Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine–pyrimidine and pyrimidine–purine dimers

Victor B. Zhurkin

Institute of Molecular Biology, USSR Academy of Sciences, 117984 Moscow B-334, Vavilov str. 32, USSR

Received 17 May 1983

Comparison of the X-ray data and results of conformational analysis reveals the sequence-dependent helical anisotropy of DNA: purine–pyrimidine (RY) and pyrmidine–purine (YR) dimers prefer bending in the opposite directions. On this basis it is suggested that in the optimal nucleosomal DNA sequence the RY and YR dinucleotides alternate at intervals of 5–6 basepairs. Examination of the cases with the known disposition of nucleosomes confirms this concept.

Nucleosome DNA conformation DNA bends Nucleosome phasing DNA primary structure, periodicity

1. INTRODUCTION

At present there is a sufficiency of evidence suggesting that the arrangement of nucleosomes on DNA is nonrandom, even in the absence of non-nucleosomal factors (see reviews [1,2]). One of the ways of explaining this phenomenon is to suppose [3] that formation of nucleosomes is dictated by thermodynamic favorability of wrapping the DNA fragment around the core, depending on DNA sequence [4,5].

Then the strategy of search for the sequences preferable for formation of nucleosomes is controlled by the nature of DNA bending flexibility. Our energetics calculations [5] have shown that the double helix is extremely anisotropic and bends into the groves much more easily than in perpendicular directions. So it was suggested [5] that the nucleosomal DNA is bent by means of 'mini-kinks' separated by 5 bp and directed into both grooves alternatively (fig. 1a). Our model has now been confirmed by the crystal structure of B-DNA dodecamer [6, 7]. As follows from fig. 1b, this oligomer has two bends (called by the authors 'annealed kinks' [7]) in accord with our predictions (fig. 1a). The upper bend is directed precisely into the major groove (CG), the lower 'mini-kink' (GC) is directed into the minor groove to the left and away from the viewer.

It is clear that due to anisotropy of DNA only those dimers are significant in which the DNA helix is bent, all the others can be neglected [8]. This is the main difference between the present approach and that [3,4] who considered the isotropic model of DNA and proposed that all dimers are equally significant. (The 10.5 bp periodicity found by them in some eukaryotic sequences [3] is related not to the packing of DNA in nucleosomes but rather to the secondary structure of the coded proteins [9].)

Based on the above data and results of computation of AAAA: TTTT, $(AATT)_2$ and $(TTAA)_2$ in [10], one can subdivide all dinucleotides into three classes [8]:

(i) Purine–pyrimidine dimers (RY) that prefer bending to the minor groove,

(ii) Pyrimidine–purine dimers (YR) bending more easily into the major groove;

(iii) RR and YY, for which both directions are nearly equivalent; RR and YY are more stable than RY or YR.

The conclusion on anisotropy of dinucleotides

Published by Elsevier Science Publishers B.V.
00145793/83/$3.00

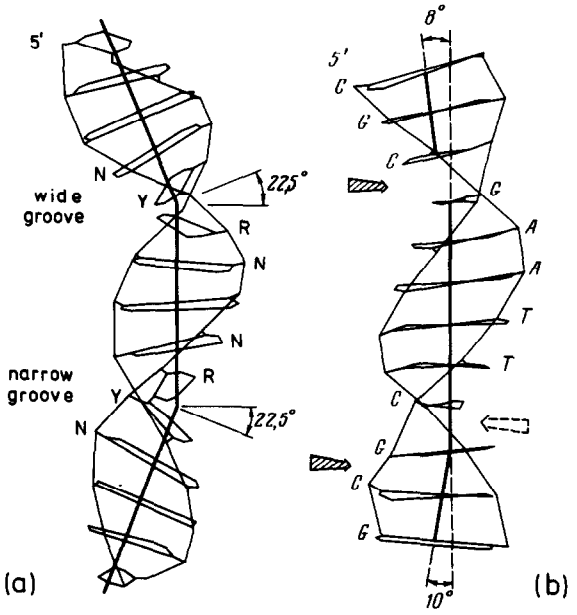

Fig. 1. Model for DNA packing in nucleosomes by means of 'minikinks' [5] (a) and the crystal structure of B-DNA dodecamer [6] (b). (a) In the optimal nucleotide sequence bends to the wide and narrow grooves occur at YR- and RY-dimers. (b): Helical axes for the 3 fragments were obtained as the least square lines for geometrical centers of basepairs, namely the centers of C8-C6 segments. The 'annealed kinks' are shown by dashed arrows; the broken arrow depicts opening of base pairs without bending of axis.

