Visible periodicity of strong nucleosome DNA sequences

Bilal Saliha, Vijay Tripathi and Edward N. Trifonov

aGenome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel; bDepartment of Computer Science, University of Haifa, Mount Carmel, Haifa 31905, Israel

Communicated by Ramaswamy H. Sarma

Fifteen years ago, Lowary and Widom assembled nucleosomes on synthetic random sequence DNA molecules, selected the strongest nucleosomes and discovered that the TA dinucleotides in these strong nucleosome sequences often appear at 10–11 bases from one another or at distances which are multiples of this period. We repeated this experiment computationally, on large ensembles of natural genomic sequences, by selecting the strongest nucleosomes – i.e. those with such distances between like-named dinucleotides, multiples of 10.4 bases, the structural and sequence period of nucleosome DNA. The analysis confirmed the periodicity of TA dinucleotides in the strong nucleosomes, and revealed as well other periodic sequence elements, notably classical AA and TT dinucleotides. The matrices of DNA bendability and their simple linear forms – nucleosome positioning motifs – are calculated from the strong nucleosome DNA sequences. The motifs are in full accord with nucleosome positioning sequences derived earlier, thus confirming that the new technique, indeed, detects strong nucleosomes. Species- and isochore-specific variations of the matrices and of the positioning motifs are demonstrated. The strong nucleosome DNA sequences manifest the highest hitherto nucleosome positioning sequence signals, showing the dinucleotide periodicities in directly observable rather than in hidden form.

Keywords: nucleosome positioning; sequence periodicity; nucleosome positioning motif; bendability matrix; chromatin code; isochore

1. Introduction

The well-documented characteristic of nucleosome DNA is its hidden sequence periodicity (Trifonov & Sussman, 1980, Trifonov, 2011), with the period 10.4 bases, which corresponds to the structural period of DNA in the nucleosome (see Results for details). The signal is very weak, and various signal-processing procedures have to be applied to the sequences to extract the periodical sequence pattern(s). These are positional correlation analysis (Bettecken & Trifonov, 2009; Saliha, Saliha, & Trifonov, 2008; Trifonov & Sussman, 1980), synchronous detection (Gabdank, Barash, & Trifonov, 2009; Mengeritsky & Trifonov, 1983), multiple alignment (Ioshikhes, Bolshoy, Derenshteyn, Borodovsky, & Trifonov, 1996), regeneration of signal from its parts (Gabdank et al., 2009) and Shannon N-gram extension (Rapoport, Frenkel, & Trifonov, 2011).

Here, we suggest a novel approach – computational extraction of the positioning pattern(s) from strong nucleosome DNA sequences in which the periodicity would be obvious, rather than hidden. The strong nucleosomes are understood here as those which are formed on the strongly periodic sequences. One experimental example is the nucleosome-forming sequence ‘601’ selected from random pool, with periodically positioned TA dinucleotides (Lowary & Widom, 1998; Vasudevan, Chua, & Davey, 2010).

As a simple sensitive measure of the sequence periodicity in this work, we use occurrence of various dinucleotides at distances which are multiples of the nucleosome DNA period. In the ideally periodical sequence of the nucleosome DNA, every dinucleotide is separated from all other like-named dinucleotides by the distances $10.4 \times n$ bases. Such sequence, therefore, would show in the positional autocorrelation functions, the maximal possible count of such distances. Not exactly ideal strong nucleosome DNA sequences would also have high counts. The search for the strong nucleosome DNA sequences in the genomes is, thus, reduced to calculation of such counts for every fragment of length 115 bases (Trifonov, 2011) in the genome sequences, and thereby selection of the winners.

Various estimates of the nucleosome DNA period converge to the values in the interval 10.36–10.40 base pairs (Bettecken & Trifonov, 2009; Cohanim, Kashi, & Trifonov, 2006; Prunell et al., 1979; Trifonov & Bettecken, 1979; Ulanovsky & Trifonov, 1983; Winter, Song, Mukherjee, Furey, & Crawford, 2013). We, thus, have chosen the period 10.4 to derive the multiples. The closest integers to the multiples are 10 bases (10.4),
21(20.8), 31(31.2), 42(41.6), 52(52), 62(62.4), 73(72.8), 83(83.2), 94(93.6), 104(104) and 114 bases (114.4), within the span of 115 bases involved in contact with the histone octamer (Trifonov, 2011). It is worth to note that, strictly speaking, frequent occurrence of the distances may not necessarily mean the sequence periodicity. However, since the ‘magic’ distances are characteristic of the nucleosome DNA, and no other sequence signals are known, unrelated to the nucleosomes, which would display such distances, there is a good chance that this simple technique would be able to, indeed, extract the strong nucleosome DNA sequences. Importantly, since such procedure counts the magic distances for all 16 dinucleotides, there is no a priori bias in the choice of the most significant ones, so that none of the signal components would be missed.

