Universal full-length nucleosome mapping sequence probe

Vijay Tripathia*, Bilal Salihb, and Edward N. Trifonova

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

(Received 1 December 2013; accepted 2 February 2014)

For the computational sequence-directed mapping of the nucleosomes, the knowledge of the nucleosome positioning motifs – 10–11 base long sequences – and respective matrices of bendability, is not sufficient, since there is no justified way to fuse these motifs in one continuous nucleosome DNA sequence. Discovery of the strong nucleosome (SN) DNA sequences, with visible sequence periodicity allows derivation of the full-length nucleosome DNA bendability pattern as matrix or consensus sequence. The SN sequences of three species (A. thaliana, C. elegans, and H. sapiens) are aligned (512 sequences for each species), and long (115 dinucleotides) matrices of bendability derived for the species. The matrices have strong common property – alternation of runs of purine–purine (RR) and pyrimidine–pyrimidine (YY) dinucleotides, with average period 10.4 bases. On this basis the universal [R,Y] consensus of the nucleosome DNA sequence is derived, with exactly defined positions of respective penta- and hexamers RRRRR, RRRRRR, YYYYY, and YYYYYY.

Keywords: nucleosome positioning; nucleosome probe; strong nucleosome; matrix of bendability; RR/YY probe

1. Introduction

Deformational properties of DNA are reflected in periodical appearance of various dinucleotide elements along chromatin DNA (Trifonov & Sussman, 1980), reflecting the same periodicity within individual nucleosomes. Thirty-three years of studies on the periodical DNA sequence pattern, also called nucleosome positioning pattern (reviewed in Trifonov, 2011) converged to a universal RRRRRYYYYY sequence (Frenkel, Bettecken, & Trifonov, 2011; Mengeritsky & Trifonov, 1983; Salih, Salih, & Trifonov, 2008; Rapoport, Frenkel, & Trifonov, 2011; Salih, Tripathi, & Trifonov, in press). Imperfect repeats of the sequence appear in the nucleosome DNA with the non-integer period 10.4 bases (reviewed in Salih et al., in press) so that actual distances between like-named dinucleotides in adjacent periods are either 10 or 11 bases. A consensus sequence of the nucleosome DNA (R_{5,6}Y_{5,6})_{11}, where 11 stands for the number of DNA periods in contact with the histone octamer (Trifonov, 2011), should consist, therefore, of intermingled periods of lengths 10 and 11 bases, making together the average period 10.4 bases. The same holds for the matrix of nucleosome DNA bendability (Gabdank, Barash, & Trifonov, 2010a, 2010b; Mengeritsky & Trifonov, 1983) where the occurrences of all 16 dinucleotides are taken into account, each preferentially located at their respective positions within the 10.4 base period. In the study of Gabdank et al. (2010a, 2010b) a full-length matrix is derived from nucleosome DNA sequences of C. elegans, the first of the kind. It was preceded, though, by an incomplete AA/TT only matrix (Ioshikhes, Bolshoy, Derenshteyn, Borodovsky, & Trifonov, 1996). In the Gabdank’s matrix the 10 and 11 base sections are intermingled, essentially, in arbitrary, though symmetrical order. The exact distribution of the dinucleotides and exact consensus sequence of the nucleosome DNA can be only derived by alignment of sufficiently large number of full-length nucleosome DNA sequences. This task is fulfilled in the current study, where the strong nucleosome (SN) DNA sequences, those with visible sequence periodicity (Salih et al., in press), are used for the derivation of the full-length nucleosome positioning probe. The idea on possible existence of the strongly periodical natural sequences followed from an elegant earlier study of Lowary and Widom (1998) in which it was experimentally demonstrated that random DNA sequences with highest affinity to histone octamers are characterized by 10–11 base sequence periodicity. They suggested the sequence (CTAGxxxxxx)n, or (AGxxxxxxCT)n, as possible nucleosome positioning motif. In (Salih et al., in press) this motif was extended to (AGAGGCCTCT)n, on the basis of original data in (Lowary & Widom, 1998). Notably, in [R,Y] alphabet the motif is (RRRRYYYYY)n. In this work the SN sequences from A. thaliana, C. elegans, and H. sapiens (total 1536 sequences), are aligned, resulting in the consensus:

*Corresponding author. Email: vtripath@campus.haifa.ac.il

© 2014 Taylor & Francis
The nucleosomes can be detected in any long sequence of interest by simple matching of the sequence to this standard.