RY and YR is further supported by a simple geometrical fact – for YR sequences purines on the adjacent basepairs come into close contact in the minor groove, while for RY sequences the clash occurs in the major groove [11]. This steric hindrance would cause bending of the DNA helix in the corresponding direction.

Thus we propose that in the optimal nucleosomal DNA the RY and YR dimers alternate at intervals of a half-pitch of the double helix (fig. 1a); RYabcYRdefRYghiYR... Below, the sequences with the known arrangement of nucleosomes are analysed on the basis of this concept and it is shown to be in agreement with the experimental data.

2. METHODS

To introduce special functions measuring favorability of a given DNA fragment to be wrapped around the core, we need several assumptions:

(1) Suppose that the DNA period is constant and equals 10 or 10.5 bp. In the first case, the minikinks are separated by 5 bp, in the second, this distance varies from 5–6 nucleotides, so that the full period is 21 bp.

(2) In the middle of the nucleosomal DNA fragment there is a bend into the wide groove. It follows from the DNases I and II digestion data: they cut the 146 bp segment at the distance of about 1 bp from the middle with the stagger of not more than 2 bp [12], so as this point the narrow groove is located on the outer surface of nucleosomal particle (see fig. 2 from [13].

This is in agreement with the data in [14] that the interaction sites of the lysine residues of histones with DNA are shifted by ~5 nucleotides relative to the nuclease cleavage sites.

Item 2 leads to a conclusion that if the DNA period is 10.5 bp, then at the edges of the 146 bp fragment, just as in the middle, there are bends into the wide groove. If the period is 10.0 bp, then those bends are situated at the distance of 3 nucleotides from both the margins.

So for the arbitrary sequence A_1, A_2,.. A_n the following functions $f(m)$ and $f'(m)$ are introduced estimating 'bendability' of the 146 bp segment beginning at A_m. For the period of 10.0 bp, $f(m)$ equals to the number of YR dimers located in positions $m+3$, $m+13$,.. $m+143$ plus the number of RY-dimers located in $m+8$, $m+18$,.. $m+138$. We say that the RY-dimer is placed in position m if dinucleotide A_mA_{m+1} is purine–pyrimidine. (Remember that YR favors bending into the wide groove, therefore it is the YR-dimer that should be placed in the middle of the core DNA fragment and at the ends of it.) Similarly, for the period of 10.5 bp, function $f'(m)$ equals the sum of YR-dimers situated in m, $m+10$, $m+11$, $m+21$, $m+31$, $m+32$,.. $m+147$ and RY positioned in $m+5$, $m+6$, $m+15$, $m+16$,.. $m+141$, $m+142$.

Note that the maximal possible value of these functions is 29. As to the mean value, this depends on N, the relative amount of RY-dimers (or YR, which is the same) in the given sequence: $\bar{f} = 29 \cdot N, \bar{f}' = 50 \cdot N$. For a random sequence $N = 1/4$, thus $\bar{f} = 7.25, \bar{f}' = 12.5$. Lastly, the functions are introduced in such a way that their maxima should correspond to the 'left' border of the most favorable core DNA fragments.