Using the counts of the $10.4^n$ bases distances between the same dinucleotides, for all 16 of them, as measure of relevant sequence periodicity, we scanned various genomic sequences in a search for those sequence fragments of length 115 bases which manifest the highest scores. These would represent the strongest nucleosomes. As the data described below demonstrate, the identified strong sequences of Arabidopsis thaliana, Caenorhabditis elegans and Homo sapiens show all major nucleosome DNA sequence characteristics revealed in earlier work (Frenkel, Trifonov, Volkovich, & Bettecken, 2011; Gabdank et al., 2009; Rapoport et al., 2011), and, thus, the calculated strong nucleosome DNA patterns are, indeed, representative of strong nucleosomes of the respective species. Utilization of only strong sequences for derivation of the nucleosome sequence patterns significantly reduces the noise level, thus, making the derived patterns more reliable. The extracted in this work, periodical strong nucleosome DNA sequences for the first time demonstrate visible, rather than hidden nucleosome DNA sequence periodicity.

2. Materials and methods

2.1. DNA sequences

In this study, we used three eukaryotic genomes. Genome sequences were downloaded from the following sources: A. thaliana (ftp.ncbi.nih.gov/genomes/Arabidopsis_thaliana), C. elegans ((ftp.ncbi.nih.gov/genomes/Caenorhabditis_elegans) and H. sapiens ((ftp.ncbi.nih.gov/genomes/H_sapiens).

The strong nucleosome SELEX sequences are taken from (Lowary & Widom, 1998).

2.2. Isochores

The isochores of a given genome are extracted by running over its DNA sequence with a non-overlapping sliding window of length 10,000 bases, and then classifying each DNA fragment to the appropriate isochore type depending on the G + C content as follows: L1 (G + C < 37%), L2 (37% ≤ G + C < 41%), H1 (41% ≤ G + C < 46%), H2 (46% ≤ G + C < 53%) and H3 (53% ≤ G + C) (Costantini, Clay, Auletta, & Bernardi, 2006).

2.3. Extraction of strong nucleosome sequences

Given a (long) DNA sequence, we screen it with an overlapping sliding window of length 115 bases (sliding with 10 base steps, instead of 1, to boost the calculation). For each fragment (skipping fragments that contain simple tandem repeats, like … ATATAT …) we calculate the autocorrelation function in the domain $[1…115]$, and consider the sum of the function’s values at the magic distances $10.4^n$ (i.e. 10, 21, 31, 42, …, 114) as the score of this specific DNA fragment. Finally, we rank all the fragments according to their scores and choose all fragments above some cut-off threshold (see Section 2.4 for the determination of the threshold). In case of overlapping fragments, the strongest are taken. The fragments selected by the procedure are called ‘strong nucleosome DNA sequences’.

2.4. Determination of strong nucleosome’s cut-off threshold

Using random sequences, appropriately generated, one can evaluate the score cut-off threshold. In this context, the null hypothesis $H_0$ would be that ‘Random sequences of base composition similar to those of the DNA sequence in question do not contain strong nucleosomes’. We use, therefore, the following algorithm: (1) generate many random sequences (say 1000 sequences of 1 million bases each) according to some base composition distribution, (2) for each sequence, independently, apply the algorithm described in previous section to calculate highest scoring fragment and (3) choose the maximum score of the highest scoring fragments over all sequences to be the cut-off threshold.

The range for the estimated thresholds is between ~90 for H3 isochore of H. sapiens and ~120 for A. thaliana (with significance level .001). With these thresholds all accepted sequences display clear visible sequence periodicity.

2.5. Derivation of optimal and near optimal linear motifs from nucleosome DNA bendability matrix

Bendability matrix ($16 \times 10$ elements) contains the information that describes deformational preferences of 16 dinucleotides in various positions along one period (10 bases) of the nucleosome DNA sequence. However, it is convenient also to look at a simpler representation
(while losing some information) – a linear sequence motif which fits best to the matrix. Reading the best motif from the matrix requires solving the optimization problem of finding the legal path with the largest sum of weights along that path.

2.5.1. Definitions

Path is a list of 10 dinucleotides chosen from the matrix, one entry per column.

Legal path is a path for which every two adjacent entries (dinucleotides) overlap (e.g. ‘AT’ and ‘TG’).

Finding the highest scoring legal path is equivalent to finding the dominant (optimal) linear motif corresponding to a given matrix of bendability. This is solved by simple Dynamic Programming (DP) (Dreyfus & Law, 1976). The same holds for close-to optimal motifs problem which is solved by the modified version of DP suggested in (Waterman & Byers, 1985).

