Applicability of the multiple alignment algorithm for detection of weak patterns: periodically distributed DNA pattern as a study case

Alexander Bolshoy¹,²,³, Ilya Ioshikhes² and Edward N. Trifonov²

Abstract

Motivation: A nucleosome DNA positioning pattern is known to be one of the weakest (highly degenerated) patterns. The alignment procedure that has been developed recently for the extraction of such a pattern is based on a statistical matching of the sequences, and its success depends on the pattern/background ratio in the individual sequences and in the generated pattern. The heuristic nature of the method and distinctive properties of the pattern bring up the question of efficiency and sensitivity in the procedure. This paper presents a method of verification for this multiple sequence alignment algorithm.

Results: To verify the applicability of the multiple alignment approach, we constructed a set of sequences carrying the hidden pattern. The pattern was presented by weak ('signal') oscillations of occurrences of AA and TT dinucleotides along otherwise random sequences. Only a few dinucleotides of any given 145 base long sequence would correspond to the signal, appearing in about the same phase within the simulated periodic pattern. The novelty of our simulation approach is that we simulated a database as a whole, as opposed to simulating each sequence separately. The correlation between the hidden pattern and a sequence from the database is negligible on average, but our statistical multicycle alignment procedure produced the pattern with attributes very close to the simulated ones. The accuracy of the procedure was tested and calibrated. The presence in a typical sequence of as little as three dinucleotides corresponding to the signal is sufficient to generate (detect) the pattern hidden in a collection of 204 sequences.

Availability: The programs of the multiple sequence alignment algorithm and database simulation are available from the authors free of charge. Requests should be accompanied by a 3.5” diskette.

Contact: E-mail: bmbolsho@dapsas1.weizmann.ac.il

Introduction

Chromatin plays a very important role in the regulatory mechanisms of gene expression (reviewed in Lu et al., 1994; Wolffe, 1994a,b). In particular, the precise positioning of nucleosomes—the structural units of chromatin—along the DNA molecule, upstream or downstream of promoters may affect the initiation of transcription. The nucleosome positioning frequently, if not always, depends on a certain sequence pattern.

Methods of multiple alignment for pattern extraction from a set of biologically related sequences have been in use since approximately 1966 [see Waterman (1989) for a review]. Many different multiple alignment algorithms have been developed [see Chan et al. (1992) for a review], and new effective approaches have been published recently (Lawrence et al., 1993; Newald and Green, 1994). The nature of DNA–protein recognition in the nucleosome, however, differs so substantially from other studied DNA–protein interactions that the applicability of the mentioned methods, as well as newly developed ones, requires additional study.

The notion which appears to be broadly accepted is that the nucleosomal segment of the DNA double helix wrapped around the histone octamer should exhibit sequence-dependent anisotropic deformability (bendability) properties that would ensure a local free energy minimum for the nucleosome three-dimensional structure. The sequence-dependent DNA bendability is not uniform along the molecule and may inhibit the nucleosome centering at some sites, while favoring some others. The nucleosome positioning pattern in eukaryotic DNA appears to be one of the weakest known patterns, still not fully extracted. Various methods of multiple sequence alignment based on pairwise algorithms (Bacon and Anderson, 1986; Taylor, 1987; Roytberg, 1992) or local similarity (Altschul et al., 1990) are not applicable in the case of nucleosomal pattern because pairwise comparisons of nucleosomal sequences show neither high global similarity nor the presence of common ‘words’, which is typical in the alignment of protein sequences. Also, methods of dynamic programming for the alignment of two sequences are not beneficial in the case of the nucleosomal pattern because the physical nature of the problem assumes a pattern without gaps that could be, perhaps, expressed by

¹Department of Membranes Research and Biophysics, Genome Project and ²Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel
³To whom correspondence should be addressed
non-random distributions of dinucleotides along sequences. For every pair of sequences, a number of possible pairwise alignments is predefined by experimental errors of nucleosome center mapping. Since the length of the DNA wrapped around a histone octamer in a nucleosome is known, and equal to 145 nucleotides, the pattern could be presented by the matrix of dinucleotide frequencies \([16 \times 144]\). It has been found that some dinucleotides display weak periodicity in their distributions along the nucleosomal DNA, with the period \(\sim 10.1-10.5\) bases (Trifonov and Sussman, 1980; Satchwell et al., 1986; Ioshikhes et al., 1992, 1996; Bolshoy, 1995). In our earlier analysis of a collection of 118 nucleosome DNA sequences (Ioshikhes et al., 1992), the AA and TT dinucleotides have been found to oscillate with the period of \(\sim 10.4\) bases and with an approximate half-period phase shift between AA and TT. The recent analysis of a larger collection of 204 sequences (Ioshikhes et al., 1996) demonstrated a period of 10.3 bases and phase shift as above. In these two studies, an original statistical multiple alignment procedure has been developed and applied. The detailed description of this algorithm is included in the Methods and algorithms section.

