains 11 DNA fragments, each 200 bases in length, with a total of 16 reported binding sites.

Nucleosomes sequences

The 177 nucleosome sequences isolated from chicken chromatin have been used to search for histone octamer patterns (20). The data set contains nucleotide sequences of DNA fragments ranging from 144 to 148 bp in length with a 145 bp long average. The sequences have been aligned about their centers as in the original data set. The sequences of the nucleosome core DNA molecules were kindly supplied by Dr A. A. Travers.

We also analyzed the positioning sequences for mapped nucleosomes (21). The data set is a heterogeneous compilation of the nucleosomal DNA sequences that were obtained from different organisms and by different experimental procedures (reconstruction and extraction). Using the reference table, we took nucleotide sequences from GenBank (Release 78.0 8/93) and created a nucleosomal DNA sequence database. The database consists of 23 sequences for which positions have been determined with accuracy less than ±6 bp. The data set has a fixed 145 bp size of DNA fragments.
experimentally identified CAP-binding sites have the highest by this matrix and the score values have been calculated. All gap-5bp. Every possible position in the data set has been tested a weight matrix based on the best pattern as described in the half site participates less in the CAP-DNA interaction (23). calculated patterns. This asymmetrical model predicts that the left are important for the protein-DNA interaction. The left half site, in six sites with a value of -ln(P) = 8.0 compared to -ln(P) = 14.8 for the number of sites described by T(G/T)T(A/G)(A/G)-gap-(A/G)(A/G)(A/G) has been some sequence. The strong signal on the edges of the nucleosomal sequence (A/G)(A/G)(A/G)-gap-(A/G)(A/G)(A/G) has been approximately symmetric peaks. The two most significant symmetric peaks are placed at the edges of the nucleosomal DNA (21). In order to accommodate the decrease of an already weak signal, arising from allowed shifts in the sequences, we have used the MACA approach in the analysis of the complex structure of the nucleosome sequences. Both the greedy approach (31) and the EM approach (32) have been modified to find multiple classes of pattern in a single data set. This can be very valuable for data where it is not known whether or not all of the fragments are bound by a single protein. Both the greedy approach (31) and the EM approach (32) have been modified to find multiple classes of binding sites within a single data set. One advantage of our approach is that no modification to the method is required, rather one just looks down the list of significant patterns to see if the entire data set can be best explained by there being more than one class of binding sites. Figure 1. The comparison of the best signals within 23 CAP-binding sites to the crystal structure (3). (a) Nucleotide sequence of the DNA fragment that was used in the crystal complex is shown. The two kinks are shown by double arrows, the nucleotides that form important hydrogen bond contacts with the protein are shown in bold. The 2-fold axis are shown with . Numbering of the base pairs is from the dyad axis. (b) The two best patterns (10 and 15 letter alphabets, respectively) obtained from the analysis of the 18 fragments are aligned to the sequence. The six dots represent the space region.

Symmetry of nucleosomal DNA sequences

There are two major problems with the analysis of nucleosomal DNA. First, the interaction between the nucleosome core and DNA has low specificity and can not produce a well defined signal on the DNA sequence. Secondly, this signal is decreased in the unaligned data set due to the experimental method of preparation of the nucleosome core. To obtain a core particle, an experimentalist uses a somewhat non-specific enzyme that cuts unprotected DNA molecules in the chromatin. Because of the dynamic and probabilistic character of such an enzymatic reaction, there is a shift ± 10 bp in the determination of the end of the nucleosomal DNA (21). In order to accommodate the decrease of an already weak signal, arising from allowed shifts in the sequences, we have used the MACA approach in the analysis of the complex structure of the nucleosome sequences. Another important feature is a symmetry of the interaction between the nucleosome hexamer and the DNA. We analyzed the set of 177 core DNA sequences isolated from chicken chromatin (20). The length of the DNA sequences varied from 142 to 146 bp. Previously, Fourier (20) and rotational analysis (24) have demonstrated the existence of a periodic signal on the nucleosomal sequence. In both cases the data sets were aligned about the center of every sequence.