3. Results

3.1. The choice of the period

Various estimates of the nucleosome DNA period converge to the values in the interval 10.36–10.40 base pairs (Bettecken & Trifonov, 2009; Cohanim et al., 2006; Gale, Nissen, & Smerdon, 1987; Ioshikhes et al., 1996; Prunell et al., 1979; Trifonov & Bettecken, 1979; Ulanovsky & Trifonov, 1983 and Winter et al., 2013), though one early estimate also suggests interval 10.1–10.3 (Satchwell, Drew, & Travers, 1986) which does not overlap with the above. The most reliable of these estimates are those which are obtained on the nucleosome material, rather than on the basis of a whole genome or overall chromatin. These are the estimates derived from nucleosome DNA databases: chicken nucleosomes, 10.10–10.30 bases (Satchwell et al., 1986), mixed nucleosome collection, 10.10–10.50 bases (Ioshikhes et al., 1996) and, more accurate ones, calculated from geometry of DNA path in the nucleosome, 10.30–10.50 (Ulanovsky & Trifonov, 1983), from data on UV dimer formation on the nucleosome, 10.20–10.40 (Gale et al., 1987), from available nucleosome crystal data, 10.36–10.44 (Cohanim et al., 2006) and most recent, estimated as average distance between DNaseI cuts on the nucleosomes, 10.30–10.40 bases (Winter et al., 2013). These on-the-nucleosome estimates also converge to the interval 10.36–10.40.

To verify, additionally, the validity of the chosen period, we conducted the same calculations as described below, using magic distances for other periods, from 6 to 15 bases, and deriving the corresponding sequence motifs. The linear motifs obtained for various periods are, essentially, shorter and longer versions of the standard (see below) 10-base long motif AAAAAATTTTT for the period 10.4 bases. For example, for the periods 10, 10.1, 10.2, 10.3 and 10.5 the resulting motif is the same AAAAAATTTTT and for the period 11 bases, the resulting motif is AAAAGGGTTTT. The occurrence of the 11 bases motif is substantially lower than one of A5T5.

3.2. Strong nucleosome-forming DNA sequences extracted from random mixture (Lowary & Widom, 1998)

The SELEX collection of the clones corresponding to strong nucleosomes contains both Watson strands of the clones and some Crick strands. We have taken only one strand of each nucleosome DNA of the collection, ending with 54 non-redundant sequences of lengths 66–100 bases.

As the distance analysis demonstrates, the major heterodinucleotide contributors to the 10.4 bases distances in the strong nucleosome DNA subset of SELEX sequences are dinucleotides TA (Table 1, first column), which confirms original observation (Lowary & Widom, 1998). Following the technique of reconstruction of the nucleosome DNA bendability matrix from its incomplete forms (Gabdank et al., 2009), we combined all occurrences of TA at distance 10 bases (total 48 TANNNNNNTA sequence fragments within the 54 sequences of the dataset) and derived the matrix of bendability (Figure 1), where the observed frequencies of various dinucleotides in various positions (phases) of the nucleosome DNA period are shown. Taking frequent homodinucleotides AA or TT for similar analysis (i.e. AANNNNNNAA and TTTNNNNNNTT fragments) would smear the final pattern, as these dinucleotides appear often as runs AAA … and TTT … As this precious ensemble of strong nucleosome DNA sequences is of rather small size, the statistics of the occurrences of the dinucleotides in the matrix is poor. Nevertheless, two continuous dominant (consensus) sequences – TAGAGTGGCTTA (sum of scores 154) and TAGAGGCCTCTTA (sum of scores 151) – are generated by DP (Waterman & Byers, 1985) that selects from all possible continuous sequences the ones with the highest sums of scores. Almost the same result can be obtained manually. Indeed, the dinucleotides TA (column 0), AG (highest score in column 1), GA (column 2) and AG (column 3) make continuous linear motif TAGA. Similarly, by fusing the neighboring elements in two last columns, one derives the motif CTA, resulting in TAGAGNNNCTTA. Manual reconstruction of the full-length continuous motif meets uncertainties, so that the reconstruction has to be made by optimization via DP which completes the partial motif above: TAGAGGCCTCTTA, the second highest scoring pattern derived by the DP. This pattern can be considered as consensus sequence of the whole collection of the one period long TANNNNNNNTA fragments.
Table 1. Contributions of various dinucleotides to the 10.4-n distances in various sets of strong nucleosome sequences. The absolute contributions are ranked, to reveal the most prominent heterodinucleotide contributors in each case (in bold). Lists of the strong nucleosome DNA sequences contain 54 sequences (Lowary & Widom, 1998), 864 (A. thaliana), 1630 (C. elegans), 4193 (H. sapiens, chromosomes 1–5), 470 (isochromes H2, H. sapiens, chromosomes 1–6), and 610 (isochromes H3, H. sapiens, chromosomes 1–20). The list of 864 A. thaliana strong nucleosome DNA sequences is provided in the Supplement.