3. RESULTS AND DISCUSSION

3.1. *AGM component*

According to the data in [15] which dealt with the repetitive sequence from African green monkey (AGM), the preferential cleavage site of micrococcal nuclease lies 126 ± 6 bp from the *Hin*dIII site (arrows in fig. 2). Mapping of nucleosomes on the basis of 10.5 bp period function $f'(m)$-places this cleavage site just in the linker region, while the 10 bp period function $f(m)$ predicts the left border of nucleosome to be ~ 5 bp to the left from the *Eco*RI site. Note that $f'(m)$ has a strong peak in this region as well. That our functions have their local maxima separated by 10–11 bp follows from the character of the functions used. It agrees with the known fact that the discrete positions of nucleosomes stand ~ 10 bp apart [1,2].

Thus, the above results are in fair accord with the data in [15] (function $f'(m)$, but they indicate also that the mode of organization of nucleosomes postulated [15] is not probably unique.

3.2 *Rat satellite I*

In [16], 3 distinct locations of nucleosomes were detected on the satellite DNA, shown by rectangles in fig. 3. In [17], 14 positions of nucleosomes were precisely mapped which proved to be in obvious

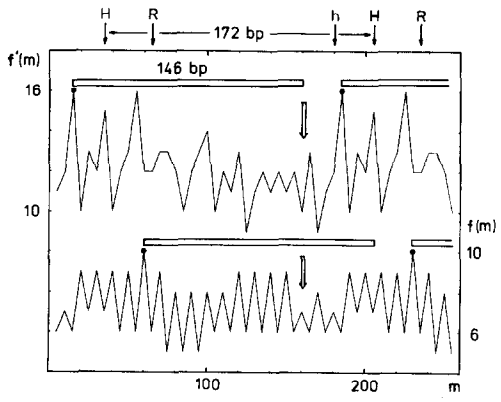

Fig. 2. Functions $f(m)$ and $f'(m)$ for the African green monkey component α. White arrows show preferential cleavage sites obtained in [15]. Here and in fig. 3, 'H', 'h' and 'R' denote restriction sites of *Hin*dIII, *Hae*III and *Eco*RI. In figs. 2–4 for simplicity of drawing each point on a curve shows maximum value of a function in 5 consecutive points: $m = 5k+1, \ldots 5k+5$. Actually, $f(m)$ varies from 1–10; $f' = 3 + 16$.

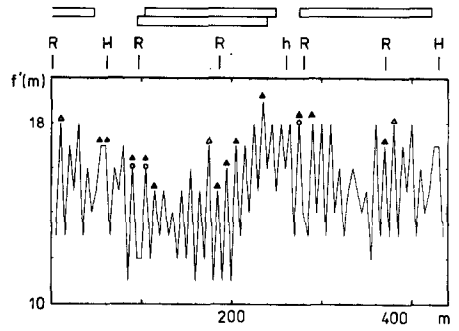

Fig. 3. Function $f'(m)$ for the rat satellite I DNA. Rectangulars in the upper part and circles on the curve denote nucleosomal positions according to ref. [16]. The new data [17] are shown by triangles. $f'(m)$ varies from 6–19.

conformity with the function $f'(m)$ of period 10.5 bp (fig. 3). Not only the strongest maximum at $m = 233$ coincides with the nucleosomal position, but other new locations are also consistent with the behaviour of $f'(m)$. Fig. 3 (▲) denote those positions which diverge from the local maxima by 0–2 bp only. Moreover, these locations are shifted to the right from the peaks; this makes the comparison far more convincing since the core DNA fragments are somewhat (3–4 bp) shorter than 146 bp due to exonuclease nibbling [17]. Only two maxima do not agree with the experimental data: at $m = 9$ instead of 7, and $m = 173$ instead of 177 (core DNA = 143 bp) (fig. 3). But in these cases the

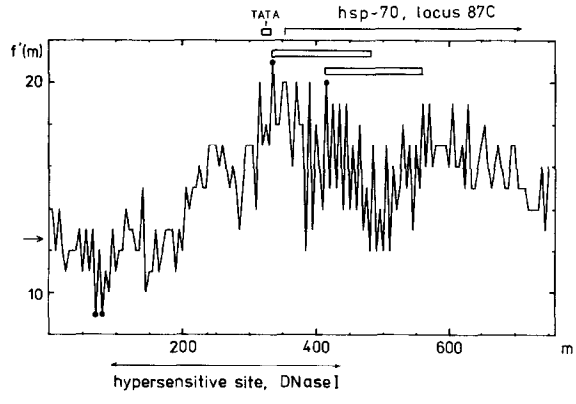

Fig. 4. 'Bendability' function predicts location of hypersensitive site [18] for the *hsp*-70 gene of *Drosophila*, locus 87 C1 [19]. Horizontal arrow outside the frame indicates the mean value of a function for random sequence.