To reveal symmetric signals in the nucleosomal DNA and to increase the sensitivity, we scanned along nucleotide sequences with windows of various sizes. In every position of the window we estimated a distribution of the 10 letter alphabet patterns. The length of the pattern is determined by the window size. The difference between the size of the window and the pattern length was twice that of the experimental error of determining the nucleosome position. Using our methods we have found some evidence for symmetrical feature on the DNA molecules. Figure 2a shows the logarithm of the probability of the most common patterns along the nucleotide sequences. The curve has a few approximately symmetric peaks. The two most significant symmetric peaks are placed at the edges of the nucleosomal DNA sequences and four helical turns from the center of the nucleosome sequence. The strong signal on the edges of the nucleosomal sequence (A/G)(A/G)(A/G)-gap-(A/G)(A/G)(A/G) has been...
To avoid ambiguity we also estimated the significance of the expected. The most random patterns are placed in the middle of the fragment (20). The results are in good agreement with the nucleosome crystal structure. The positions of the peaks correlated with the positions where the sharp bends of the DNA molecule wrap around the core particle (26,27) and where some sequence conservation is expected. The most random patterns are placed in the middle of the nucleosome.

As was shown by different authors (28,29) the variance of an overlapping pattern is dependent on the autocorrelation coefficient. To avoid ambiguity we also estimated the significance of the obtained pattern by the method described in (28). Our calculation demonstrates that the correction of the deviation for the overlapping pattern \((A/G)(A/G)(A/G)\sim n\sim (A/G)(A/G)(A/G)\) does not change it from being the most significant compared to the other patterns.

**Distribution of the phased patterns**

The main feature of the nucleosomal particle is a sharp bend of the DNA molecule. The DNA molecule is wrapped around a histone core as a 1.8-turn left-handed solenoid 110 Å in a diameter. The protein–DNA complex is a symmetric particle with a dyad axis through the middle of the DNA molecule. It was theoretically suggested (25) and experimentally shown on the artificial positioning sequences (6,7) that the wrapped DNA sequence has a phased pattern of \((A/T)^3\sim n\sim (C/G)^3\). In addition, periodic signals of dinucleotides were reported in the work of Sutchwell et al. (20).

The MACA method allows us to test the significance of different models of protein–DNA interaction based on various theoretical models. Travers and Klug (25) propose the phasing model of the nucleosomal hexamer on the DNA molecule, based on the dinucleotide occurrences and Fourier analysis of the nucleosome data set (20). The two most significant dinucleotides have been used in the model. AA/TT and GC dinucleotides have a significant anti-phased amplitudes of Fourier transformation. The authors proposed that the pattern \((AA/TT)n(GG/GC)\) can play a major role during the phasing of the nucleosome particle (25). Shrader and Crothers also showed that repeated \((A/T)^3n(C/G)^3\) motifs are most favorable for nucleosome formation (6).

To analyze phasing signals on the chicken core particles and compare them with the proposed patterns we created a ‘synchronized’ data set consisting of phased fragments from the original data set (20). We cut the 177 direct and 177 complementary sequences of DNA molecules into six 14 bp fragments. The six fragments contained positions 1-26, 23-36, 33-46, 44-57, 53-66 respectively and aligned all fragments together. The cut position was chosen such that the peaks of AA/TT and GC frequencies were placed in the middle of the fragment (20). The resulting data set contained 2124 14 bp-long fragments. The aim of this compilations is to increase the relative value of the phased signal on the nucleosome DNA. Because our method is not sensitive to the local shifts of DNA fragments, we can calculate the occurrences and estimate the probabilities of all the patterns even in the case when the original fragments are not aligned. To test the model of Travers and Klug we compared the occurrences of all available patterns that have motif XXX-gap-XXX or ZZ-gap-ZZ, where X is any letter from the alphabet A Z is one of the 10 dinucleotide steps as in (20), and gaps vary from 2 to 5 bp.