|                  | SELEX | A. thaliana | C. elegans | H. sapiens (chrom. 1–5) | H. sapiens H2 isochromes (chrom. 1–6) | H. sapiens H3 isochromes (chrom. 1–20) |
|------------------|-------|-------------|------------|------------------------|---------------------------------------|---------------------------------------|
| Dinu. Cont.      |      |             |            |                        |                                       |                                       |
| TA 255           | TT 56,734 | AA 111,801 | TT 137,837 | AA 8598                | GG 18,979                             |                                       |
| AG 246           | AA 55,947 | TT 102,186 | AA 125,022 | TT 7467                | CC 17,018                             |                                       |
| GC 232           | TA 6649 | AT 10,591   | AT 17,496  | GG 5902                | GC 2742                               |                                       |
| CT 230           | AT 4508 | GA 6082     | TA 15,229  | CC 5831                | TG 2337                               |                                       |
| CG 202           | AG 3132 | TC 5906     | TG 10,214  | CT 1528                | AG 2206                               |                                       |
| GG 186           | CT 2977 | CA 4706     | GG 10,314  | CT 1615                | CA 2289                               |                                       |
| CC 178           | GA 1765 | TG 4648     | TG 10,214  | CT 1528                | AG 2206                               |                                       |
| GA 151           | TC 1669 | TA 3169     | CT 10,056  | TG 1502                | TT 1941                               |                                       |
| TT 137           | GG 1501 | CC 2933     | CA 9754    | GA 1318                | AA 1801                               |                                       |
| AA 132           | GT 1367 | GG 2546     | AG 9714    | AT 1260                | TC 1747                               |                                       |
| AC 129           | CC 1361 | CT 2493     | GC 8349    | TC 1237                | GA 1586                               |                                       |
| TC 120           | AC 1321 | AG 2424     | GA 7671    | GC 1081                | GT 1293                               |                                       |
| GT 110           | TG 1095 | GT 1855     | GT 6661    | AC 1048                | AC 1124                               |                                       |
| TG 97            | CA 1062 | AC 1847     | AC 6441    | TA 1027                | CG 1036                               |                                       |
| CA 95            | GC 117  | GC 1763     | GC 2612    | GT 767                 | AT 480                                |                                       |
| AT 47            | CG 73   | CG 1426     | CG 692     | CG 396                 | TA 257                                |                                       |

Figure 1. Matrix of bendability, calculated (this work) for SELEX collection of DNA clones forming strong nucleosomes (Lowary & Widom, 1998). The strongest elements in the columns contributing to dominant continuous motif TAGAGGCCTCTA (see text) are in bold.

The complementary TAG and CTA are the most prominent elements of the above-pattern, as it follows from the matrix. This has been also noted in (Lowary & Widom, 1998), where the periodical nucleosome positioning motif CTAGNNNNNNCTAGNNNNNN ... was suggested. Remarkably, the self-complementary sequence TAGAGGCCTCTA not only agrees with the earlier results – alternation of runs of purines and pyrimidines (YRRRRRRYYYYY) in the nucleosome DNA (Mengeritsky & Trifonov, 1983; Rapoport et al., 2011; Salih et al., 2008; Trifonov, 2010), but also strictly conforms to the complementary symmetry of the bendability pattern, which follows from DNA duplex symmetry (Trifonov, 2010).

3.3. Strong nucleosome sequences of A. thaliana

In the Figure 2, a randomly chosen subset of strong nucleosome DNA sequences extracted from A. thaliana genome is shown. Simple manual alignment reveals somewhat distorted, but clearly periodic character of the sequences. The periodically repeated runs of adenines are highlighted (in red), as well as the periodically reappearing TA dinucleotides, in the same sequences (in green). Note that the periodicity as such was not searched for by the procedure, which only counted the occurrences of 'magic' distances of 10.4-n bases. The procedure, therefore, attests to the fact that the magic distances appear to be delivered by, indeed, 10.4-n periodic DNA sequences characteristic of the nucleosome DNA.

The Figure 2 presents the nucleosome DNA sequence periodicity in the most obvious way. Even the strongest hitherto known periodical nucleosome DNA sequence, the clone 601 (Lowary & Widom, 1998), does not show periodicity clearly. Majority of the natural nucleosomes are positioned on the sequences with no
visible period, and special means are required to detect the periodicity, as illustrated well by the whole history of the studies on the nucleosome DNA periodicity (Trifonov, 2011), starting with the positional autocorrelation analyses (Trifonov & Sussman, 1980). Community of chromatin researchers is generally not familiar with the weak signals buried in noise. They may accept the existence of the periodicity but remain skeptical. The clearly periodic nature of the strong nucleosome sequences, as in the Figure 2, perhaps, would be more convincing than the whole arsenal of indirect demonstrations of the periodicity.

It is clear, however, that even the strong sequences are not ideal repetitions of whatever pattern, i.e. there are no ideal strong nucleosome sequences in the genome of A. thaliana, nor in other genomes. Such ideal nucleosome sequences, the strongest ones, if at all existing, would be a very rare exception, as they, perhaps, will create severe obstacles for transcription and replication. The obvious similarity of the sequences above, however, is suggestive of existence of some ideal (consensus) pattern with repeating runs of A and of T. Mere inspection of the sequences suggests that the most prominent frequently occurring sequence motif is TTTAAA ...

Since heterodinucleotide TA is the major contributor to the magic distances in A. thaliana (Table 1), we

Figure 2. Subset of strong nucleosome DNA sequences of A. thaliana. The sequences are aligned to illustrate the almost uninterrupted though imperfect periodicity of AA dinucleotides (top, in red, bold), and of TA dinucleotides (bottom, in green, bold).