In this iterative procedure, the sequences have been best matched consecutively by the AA and TT dinucleotides to the two-line matrix of AA and TT frequencies obtained for previous sequences. The procedure of best fitting a sequence to a matrix is rather simple since gaps and deletions are not allowed. The best match provides maximal correlation between the current matrix and the sequence. This iterative procedure is dependent upon the order in which the sequences are matched, therefore one cycle of the procedure cannot provide a final pattern. Thus, this procedure has been repeated many times, choosing different sequence orders. The resulting matrices have been averaged. It was assumed that the weak pattern present in the sequences would result in the correct match of at least some; being the only coherent pattern present in the sequences, it may show up after many alignment cycles and averaging of the output matrices.

The expectation was, indeed, well met and the pattern revealed, with its basic features consistent with earlier studies. In Figure 1 (taken from Ioshikhes et al., 1996), the distribution of AA dinucleotides obtained by this procedure is presented. Spectral analysis shows that the distribution of AAs can be presented by the sum of two components: random oscillations around the average content level of 21 AAs on 204 sequences per position and modulated periodic oscillations with a period of 10.15 bases, phase of about three bases and amplitude of 0.019 (corresponds to an oscillation with a magnitude of \(\sim 4\) AAs: \(204 \times 0.019 = 4\)). In one of our recent studies (Ioshikhes et al., 1996), we applied five different procedures of multiple alignment to the nucleosome database. All 10 distributions (five for AAs and five for TTs) possess quite similar quasi-periodic features, though with somewhat different periods and other distribution parameters. The heuristic nature of the methods and distinctive properties of the pattern gives rise questions of the efficiency and robustness of procedures. How does one exactly evaluate the proportion of the signal dinucleotides (the dinucleotides correlating well with the pattern) in a typical nucleosome DNA sequence (signal-to-noise ratio) while it is so difficult to estimate the quality of the extracted pattern? With such a weak signal, the calibration of the alignment procedure is necessary to determine its efficiency (how much of the signal is extracted) and sensitivity (the minimal detectable amount of signal elements). In the present work, the statistical multicycle consecutive multiple alignment technique is calibrated on a simulated system of random sequences in which a weak periodic signal of specified magnitude is introduced. A periodic variation of occurrences of the dinucleotides AA and TT with the period 10.33 bases and approximately a half-period relative phase shift is taken as a signal component imitating the nucleosome DNA pattern.

**Methods and algorithms**

**The pattern presentation**

There are different methods of pattern description with different levels of generalization (Waterman, 1989; Krogh et al., 1994; Bucher et al., 1996). To describe the DNA nucleosomal pattern, we could use the Hidden Markov Model (Krogh et al., 1994) or profile (Waterman and Perlwitz, 1986; Gribskov et al., 1990; Bucher et al., 1996). However, since the physical nature of the problem does not allow for gaps or deletions in pattern matching, all presentations become equivalent and can simply be presented by a matrix of distributions (Brendel and Trifonov, 1984; Stormo, 1990). The simplest nearest-neighbor model of DNA bendability (sequence-dependent anisotropic deformability) assumes that a pattern may be expressed in terms of dinucleotide distribution along the sequence. The length of the nucleosome core is known to be 145 ± 2 bases, a fragment of DNA with this given length is wrapped around the complex of proteins called histones. For each of the 144 dinucleotide positions along the DNA molecule, there is a probability distribution over the 16 dinucleotides occurring in that position in a typical nucleosome site. Since we are dealing with a pattern expressed only by distributions of the dinucleotides AA and TT, the pattern can be presented by a matrix \([2 \times 144]\) of AA and TT frequencies. We assume that a dinucleotide distribution can be simulated by the sum of two components: random background (noise) and periodic oscillations. Thus, the simulated signal was chosen as:

\[
F_d(i) = B_d + A_d \times \cos(2\pi \times (\Omega_d + i)/p) \tag{1}
\]