Table 1 shows the estimations of the occurrences of multi-alphabet and simple patterns. For both approaches the significance only of two the best patterns are compared with the proposed model. The analysis shows that both the simple alphabet and the multi-alphabet models reveal patterns that are different from the patterns \((AA/TT)X_{CA/TG}\) or \((A/T)^3X_{CA/TG}\). In the case of the simple alphabet model the two dinucleotides that compose the pattern (AA/TT)X_{CA/TG}, also have significant anti-phased amplitudes of Fourier transformation (results not shown) and the highest score is when the gaps equal 3 bp long. However, the other

---

**Figure 2.** Plots of scores, \(-\log(P)\) for highest values of the score as a function of window position. The window size is 12 bp. The search space have been restricted by patterns XXX\sim nXXX\sim n, where X, any letter from alphabet A and n, 2 bp gap. Two best symmetrical signals are shown on the diagrams. (a) The scores obtained from original 177 nucleosome sequences. (b) The scores when complimentary sequences have been added. We used standard nomenclature for 2-fold degeneration: R is \((A/G)\), Y is \((C/T)\) and W is \((A/T)\).
highest scoring pattern is (AA/TT)Xn(AA/TT) which is impossible to consider by Fourier analysis. Compared to those two patterns, the (AA/TT)Xn(CG) has a −Ln(P) of 30-fold less. The multi-alphabet analysis almost reveals the same motifs and the occurrence of the proposed (A/T) (A/T) Xn (C/G) (C/G) (C/G) model occurs even little less than the expected value.

Table 1. Distribution of the patterns in the multi-alphabet and simple approaches of the phased model of the nucleosome DNA

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Size of gap (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Simple model</td>
<td></td>
</tr>
<tr>
<td>(AA/TT)Xn(AA/TT)</td>
<td>32.05</td>
</tr>
<tr>
<td>(AA/TT)Xn(CA/TG)</td>
<td>29.03</td>
</tr>
<tr>
<td>(AA/TT)Xn(CG)</td>
<td>1.44</td>
</tr>
<tr>
<td>Multi-alphabet model</td>
<td></td>
</tr>
<tr>
<td>(CT/CT)Xn(AA/AA)(AA/AA)(AA/AA)</td>
<td>11.14</td>
</tr>
<tr>
<td>(AT/AT)Xn(AC/A)(AC/A)(AC/A)</td>
<td>7.54</td>
</tr>
<tr>
<td>(AT/AT)Xn(CG/C)(CG/C)(CG/C)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*The values of −Ln(P) are calculated for the two most significant patterns in both multi-alphabet and simple models. The third pattern was proposed to contain a nucleosome phasing motif by (25). The zero value of probabilities mean significance of rare occurrences of the pattern. In all calculations, the size of the window was 4 bp longer than the size of particular pattern (that provides a ±2 bp available shift) and observed frequencies of the dinucleotides have been used because they deviate from the value expected from equiprobable frequencies.

Mapped nucleosomal sequence database

The same type of analysis that was described in the previous section was conducted on the 23 sequences from the database that was located with experimental error less than ±6 bp (21). Figure 3 shows the score distribution of the best patterns along the DNA sequences. As in the prior analysis, the two significant peaks occur on each side of the dyad where 1.5 and 4 helical turn kinks are placed on the DNA molecule wrapped around histone octamer. The pattern (A/G) (A/G) (A/G)nn(A/G)(A/G)(A/G) is placed 15 bp from the dyad and the pattern (A/C)(A/C)(A/C)(A/C)(A/C)(A/C) is placed 40 bp from the dyad.

The analysis of phased patterns show almost the same distribution of phased patterns as for the chicken chromatin (result not shown). The most significant dinucleotide phased patterns are (AA/TT)Xn(AA/TT) and (AA/TT)Xn(CA/TG). Moreover the behavior of the probabilities of the gap size is exactly the same. For example, the minimum for (AA/TT)Xn(AA/TT) and the maximum for (AA/TT)Xn(CA/TG) of −Ln(p) for the two patterns have been observed for 3 bp gap size. As for the chicken chromatin the (AA/TT)Xn(CG) pattern is not represented enough to play the role of the phasing signal.