Figure 3. Matrix of bendability, calculated for collection of strong nucleosome sequences extracted from A. thaliana (complete genome). The strongest element in the columns contributing to dominant continuous motif TAAAAATTTTTA (see text) are in bold.
A. thaliana (Figure 3). The strongest elements in the left part of the matrix (columns 1–4) are AA dinucleotides (TAAAAA) while on the right half (columns 7–10), these are rather TT dinucleotides (TTTTTA). This suggests the linear form of the matrix (consensus sequence): TAAAAATTTTTA. The highest scoring pattern derived by DP procedure applied to the matrix is, indeed, TAAAAATTTTTA. The motifs with A4 or A6 instead of A5 are the next highest scoring sequences.

The dominant pattern, TAAAAATTTTTA is a good match to the motifs obtained earlier for A. thaliana genome and other sequences (Rapoport et al., 2011), by other approaches.

3.4. Matrix of bendability for genome of C. elegans

In the Figure S1 (supplement), the matrix is shown, based on the starting pattern ATNNNNNNNNAT, since the AT heterodinucleotide is the major contributor to the observed 10.4n distances in C. elegans genome (Table 1). The linear motif – one complete 10-base period – derived from the elements of the matrix, the same way as in case of A. thaliana, is ATTTTAAAAAT, or in AT-central form TAAAAATTTTTA, i.e., identical to the strongest pattern in A. thaliana. Interestingly, the next strongest motifs, according to DP analysis of the C. elegans matrix are very similar to the dominant one: ATTTTAAAAAT, ATTTAAAAAAAT, ATTTTTGAAAAT, ATTTTTTAAAAT, ATTTTGAACACT, ATTTTCTAAAAAT and ATTTTTCGAAAAAT.

Comparison of the dominant pattern TAAAAATTTTTA in (RY-central form) with the one calculated earlier by a different technique (Gabdank et al., 2009), CGRAATTTTCYG, shows 80% match (per 10 bases). The difference, CG instead of TA, can be attributed to different ways of evaluation of the ‘strengths’ of various dinucleotides contributing to the patterns. In the regeneration of signal approach (ibid) as the measure of strength, the selectivity of a given dinucleotide to the best position within the period was chosen. In that work, the dinucleotides CG and AT appeared as the strongest. In the current work, rather the absolute contribution of the dinucleotides to the 10.4n distances is selected. Only eventual rigorous energy calculations of the importance of various dinucleotides for DNA bending in the nucleosomes may provide more appropriate measures. In the mean time the most frequently used (consensus) pattern, i.e., TAAAAATTTTTA in case of C. elegans, seems to be a right choice, though the CGRAATTTTCYG pattern can be read from the matrix (Figure S1) as well. CGRAATTTTCYG is the eighth strongest motif (see above) calculated by DP from the C. elegans matrix (Figure S1). Both the TA and CG dinucleotides show highest preference to the central YR position of the matrix. That is, both C. elegans patterns, as above, can be used for nucleosome mapping, and they both satisfy the most general pattern YRRRRRYYYYYR (Mengeritsky & Trifonov, 1983; Rapoport et al., 2011; Salih et al., 2008; Trifonov, 2010).

To quantitatively characterize contributions of all dinucleotides to the DNA bendability in the nucleosome, the matrix description is more appropriate, since the columns of the matrix contain other elements also, in addition to the dominant ones. Those which appear in the same column have the same rotational setting in the DNA molecule on the surface of the histone octamer. The simple linear presentations of the matrices of DNA bendability, the consensus motifs are convenient for the purpose of quick comparisons.

3.5. H. sapiens

For derivation of the matrix of bendability, sequences of chromosomes one to five are taken. 4193 strong nucleosome sequences are extracted. The heterodinucleotide that is found to contribute most to the magic distances is AT (Table 1). After collecting all occurrences of ATNNNNNNNNAT in the strong nucleosome DNA sequences of H. sapiens, the matrix of bendability is derived as shown in Figure S2. The strongest elements of the matrix (TT in positions 1–4, and AA in columns 6–9) make the highest score one-period motif, already familiar, ATTTTAAAAAT, or in AT-central form: TAAAAATTTTTA. The DP generates the same pattern. It coincides with the motif derived for H. sapiens earlier (Frenkel, Trifonov, et al., 2011; Rapoport et al., 2011). The next strongest motifs, by DP, contain runs of T and A of lengths 3–7 (7–3) instead of T5 and A5.