where \(F_d(i)\) is the frequency of dinucleotide \(d\) (\(d = AA, AC, \ldots\)) in position \(i\), \(B\) is a background, \(A\) is an amplitude, \(\Omega\)
Fig. 1. Authentic and simulated AA dinucleotide distributions along DNA. (A) AA pattern revealed by the multiple alignment from the set of 204 sequences of the actual nucleosome DNA sequence database (taken from Ioshikhes et al., 1996). (B) Input AA simulated distribution: sinusoidal pattern with an amplitude of 0.029 (see Methods and algorithms) introduced to the set of 204 random sequences. (C) Output AA pattern revealed by the multiple alignment from the simulated database.
is a phase of a periodic component and $p$ is a period of it. We performed simulations with varying amplitudes, and constant values of $\Omega_{AA}$, $\Omega_{TT}$ and $p$. The period $p$ was chosen to be 10.33 bases, the phases for AA and TT were taken to be equal to 2.30 and 7.46, respectively, with the difference 5.16 bases, about a half-period; all values are similar to those presented in Ioshikhes et al. (1996). Background components $B_{AA}$ and $B_{TT}$ were simulated on the basis of the uniform distribution of all nucleotides. One realization of such hidden simulated AA distribution with the amplitude 0.029 is presented in Figure 1B. The spectral analysis of the aligned output patterns (Figure 2) was performed essentially by Fourier formalism. A virtually continuous spectrum was calculated rather than the Fourier series (Kolker and Trifonov, in preparation).

The simulated sequence database construction

The set of sequences was constructed to imitate as closely as possible the experimental nucleosome DNA sequences. Ideally, an experimental technique of nucleosome mapping should precisely indicate the nucleosome center that, taking 72 bases downstream and 72 bases upstream, produces exactly 145 nucleotides of DNA nucleosomal site. However, only a few of the nucleosome DNA sequences available from literature have been mapped with a high degree of accuracy with regard to the nucleosome center (±1 base, three possible positions of the center). For others, the uncertainty of mapping is higher, up to 55 nucleotides (111 possible positions of nucleosome DNA sequence midpoint). The latest database (Ioshikhes and Trifonov, 1993, 1994) was supplemented by new entries, making a total of 204 sequences, sorted in order of descending accuracy of their experimental mapping (Ioshikhes et al., 1996, supplementary material). We designed several sets, each with the same amount of simulated sequences (204) with the same distribution of experimental errors. Each set was produced in two steps. First, all the sequences were generated randomly with the uniform distribution of four bases A, C, G, T. The multiple alignment algorithm was applied to these random sets to test the method, to avoid possible artifacts of the technique, to ensure that the distribution of dinucleotides is indeed random and can suit background from equation (1). Following this, sinusoidal signal distributions of AA and TT were introduced to the random set in the next way. If the simulated signal for AA distribution assumes the frequency of $k$ dinucleotides over background in position $i$, the procedure randomly chooses $k$ sequences out of the total 204 and places nucleotides A in positions $i$ and $i + 1$. Similarly, for negative values of $k$ (under background) the AA dinucleotides in $k$ sequences are changed to non-AA ones. Modified sequences were placed back to the simulated database with randomly chosen shifts according to error limits indicated in the experimental database. Global amounts of such changes, both positive and negative, correspond to the amplitude of a signal from equation (1). For instance, to realize the signal with the amplitude equal to 0.02 (oscillation of ±4 bp), about 760 changes of dinucleotides (both AAs and TTs) in 204 database sequences should be done. This means that an average of 3.6 dinucleotides of 145 bases were changed per sequence (or to approximate less accurately, every sequence has 3–4 signal dinucleotides on average).

The multiple alignment algorithm

In our analysis, each $K$th sequence of the database has been represented by a $[2 \times L]$ matrix $M_K$, where $L + 1$ is the length of the sequence. The first row of the matrix corresponds to AA dinucleotides, the second row to TT dinucleotides. The rows elements are equal to 1 or 0, depending on the presence (absence) of the dinucleotide in the given position. The alignment procedure does not allow for gaps or deletions, it allows shifts of assumed nucleosome core midpoint within bounds of claimed accuracy $a_K$ of the $K$th sequence. To find the final alignment in this case means to generate a vector of individual shifts that would correspond to the best alignment of all the sequences according to a certain criterion. The exhaustive search for such a vector (comparing all possible alignments) is not feasible, and one has to develop an algorithm in its initial form was described earlier (Ioshikhes et al., 1992). The hidden criterion of the best alignment for this algorithm is a maximization of the sum of scores of pairwise alignments among all database sequences. For the score, measure of the pairwise alignment was chosen as a criterion of the maximal likelihood. It cannot be rigorously proven that the multicycle consecutive alignment algorithm provides the optimal or even suboptimal alignment according to this criterion, the method is absolutely heuristic; however, outcome of its application demonstrates that it works.