DISCUSSION

In this study we have developed a new approach that can be applied to the general problem of DNA–protein interaction. The ability of the MACA algorithm to analyze an unaligned data set without prior assumptions in the space of degenerate patterns gives us a new feature that is not available in previous approaches.

CAP-binding sites

To bind selectively to DNA the CAP protein should recognize a specific sequence of hydrogen-bonding donor–acceptor positions and steric peculiarities of the binding site. The most significant 10-letter alphabet consensus that was obtained in this work permits five ambiguities in the sequence, two of them among 4 bp that have the hydrogen bonds with the protein (3). The right half site has two conserved cytosines that form hydrogen-bonds with the protein. The left half site has the degeneracies of G/T and C/T in positions 7 and 5, respectively. In the first case, theoretically one hydrogen bridge is still available instead of the two that was supposed from crystallographic data. In the second case, the C→T mismatch has absolutely different recognition pattern of the donor-acceptor positions. Meanwhile the left half site has the other two conserved thymines that are, probably, also important for the specific protein–DNA recognition. In other words, the calculated consensus is degenerate enough to involve all known CAP binding sites and are not in conflict with the crystallographic data. It will
be interesting to model those two mismatches by molecular modeling and obtain their energy cost.

**Nucleosome sequences**

In this work we make a systematic search for the existence of a phasing signal on the nucleosome DNA. Both a phasing signal and linear symmetric patterns have been revealed. Using the method developed in this work, we have eliminated the problem of small errors in the alignment of the unshifted nucleosomal DNA data set. This allowed us to increase the sensitivity of the analysis and to detect two main symmetric signals.

The theoretical model (25) of the phasing signal was based on the Fourier analysis of dinucleotides distributions. The distribution of AA/TT and CG dinucleotides have the highest amplitudes and are in anti-phase. Meanwhile the absolute occurrence of CG is under-represented relative to the others dinucleotides. In spite of the attractiveness of the proposed model (25) the suggested phasing signal pattern (A/T)^3gap(C/G) occurs less often than expected. The CA/TG dinucleotide has the third highest value of the Fourier amplitude and also is in opposite phase with AA/TT distribution (result not shown). In contrast, the dinucleotide CA/TG occurs only a little less than AA/TT and our analysis has demonstrated a significant occurrence of the pattern (AA/TT)-gap-(CA/TG) with maximum score where gap size is 3 bp. The gap size of 3 bp orients the two dinucleotides pairs exactly on opposite sides of the DNA helix surface and allows the synchronization of the wrap peculiarities of the nucleosome DNA molecule. We also found that the (AA/TT)-gap-(AA/TT) pattern has a remarkably high occurrence but does not have a preferred gap length and probably reflects just the A/T-rich content of the nucleosomal DNA molecules.

The multi-alphabet analysis of the phased nucleosome data set reveals an agreement with the simple model. Again, the (A/T)^3gap-(C/G) has a significantly rare occurrence. In the multi-alphabet approach we have observed two outstanding patterns that have a maximum occurrence when the right and left patterns are placed on the opposite sides of the DNA. The (A/T)Xalpha(C/G)(A/T)Xgamma(T/G) is a generalization of the (AA/TT)Xalpha(CA/TG) pattern.

The two sets of the nucleosomal DNA sequences that were compared in this paper have different origins. The 177 nucleosome sequences were extracted from the single genome of chicken in a random manner while the other 23 sequences represent different organisms but all have been found to play a role in gene regulation and/or have strong positioning preference. It is intriguing that both sets have the same positional features on the DNA sequence although we found some different but significant preference in the patterns. This probably reflects the common structural requirement on the DNA organization around the nucleosome core. Both sets have a periodic distribution of the multi-alphabet pattern score among a nucleosomal sequence with a period of 10–11 bp.

**ACKNOWLEDGEMENTS**

We thank G. Hertz for several helpful discussions and comments on the manuscript. This work was supported by DOE grant ER61066 and an equipment grant from Sun Microsystems.

**REFERENCES**