3.6. Isochores

The motifs described above, all TAAAAATTTTTA for three different genomes, essentially, represent the prevailing A + T rich isochores L1 and L2 of these genomes (Cammarano, Costantini, & Bernardi, 2009; Costantini et al., 2006; Zhang & Zhang, 2004). Other isochores are expected to generate different patterns, with dinucleotides containing G and C. However, the bendability matrix calculated for the isochores H1 boils down to the same dominant TAAAAATTTTTA pattern as for isochores L1 and L2 (not shown). This is due to higher than 50% A + T composition of the isochores L1–H1 (all < 46% G + C). The matrix for strong nucleosome sequences of isochores H2 (46–53% G + C) is shown in the Figure S3. The linear pattern derived from this matrix is different. The major dinucleotide contributor in this case is CA, and the DP calculation generates from the matrix in Figure S3, the predominant motif CAAAACCCCTCA showing the alternation of runs of purines and pyrimidines. The standard R5Y5 variant
from the above C-rich motifs, apparently, re-
symmetrical pattern ATTTTTAAAAAT. The difference
taken, AT in this case (Table 1), the DP generates also
cal. If, however, the strongest symmetrical contributor is
in this case, since the CA element itself is not symmetri-
cal. Its advantage is in its universality, as all the
alphabet. Its advantage is in its universality, as all the
motif-forming CA and AT dinucleotides.

In the Figure 4, the matrix for strong nucleosome
sequences of isochores H3 (> 53% G + C) is shown.
The linear patterns derived from this matrix are substan-
tially different: GCCCCCGGGGGC, and, of nearly the
same strength, GCCCCCGGGGGC, the latter showing
also complementary symmetry, in addition to the
standard alternation of runs of R and Y.

4. Discussion

4.1. Nucleosome positioning sequences, various levels
of their description

The most appropriate description of the positioning
pattern is listing which dinucleotide elements have
physical (deformational) preference to which positions
along the period of the DNA in the nucleosome, and
how strong (selective) the preference is. This is called
the matrix of DNA bendability (Gabdank et al., 2009;
Mengeritsky & Trifonov, 1983; Trifonov, 1980). Currently
the ultimate final matrix of bendability, general
for all cases, is unavailable. Rather there are several
more or less similar matrices, different for different
species and for DNA with different nucleotide composi-
tion (Frenkel, Bettecken, & Trifonov, 2011; this work).
The matrices described in this paper are typical exam-
pl es. The matrix of nucleosome DNA bendability can be
expressed as frequencies of occurrences of various dinu-
cleotides, as in the examples, or as matrix of energy
parameters and values of deformational preferences of
the dinucleotides. This last form of the bendability
matrix at the moment can only be imagined, as the pref-
ferences are not evaluated yet to build a set of 160 values
(16 dinucleotide elements, 10 positions along the period).
Rigorous estimation of the values is rather problematic
(e.g. Cui & Zhurkin, 2010; Tolstorukov, Colasanti,
McCandlish, Olson, & Zhurkin, 2007).

A simplified presentation of the matrix of bendability
is the continuous sequence motif, 10 bases long, derived
by connecting the topmost dinucleotides of each column
– the consensus motif (one-period-long consensus or
dominant nucleosome positioning sequence). In the
current and in earlier papers these motifs are given rather
as 12-base sequences, with identical dinucleotides (e.g.
YR) at both ends of the period.

As the Figure 2 illustrates, the consensus nucleo-
some positioning pattern, TAAAAATTTTTA in case of
A. thaliana, is not readily seen as such even in the
strongest nucleosome DNA sequences. It is compro-
mised almost in every period by the bases which do
not belong to the consensus – an obvious noise in the
pattern. In typical nucleosome DNA with only marginal
periodicity, the noise, actually, dominates. However, the
signal in these sequences is present – the sequence ele-
ments of the consensus, appearing here and there in
phase with the hidden repeat, collectively result in eas-
ier deformation in certain direction and in the recogni-
tion of the DNA by the histone octamers. In other
words, one does not have to have an exact match of
the sequence to the ideal repetition of the consensus.
Even few matching signal dinucleotides may be
sufficient for the accurate recognition, i.e. nucleosome
positioning. Now, with the consensus patterns available,
the critical amount of the signal elements can be
experimentally determined.

Yet, even simpler presentation of the matrix of
bendability is linear form YRRRRYYYYYR, in binary
alphabet. Its advantage is in its universality, as all the
consensus linear patterns described above match the
form. It can be applied to any sequence for preliminary
mapping of the nucleosomes. More accurate evaluation
of the nucleosome strengths can be achieved by taking
into account the dinucleotide composition, i.e. using the
full matrix of bendability for the given species or
composition.