function *f(m)* has its peaks in the proper positions ($m = 5$ and 175, not shown). However, generally function *f(m)* with a period of 10 bp is in a poorer accordance with the detected locations of nucleosomes 5 peaks of 14 are shifted in the wrong direction, and those which are not, are comparatively low.

So, the latest data [17] give preference to the 10.5 by periodicity. Besides, some new locations might be expected on the basis of our investigation, e.g., at $m = 357$, 30, 223, function *f' (m)*, and $m = 125$, 309, function *f(m)*.

3.3 *Heat shock genes of* Drosophila

The 5'ends of *hsp*-70 genes in chromatin are known [18] to be hypersensitive to DNase I; it is supposed that these sites are free from nucleosomes. Analysis of the DNA sequence has shown (fig. 4) that the absolute minimum of *f' (m)* at $m = 85$, which indicates the least probable position of a nucleosome, coincides with the left margin of the hypersensitive site discovered in [18]: $m = 90$–430 in our notation. The right end of this site goes beyond the total maximum of *f' (m)* at $m = 355$, but this discrepancy can be easily avoided if we assume that the nucleosome is located alternatively at $m = 335$–415.

It is of interest that in another hypersensitive site, near the replication origin of simian virus 40 [20], the functions also have sharp minima: *f' (m)* attains an extremely small value of 3. No doubt this result is statistically reliable: *f' (m)* varies from 3–21, while the mean value and RMS deviation for the random sequence are 12.5 and 2.5, respectively.

3.4 *Other cases*

On the *lac*-UV5 fragment both the functions *f* and *f' (m)* reach their total maxima only 2 bp apart from the position located in [21] (see [8]). For the three 5 S RNA-genes [22–24] and the histone genes of *Drosophila* [25] the experimental data can be compared with our calculations only qualitatively, as in the case of AGM component α (fig. 2). Such a comparison was made and it gave reasonable agreement.

4. CONCLUSION

The above analysis has shown that the concept proposed here is consistent with the known cases of specific alignment of nucleosomes. Generally the experimental data are in a better agreement with the function of period 10.5 bp, though sometimes period 10.0 is more suitable (see section 3.2). Perhaps this means that periodicity of the DNA secondary structure in a nucleosome depends on its sequence.

We consider here an oversimplified model; for instance, the role of nucleotides adjacent to putative 'mini-kinks' is not taken into account. But as a first step it is a proper thing to do. For the final test of our scheme experiments are needed where all possible positions of nucleosomes are defined with the precision of 1–2 bp. Now we can only state that the vast majority of nucleosome positions localized precisely deviate from the peaks of the functions by no more than 2 bp (*lac*-UV5 and rat satellite I).

There is also some indirect evidence in favor of the concept presented here, namely preferable formation of nucleosomes on the alternating sequences poly[d(A–T)] and poly [d(G–C)] in contrast to poly(dA): poly(dT) and poly(dG): poly(dC) which inhibit nucleosome formation [26, 27]. These data can be easily explained by our scheme, since in the alternating polymers the RY- and YR-dimers are separated by 5 bp, while in the homopolymers there are neither RY nor YR; consequently, the double helix resists wrapping around the core.

Compare this concept with the approach in [3, 4]. They suppose that the strong binding sites of nucleosomes are characterized by a pronounced periodicity in the nucleotide sequence, which may exist only as an exception. On the contrary, here it is suggested that specific alignment of nucleosomes needs only degenerative periodicity:

(i) All dinucleotidues are divided into 3 classes according to their 'mechanical' properties;
(ii) Only each fifth dimer is meaningful.