The main basis of the algorithm is that the DNA sequences represented by AA and TT dinucleotide matrices are aligned one at a time with the pattern derived from alignments at earlier steps. In other terms, at a step $K$ the shift of the $K$th nucleosomal sequence is determined so as to maximize a sum of scores of pairwise alignments of the $K$th sequence with $K - 1$ sequences located at previous steps. We can suppose intuitively that an obtained alignment is a suboptimal one for the aligned $K$ sequences. This is a general description, a more detailed one is given below.
The sequences are preliminarily sorted in ascending order of claimed experimental errors \( (a_k = \pm 1, \pm 2, \pm 3, \ldots, \pm 50 \text{bp}) \) and clustered into several groups uniting nucleosomes characterized by the same value of \( a_k \). Then the sequences are aligned one at a time by the following scheme.

First, the initial pattern \( P_1 \) \([2 \times 144]\) matrix is formed for the first sequence in the database, one with an accuracy \( a_1 = \pm 1 \), thus \( P_1 = M_1 \).

At every step \( K (K = 2, \ldots, 204) \), the \( K \)th sequence is shifted to the position \( s_K \) which provides maximal correlation between the current pattern matrix \( P_{K-1} \) and the \([2 \times 144]\) matrix \( M_K \) of the sequence \( K \), as much as indicated error bars allow. As the correlation measure, the product \( P_{K-1}^T \times M_K \) of the transposed \((^T)\) current pattern matrix \( P_{K-1} \) and the current sequence matrix \( M_K \) has been considered.

A new pattern matrix \( P_K \) is obtained by combining the aligned matrices: \( P_K = P_{K-1} + M_K \). In this work (like in Ioshikhes et al., 1996), we used the version of the algorithm that utilizes only oscillating components of the AA and TT positional frequency distributions: \( M_K = M_K - <M_K> \), where \( <M_K> \) is the mathematical expectation of elements of the AA (TT) strings of \( M_K \).

The pattern derived by one full cycle of such an alignment procedure depends on the order in which the sequences are taken. For this reason, the order of the sequences in every group of accuracy has been changed randomly and the results of 10,000 different alignment cycles were averaged.

**Hardware and software**

The sequence alignment algorithm and simulating programs were written using Turbo PASCAL (Version 6.0) and implemented on IBM PC compatible computers. The programs are available from the authors free of charge. Requests should be accompanied by a 3.5″ diskette.

**Results and discussion**

As described in Methods and algorithms, all intermediate and final pattern distributions have been obtained for the AA dinucleotide distributions as well as for the TT dinucleotide distributions. We illustrate our results with the figures presenting only AA distributions, TT distributions are essentially consistent.

Figure 1A is taken from Ioshikhes et al. (1996) and presents the distribution of AA dinucleotides along the natural nucleosome sequences, calculated by the multicycle consecutive alignment of 204 sequences compiled from the literature. We can easily see how weak the signal is: the maximal presence of the AA dinucleotide in a certain position is only \(~40\) AAs, meaning that only \(20\%\) of all sequences match the pattern in this position. The distribution of TT dinucleotides (not shown) is essentially symmetrical to the AA distribution with the center of symmetry at the midpoint of the nucleosome DNA sequence. The final AA pattern appears as a rather regular oscillation with the period 10.3 bases. That is, the strictly periodic (sineoidal) oscillation would be a good first approximation of the pattern. A signal of this form was introduced in the model set of random sequences, as described in Methods and algorithms, so that each sequence differs from a randomly generated prototype by only a few AA and/or TT dinucleotides introduced at some of their preferred periodical signal positions. The introduced pattern [see equation (1)] with an amplitude equal to 0.029 is shown in Figure 1B. Comparing Figure 1A and B, we can see that the model is a simplification of a real case: the background level of noise in the model (Figure 1B) corresponds to the uniform composition of dinucleotides and is equal to 1/16, whereas the background level of AAs in eukaryotic DNA (Figure 1A) is substantially higher (1/10) because of a higher A + T content and over-representation of runs of adenines. The other difference is that in the simulated pattern, all 14 peaks are present, while in the extracted nucleosomal pattern (Figure 1A) some peaks are missing. Despite these differences, overall periodic characteristics of these distributions are similar and, thus, the sinusoidal model could be a good first approximation. In Figure 1C, the extracted periodic AA pattern is shown. We can compare the introduced (hidden) pattern (Figure 1B) and the pattern obtained by the multiple alignment procedure. Positions of major peaks and a general character of the extracted pattern correlate well between input and output distributions. The spectral analysis was applied to clarify this strong correlation.