|     | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 0   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| AA  | 0   | 0   | 8   | 4   | 4   | 11  | 11  | 13  | 8   | 0   | 0   |
| AC  | 0   | 0   | 21  | 17  | 23  | 11  | 11  | 7   | 7   | 0   | 0   |
| AG  | 0   | 0   | 31  | 23  | 23  | 31  | 32  | 37  | 34  | 60  | 0   |
| AT  | 0   | 0   | 9   | 12  | 7   | 6   | 9   | 4   | 8   | 0   | 0   |
| CA  | 0   | 69  | 34  | 22  | 19  | 23  | 16  | 19  | 25  | 0   | 0   |
| CC  | 0   | 185 | 88  | 62  | 55  | 66  | 79  | 71  | 28  | 0   | 0   |
| CG  | 0   | 44  | 23  | 20  | 21  | 16  | 23  | 12  | 27  | 49  | 0   |
| CT  | 0   | 71  | 40  | 33  | 29  | 17  | 18  | 22  | 36  | 0   | 0   |
| GA  | 0   | 0   | 8   | 21  | 28  | 25  | 21  | 24  | 24  | 0   | 0   |
| GC  | 369 | 0   | 5   | 24  | 23  | 35  | 17  | 25  | 5   | 0   | 360 |
| GG  | 0   | 0   | 20  | 46  | 60  | 51  | 69  | 81  | 104 | 191 | 0   |
| GT  | 0   | 0   | 11  | 15  | 19  | 11  | 11  | 16  | 0   | 0   | 0   |
| TA  | 0   | 0   | 6   | 10  | 8   | 4   | 3   | 3   | 3   | 0   | 0   |
| TC  | 0   | 0   | 23  | 21  | 21  | 24  | 17  | 13  | 7   | 0   | 0   |
| TG  | 0   | 0   | 32  | 32  | 26  | 20  | 24  | 29  | 29  | 69  | 0   |
| TT  | 0   | 0   | 10  | 7   | 12  | 10  | 8   | 1   | 9   | 0   | 0   |

Figure 4. Matrix of bendability, calculated for collection of strong nucleosomes extracted from H3 isochores, H. sapiens. The strongest elements in the columns contributing to dominant continuous motif GCCCCCGGGGGC (see text) are in bold.
4.2. Optimal positions of various dinucleotides within nucleosome DNA period

The dinucleotide stacks YR•YR are located at the local axes of the minor grooves oriented toward the histone octamer (Cui & Zhurkin, 2010; Vasudevan et al., 2010) – orientation ‘IN’. The stacks RY•RY, half-period away, are oriented ‘OUT’. Accordingly, all YR dinucleotides (TA, CG, TG and CA) should concentrate in central columns of RYNNNNNNNRY matrices, which is the case, with only occasional deviations due to statistical reasons (see Figures S1 and S2). Similarly, in YRNNNNNNNYR matrices the dinucleotides AT, GC, AC and GT are found preferably in central or close-to-central positions (Figure 3). All the RR dinucleotides concentrate in one half-period, while all YY dinucleotides are found in another half-period of the positioning pattern. This has a general explanation based on the asymmetry of deformation of the RR•YY stacks (Trifonov, 2010). Some of the particular RR and YY dinucleotides show more specific positional preferences within the period, as e.g. in case of strong SELEX nucleosomes (Figure 1). Interestingly, all the matrices of bendability demonstrate substantially stronger elements close to the pattern-forming ‘dominant’ dinucleotides, with rather uncertain preferences around the central positions. This, probably, reflects the avoidance of having excessively stable nucleosomes, with all key elements of the matrix equally present.

4.3. Towards high resolution mapping of the nucleosomes on any sequence of interest

The simplest and most general nucleosome positioning periodical pattern would be (RRRRRRYYYYY)\(n\), with occasional extension of the 10 base sequence repeat to 11 bases, to satisfy the observed average period 10.4 bases of DNA in the nucleosome, as in (Gabdank, Barash, & Trifonov, 2010a, 2010b). More accurate mapping would be based on the periodically repeated matrices of bendability, specific for given sequence type (species or isochores). One obvious way to derive such full length matrices is alignment of the strong nucleosome DNA sequences and derivation of its 16 × 114 elements (values), corresponding to respective orientations of 16 dinucleotide stacks in the 115 base long nucleosome DNA (work in progress). In the mean time one can use instead the matrix for C. elegans (Gabdank et al., 2009) the first of the kind, already generating single-base resolution maps (http://www.cs.bgu.ac.il/~nucleon) to study orientation of important sites in DNA on the surface of the histone octamers (Hapala & Trifonov, 2011, 2013; Trifonov & Hapala, 2013).

5. Conclusions

Computational extraction of strong potential nucleosome DNA sequences from genomes demonstrates for the first time that the strong nucleosome sequences possess obvious periodicity rather than mere hidden periodicity as indirectly demonstrated in earlier studies. Species-specific matrices of bendability are derived, with respective dominant nucleosome positioning sequence motifs. The universal nucleosome positioning motif (RRRRRRYYYYY)\(n\) is confirmed.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/19443994.2013.855143.

Funding

This work has been supported in part by fellowship of SoMoPro (South Moravian Program, Czech Republic) with financial contribution of the European Union within the seventh framework program (FP/2007–2013, grant agreement No. 229603), and by the project ‘CEITEC – Central European Institute of Technology’ (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund. V. T. is recipient of the fellowship of Indian-Israeli Government Scholarships 2012–2013.