2. Results and discussion

2.1. Full-length matrices of bendability derived from SN sequences

Alignment (see Methods) of 512 SNs of *A. thaliana* and of the same number of SN sequences of *C. elegans* and *H. sapiens*, chromosomes 1–5 (Salih et al., in press) resulted in very similar distributions. One of them, for *A. thaliana*, is shown in Figure 1, where the graphs (lines of the matrix) are sorted by the amplitude of 10.4 base oscillations. The matrices for two other species are deposited in the Supplement (Figures S1(a) and S1(b)). (The matrices in numerical form can be provided by request)

The full-length matrix of the nucleosome DNA bendability in Figure 1 is presented in complementarily symmetrized form, which results from summing together the 5′→ 3′ Watson strands and 5′→ 3′ Crick strands, since both represent the nucleosome which harbors the respective nucleosome DNA. Such symmetry is also required by the dyad symmetry of DNA (Mengeritsky & Trifonov, 1983; Trifonov, 2010). The component Watson strands and Crick strands matrices have only minute differences (not shown). The AA and TT dinucleotides are the main contributors of the matrix. Reading only topmost matrix elements consecutively results in the following linear form of the matrix (consensus sequence for the full-length 116 base nucleosome probe for *A. thaliana*):

```
TAAAAATTTTTAAAAATTTTTTTTTTTTTAAAA
TTTTAAAAATTTTTAAAAATTTTTTTAAAA
TTTTAAAAATTTTTAAAAATTTTTTTTTTTAA
```

Inspection of the matrix reveals that every 10–11 base period is essentially the same, with the same dinucleotide distributions. The periods of the probe above are identical as well except for two extra A (in 5th and 10th periods) and two extra T bases (in 2nd and 7th periods). This near identity of the periods is not a trivial result, since some positions within the nucleosome DNA could systematically deviate from the repeating AAAAAATTTTT consensus.

2.2. Matrices of bendability in [R,Y] alphabet

Although somewhat different in details, the matrices for three species have common dominant oscillation of AA and TT dinucleotides with the consensus period AAAAAATTTTT confirming earlier results for the same species, obtained from respective genomic sequences by Shannon N-gram extension (Rapoport et al., 2011). As the typical matrix shown in Figure 1 demonstrates, the RR dinucleotides AG, GA, and GG display oscillations nearly in phase with AA, while CT, TC, and CC oscillate in phase with TT. Such behavior has been observed first in (Mengeritsky & Trifonov, 1983), suggesting the universal pattern (RRRRYYYYYY)ₙ.

The same three full-length matrices, but in [R,Y] alphabet, are shown in Figure 2. They clearly demonstrate oscillations of RR and YY dinucleotides as the main feature common for all three species. That is, the consensus (RRRRYYYYYY)ₙ is confirmed.

2.3. Linear [R,Y] nucleosome mapping probe

In order to derive more exact linear form, reflecting the average 10.4- base rather than 10- base periodicity, we combined the plots for three species in the Figure 3(A). The plot for difference RR–YY (Figure 3(B)) allows reading the exact consensus. Indeed, while positive peaks correspond to runs of R and negative ones – to runs of Y, the crossings with the x-axis give the sequence coordinates of the transition dinucleotides RY and YR. These are -52, -42, -31, -21, -10, 0, 10, 21, 31, 42, and 52 for RR, and -47, -36, -26, -16, -5, 5, 16, 26, 36, and 47 for YR. The exact positions for YR dinucleotides are confirmed by the maxima in the YR plot (Figure 3(A)), while the maxima in more noisy plot for RY also fit fairly well to the calculated positions of RR. Remarkably, the YR and RR coordinates are all nearest integers to 10.4*n and 10.4*n + 5.2 values: 5 (5.2), 10 (10.4), 16 (15.6), 21 (20.8), 26 (26.0), 31 (31.2), 36 (36.4), 42 (41.6), 47 (46.8), and 52 (52.0). The resulting linear [R,Y] consensus form of the nucleosome DNA bendability matrix is, thus,

```
YRRRRRRYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
```

where the RY (bold) corresponds to the central dinucleotide (coordinate 0), while the YR dinucleotides at the ends correspond to the first and last contact points of the nucleosome DNA with the histone octamer. The second and seventh periods contain hexamers YYYYYY rather than pentamers. The hexamers RRRRRR are located in symmetrical to YYYYYY positions, within the fifth and tenth periods. The four extra nucleotides introduce the correction to the average period of the sequence: 114/11 = 10. (36), instead of 10, if all runs of R and Y would be of the same length 5. For the period 10.4, the positions of the outermost YR would be at 57.2 bases from the nucleosome dyad. The difference,.2 bases (6.8º), is of no significance, since the histone octamers and DNA can be mutually adjusted so that local DNA
Figure 1. The full-length matrix of nucleosome DNA bendability, 16 dinucleotides in 115 sequence positions, for SNs of *A. thaliana*.
periods may be anything between 10 and 11 bases, that is by as much as one base off (33°–36°) (Ong, Richmond, & Davey, 2007). It is also understood that small changes in the [R,Y] consensus, like change of R5Y6–R6Y5 or Y5R6–Y6R5 here and there, would lead to only minor change in the match of a given sequence to the probe.