The amplitude spectrum of the pattern from Figure 1C is shown in Figure 2 (continuous line A). It has a major maximum at 10.35 bases, very close to the period of the simulated signal introduced into the sequences. The amplitude of the major peak of the extracted pattern is equal to 0.027, while the amplitude of the introduced signal is equal to 0.029. The signal with such amplitude corresponds to 5–6 AA and/or TT signal dinucleotides per sequence on average. For comparison, a typical amplitude spectrum is shown for the output generated by the alignment procedure from one of the sets of completely random sequences (dotted line D). The amplitude of the signal peak (0.027) is seven times larger than the average noise amplitude calculated for several independent sets of random sequences (0.004, data not shown). This means that a substantially weaker signal could still be extracted by the procedure. Two more simulations were performed with the input amplitudes equal to 0.022 and 0.014 (dashed lines B and C, respectively). The maximal value of B line is equal to 0.020 and corresponds to the input of 0.022. The procedure is able to detect a signal with as low an amplitude as 0.014 (the output is equal to 0.010), although the period of 10.3 does not provide the major peak—there is a larger peak in the neighborhood of a period with a value of about 13.7. Of course, this is an artifact: a small peak in the
The amplitude spectra of the extracted patterns. All databases have the same size of 204 nucleotide sequences. (A) From the simulated database with an introduced sinusoidal pattern, amplitude 0.029. (B) From the simulated database with an introduced sinusoidal pattern, amplitude 0.022. (C) From the simulated database with an introduced sinusoidal pattern, amplitude 0.014. (D) From the database of uniformly random sequences, a typical example.

The signal is not fully extracted, however, since the amplitude of the output pattern is smaller than in the sequence (an amplitude = 0.017).

The signal is not fully extracted, however, since the amplitude of the output pattern is smaller than in the simulated periodic input. The efficiency of signal extraction is 93% for a signal with an amplitude about 0.03 and 71% for a signal with an amplitude close to 0.015. As one would expect, a larger signal is extracted more efficiently.

A nucleosome DNA sequence database (or any other set of collected DNA sequence fragments mapped with experimental errors) should be characterized by its size and distribution of experimental errors in mapping of nucleosome sites, the sensitivity of the statistical multiple alignment technique is about three AA and/or TT dinucleotides per 145 base sequence (an amplitude ≈ 0.017).

1. Ioshikhes, I. (1992) Preferred positions of AA and TT dinucleotides in natural nucleosome DNA correlating to the nucleosome pattern: 4–6 dinucleotides. Being scattered in 2 x 14 alternative preferred positions, these few dinucleotides are clearly not recognizable when two typical sequences are compared. One may have a proper match if some sequences in the collection actually have a larger number of the signal dinucleotides as fluctuation from the average. Owing to these sequences, the statistical alignment procedure is apparently also able to align poorer sequences and extract the signal, weak as it is.

Conclusions
The goal of this study was to check the sensitivity of the statistical multicycle consecutive multiple alignment technique by applying it to model systems. A periodic variation of occurrences of the dinucleotides AA and TT with the period 10.33 bases and about a half-period phase shift between them was taken as a signal component imitating the nucleosome DNA pattern.

The analysis above verifies the applicability of the statistical multiple alignment technique to signals as weak as the nucleosome DNA pattern. Taking into account the efficiency of the extraction, one can estimate the average number of AA and TT dinucleotides in natural nucleosome DNA correlating to the nucleosomal pattern: 4–6 dinucleotides. Being scattered in 2 x 14 alternative preferred positions, these few dinucleotides are clearly not recognizable when two typical sequences are compared. One may have a proper match if some sequences in the collection actually have a larger number of the signal dinucleotides as fluctuation from the average. Owing to these sequences, the statistical alignment procedure is apparently also able to align poorer sequences and extract the signal, weak as it is.

Acknowledgements
The authors are thankful to E.Kolker for kindly providing the original program of spectral analysis. A.B. is supported by the 'B. de Rothschild Fund for the Advancement of Science in Israel' and the National Laboratory for Bioinformatics and DNA Sequencing of the Israel Council for Higher Education. I.I. is supported by an L.Bein WIS scholarship.

References
Ioshikhes, I. and Trifonov, E.N. (1994) Nucleosome DNA sequence database. EMBL deposit on a FTP server FTP.EBI.AC.UK in a directory 'nucleosome dna'.