References

Bettecken, T., & Trifonov, E. N. (2009). Repertoires of the nucleosome-positioning dinucleotides. PLoS ONE, 4, e7654.
Cammarano, R., Costantini, M., & Bernardi, G. (2009). The isochore patterns of invertebrate genomes. BMC Genomics, 10, 538.
Cohanim, A. B., Kashi, Y., & Trifonov, E. N. (2006). Three sequence rules for chromatin. Journal of Biomolecular Structure and Dynamics, 23, 559–566.
Costantini, M., Clay, O., Auletta, F., & Bernardi, G. (2006). An isochore map of human chromosomes. Genome Research, 16, 536–541.
Cui, F., & Zhurkin, V. B. (2010). Structure-based analysis of DNA sequence patterns guiding nucleosome positioning in vitro. Journal of Biomolecular Structure and Dynamics, 27, 821–841.
Dreyfus, S., & Law, A. (1976). The art and theory of dynamic programming. New York, NY: Academic Press.
Frenkel, Z. M., Bettecken, T., & Trifonov, E. N. (2011). Nucleosome DNA sequence structure of isochores. BMC Genomics, 12, 203.
Frenkel, Z. M., Trifonov, E. N., Volkovich, Z., & Bettecken, T. (2011). Nucleosome positioning patterns derived from human apoptotic nucleosomes. Journal of Biomolecular Structure and Dynamics, 29, 577–583.
Gabdank, I., Barash, D., & Trifonov, E. N. (2009). Nucleosome DNA bendability matrix (C. elegans). Journal of Biomolecular Structure and Dynamics, 26, 403–411.
Gabdank, I., Barash, D., & Trifonov, E. N. (2010a). Single-base resolution nucleosome mapping on DNA sequences. *Journal of Biomolecular Structure and Dynamics*, 28, 107–121.

Gabdank, I., Barash, D., & Trifonov, E. N. (2010b). Fine STR: A web server for single-base-resolution nucleosome positioning. *Bioinformatics*, 26, 845–846.

Gale, J. M., Nissen, K. A., & Smerdon, M. J. (1987). UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. *Proceedings of the National Academy of Sciences of the United States of America*, 84, 6644–6648.

Hapala, J., & Trifonov, E. N. (2011). High resolution positioning of intron ends on the nucleosomes. *Gene*, 489, 6–10.

Hapala, J., & Trifonov, E. N. (2013). Nucleosomal TATA-switch: Competing orientations of TATA on the nucleosome. *Gene*, 527, 339–343.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., & Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology*, 262, 129–139.

Lowary, P. T., & Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology*, 276, 19–42.

Mengeritsky, G., & Trifonov, E. N. (1983). Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Research*, 11, 3833–3851.

Prunell, A., Kornberg, R., Lutter, L., Klug, A., Levitt, M., & Crick, F. (1979). Periodicity of deoxyribonuclease I digestion of chromatin. *Science*, 204, 855–858.

Rapoport, A. E., Frenkel, Z. M., & Trifonov, E. N. (2011). Nucleosome positioning pattern derived from oligonucleotide compositions of genomic sequences. *Journal of Biomolecular Structure and Dynamics*, 28, 567–574.

Salih, F., Salih, B., & Trifonov, E. N. (2008). Sequence structure of hidden 10.4-base repeat in the nucleosomes of *C. elegans*. *Journal of Biomolecular Structure and Dynamics*, 26, 273–282.

Satchwell, S. C., Drew, H. R., & Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology*, 191, 659–675.

Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K., & Zhurkin, V. B. (2007). A novel roll-and-slide mechanism of DNA folding in chromatin: Implications for nucleosome positioning. *Journal of Molecular Biology*, 371, 725–738.

Trifonov, E. N. (1980). Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Research*, 8, 4041–4053.

Trifonov, E. N. (2010). Base pair stacking in nucleosome DNA and bendability sequence pattern. *Journal of Theoretical Biology*, 263, 337–339.

Trifonov, E. N. (2011). Cracking the chromatin code: Precise rule of nucleosome positioning. *Physics of Life Reviews*, 8, 39–50.

Trifonov, E. N., & Bettecken, T. (1979). Noninteger pitch and nuclease sensitivity of chromatin DNA. *Biochemistry*, 18, 454–456.

Trifonov, E. N., & Hapala, J. (2013). Single-base resolution sequence-directed nucleosome mapping. *Israel Journal of Chemistry*, 53, 144–146.

Trifonov, E. N., & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 77, 3816–3820.

Ulanovsky, L. E., & Trifonov, E. N. (1983). Superhelicity of nucleosomal DNA changes its double-helical repeat. *Cell Biophysics*, 5, 281–283.

Vasudevan, D., Chua, E. Y., & Davey, C. A. (2010). Crystal structures of nucleosome core particles containing the ‘601’ strong positioning sequence. *Journal of Molecular Biology*, 403, 1–10.

Waterman, M. S., & byers, T. H. (1985). A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Journal of Mathematical BioSciences*, 77, 179–188.

Winter, D. R., Song, L., Mukherjee, S., Firey, T. S., & Crawford, G. E. (2013). DNase-seq predicts regions of rotational nucleosome stability across diverse human cell types. *Genome Research*, 23, 1118–1129.

Zhang, C. T., & Zhang, R. (2004). Isochore structures in the genome of the plant *Arabidopsis thaliana*. *Journal of Molecular Evolution*, 59, 227–238.