2.4. Advantages of using the [R,Y] consensus probe for nucleosome mapping

The full-length matrix of nucleosome DNA bendability may take several forms. First, the ideal matrix would list energy values for all 16 dinucleotides in all positions along the nucleosome DNA. Such matrix, unfortunately, will become available, perhaps, in decades from now. No reliable deformation energy parameters are known today even for a single specific base-pair stack. The dinucleotide frequency full-length matrices, as the ones in this work, and earlier versions (Gabdank et al., 2010a, 2010b; Ioshikhes et al., 1996) are useful approximations, since these take into account the species-specific dinucleotide frequencies. The mapping of the nucleosomes using such matrices as mapping probes is at the moment the only way to computationally locate the nucleosomes on the sequences, with single-base resolution (Hapala & Trifonov, 2011, 2013). There is one seemingly

Figure 2. The full-length matrices of nucleosome DNA bendability, 4 dinucleotides (RR, YY, RY, YR) in 115 sequence positions, for SNs of three species. (A) A. thaliana, (B) C. elegans, (C) H. sapiens (chromosomes 1–5).
insurmountable problem: individual nucleosome sequences may have their individual biases in choice of a repertoire of periodically placed dinucleotides. In other words, the individual nucleosome would need its own specific matrix to adequately evaluate the nucleosome-forming quality of the sequence, which becomes impractical. The composition-specific matrices for different isochores of the same species are a good example (Frenkel et al., 2011).

The species- and composition-specific biases of individual nucleosome DNA sequences are leveled off if 16 different dinucleotides are combined in RR, YY, RY, and YR groups. Indeed, it is known since 1983 (Mengeritsky & Trifonov) that the consensus repeating motif in the nucleosome DNA is (RRRRYYYY)\textsubscript{\alpha}, rather than just (AAAAATTTTT)\textsubscript{\alpha}. This is confirmed in later sequence analysis studies (Salih et al., 2008, in press), and strongly supported by deformational properties of various [R,Y] type base-pair stacks of DNA (Cui & Zhurkin, 2010; Trifonov, 2010). This justifies the aim of this work — derivation of the full-length [R,Y] sequence consensus of the nucleosome DNA. The consensus of 116 R and Y bases, as above, can serve as the nucleosome DNA sequence probe, equally sensitive to AAAAAATTTTT and GGGGGCCCCC periodical motifs, as in \textit{A. thaliana} (Salih et al., in press) and in H3 isochores (Frenkel et al., 2011; Salih et al., in press), respectively, as well as to any other RRRRRYYYYY motif, like the consensus GGAAATTTCC for \textit{C. elegans} (Gabdank, Barash, & Trifonov, 2009). The [R,Y] sequence probe derived in this work is superior to full-length matrix of bendability for \textit{C. elegans} hitherto used for mapping purposes in various sequences. The [R,Y] probe is free of specific \textit{C. elegans} dinucleotide biases. In particular, it does not give any special preference to CG dinucleotides.

It may well be, however, that certain dinucleotides of the RR and YY families still do have their specific preferences to some positions within the repeating RRRRYYYYY motif due to their individual deformation properties. Future studies may clarify this point.

Figure 2. (Continued).
Figure 3. DNA bendability functions for all three species in [R,Y] alphabet. (A) Combined matrix of bendability, (B) difference plot RR–YY for derivation of linear form (consensus sequence) of the matrix.
2.5. Implementation of the [R,Y] probe for the nucleosome mapping

Usage of the [R,Y] consensus sequence rather than whole 16-line matrix, as in (Gabdank et al., 2010a, 2010b), for the mapping is both justified and more convenient. It is justified because the sequence in the [R,Y] alphabet is not sensitive to neither species-specific, nor individual compositional biases. It is equally applicable to A+T-rich and to G+C-rich sequences, since the RRRRRYYYYY pattern is common for all (Frenkel et al., 2011; Rapoport et al., 2011; Salih et al., in press).

It is more convenient, as the mapping is reduced to counting simple dinucleotide match to the probe at every sequence position. Maximal match is 115, according to the size of the probe, statistical match is 115/4≈29. The values in this range, from 29 to 115, provide a simple measure of the nucleosome “strength”. The strongest nucleosomes of C. elegans, for example, are characterized by match value 91 (Salih & Trifonov, in press).

It is important to keep in mind that the sequence is not the only determinant of the nucleosome position (Struhl & Segal, 2013; Valouev et al., 2011). It is, clearly, a contributor, as the nucleosomes do display the sequence periodicity (Trifonov & Sussman, 1980), especially stronger nucleosomes (Lowary & Widom, 1998). Other important actors in the game are transcription factors, barriers for nucleosome sliding, and proteins involved in chromatin remodeling (Teif & Rippe, 2009).

Any sequence-directed nucleosome mapping procedure only offers the sequence component of the positioning.

The particular choice of RR and YY dinucleotides by different species depends on many factors, G+C content first of all. The follow-up studies may reveal further details of the universal pattern. The publicly accessible server for the mapping with the [R,Y] probe is under construction. The code for the mapping can be also provided by request.

3. Materials and methods

3.1. Sequences of SNs

The sequences of A. thaliana SNs and C. elegans SNs are taken from the paper of Salih et al. (in press). The 512 sequences of SNs of H. sapiens (chromosomes 1–5) generated anew, using the same “magic distance” technique (ibid.), that is, by selecting DNA fragments with maximal number of distances multiples of the period 10.4 bases, between like-named dinucleotides. Human genome sequences are taken from: ftp.ncbi.nih.gov/genomes/H_sapiens.

3.2. Alignment of SN DNA sequences

The sequences of SNs of the three species have visible periodicity of obviously the same motif AAAAAATTTTT (ibid.). Therefore, the choice of the first and subsequent sequences to which the other sequences should be consecutively aligned, is not as crucial, as for the nucleosome DNA sequences with weak, hidden periodicity (Ioshikhes et al., 1996). Nevertheless, the precaution was taken, to ensure that none of the sequences of the whole ensemble would have an advantage and excessive influence on the outcome. Each of the three sets of the SN sequences for three species was treated separately, thus, generating species-specific nucleosome DNA bendability matrices.

Definitions: the 20-matrix, of 16 lines (dinucleotides) in 115 positions, represents single SN sequence, and consists of matrix elements 0 and 1.

2n-matrix consists of 16 lines (dinucleotides) and up to 153 columns, 115 initial ones, and additional flanking columns, originally empty. Each 2n-matrix is result of alignment of two 2n−1-matrices, and corresponds to 2n aligned sequences.

Flanking columns are introduced to accommodate ±15 base shifts during the pairwise alignments of the 2n-matrices.

Alignment is result of sliding of one 2n-matrix along another one, and calculating sums of products of corresponding matrix elements in register. When the line position with maximal sum of products is reached, the matching elements of two matrices are summed, generating the 2n+1-matrix.

The alignments were conducted in the following steps:

1. Pairwise alignment of the SN sequences, in arbitrary order, resulting in 256 21-matrices
2. Consecutive pairwise alignments of 2n-matrices, generating 2n+1 matrices, until exhaustion of all 512 (29) sequences. Last two matrices aligned are 28-matrices.
3. After each alignment the zero flanking columns are removed.

All computations have been conducted on Matlab.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2014.891262.

References

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.

Frenkel, Z. M., Bettecken, T., & Trifonov, E. N. (2011). Nucleosome DNA sequence structure of isochores. BMC Genomics, 12, 203.
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). FineSTR: A web server for single-base-resolution nucleosome positioning. *Bioinformatics, 26*, 845–846.

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.

Ong, M. S., Richmond, T. J., & Davey, C. A. (2007). DNA stretching and extreme kinking in the nucleosome core. *Journal of Molecular Biology, 368*, 1067–1074.

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*, 273–281.

Salih, B., & Trifonov, E. N. (in press). Strong nucleosomes reside in meiotic centromeres of C. elegans. *Journal of Biomolecular Structure and Dynamics.*

Salih, B., Tripathi, V., & Trifonov, E. N. (in press). Visible periodicity of strong nucleosome DNA sequences. *Journal of Biomolecular Structure and Dynamics.* doi: 10.1080/07391102.2013.855143

Struhl, K., & Segal, E. (2013). Determinants of nucleosome positioning. *Nature Structural & Molecular Biology, 20*, 267–273.

Teif, V. B., & Rippe, K. (2009). Predicting nucleosome positions on the DNA: Combining intrinsic sequence preferences and remodeler activities. *Nucleic Acids Research, 37*, 5641–5655.

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

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., & Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature, 474*, 516–520.