Consequently, a random nucleotide sequence can have some sites with the increased affinity for nucleosomes (as in the case of procaryotic *lac*-UV5 fragment). Indeed, a general pattern of the functions used here proved to be similar for the SV40 sequence and the pseudo-random 5000 bp sequence generated by a computer. The main difference is that the random sequence has no such wide and deep declines in the function profile as

SV40 has near the replication origin or the *hsp*-genes have at their 5′-ends. As it follows from section 2, the strong declines in the profile of *f(m)* and *f′(m)* are caused by the diminishing of the number of RY- and YR-dimers in these areas, in other words, by numerous polypurine or polypyrimidine blocks. It is tempting to suppose that these fragments can play a regulatory role decreasing the probability of nucleosome formation.

ACKNOWLEDGEMENTS

The author is thankful to Drs V.I. Ivanov, A.D. Mirzabekov, S.A. Nedospasov and N.B. Ulyanov for valuable discussions and to Drs R.E. Dickerson and T. Igo-Kemenes for sending their data prior to publication.

REFERENCES

[1] Kornberg, R.D. (1981) Nature 292, 579–580.
[2] Zachau, H.G. and Igo-Kemenes, T. (1981) Cell 24, 597–598.
[3] Trifonov, E.N. and Sussman, J.L. (1980) Proc. Natl. Acad. Sci. USA 77, 3816–3820.
[4] Trifonov, E.N. (1980) Nucleic Acids Res. 8, 4041–4053.
[5] Zhurkin, V.B., Lysov, Yu. P. and Ivanov, V.I. (1979) Nuclei Acids Res. 6, 1081–1096.
[6] Dickerson, R.E. and Drew H.R. (1981) J. Mol. Biol. 149, 761–786.
[7] Fratini, A.V., Kopka, M.L., Drew, H.R. and Dickerson, R.E. (1982) J. Biol. Chem. 257, 14686–14707.
[8] Zhurkin, V.B. (1982) Studia Biophys. 87, 151–152.
[9] Zhurkin, V.B. (1981) Nucleic Acids Res. 9, 1963–1971.
[10] Ulyanov, N.B. and Zhurkin, V.B. (1983) in preparation.
[11] Calladine, C.R. (1982) J. Mol. Biol. 161, 343–352.
[12] Lutter, L. (1981) Nucleic Acids Res. 9, 4251–4265.
[13] Klug, A. and Lutter, L. (1981) Nucleic Acids Res. 9, 4267–4283.
[14] Shick, V.V., Belyavsky, A.V., Bavykin, S.G. and Mirzabekov, A.D. (1980) J. Mol. Biol. 139, 499–518.
[15] Musich, P.R., Brown, F.L. and Maio, J.J. (1982) Proc. Natl. Acad. Sci. USA 79, 118–122.
[16] Igo-Kemenes, T., Omori, A. and Zachau, H.G. (1980) Nucleic Acids Res. 8, 5377–5390.
[17] Seligmann, H. and Igo-Kemenes, T. (1983) personal communication.
[18] Wu, C. (1980) Nature 286, 854–860.
[19] Török, I. and Karch, F. (1980) Nucleic Acids Res. 8, 3105–3123.
[20] Scott, W.A. and Wigmore, D.J. (1978) Cell 15, 1511–1517.
[21] Chao, M.V., Gralla, J. and Martinson, H.G. (1979) Biochemistry 18, 1068–1074.
[22] Gottesfeld, J.M. and Bloomer, L.S. (1980) Cell 21, 751–760.
[23] Louis, C., Schedl, P., Samal, B. and Worcel, A. (1980) Cell 22, 387–392.
[24] Simpson, R.T. and Stafford, D.W. (1983) Proc. Natl. Acad. Sci. USA 80, 51–55.
[25] Samal, B., Worcel, A., Louis, C. and Schedl, P. (1981) Cell 23, 401–409.
[26] Bryan, P.N., Wright, E.B., Hsie, M.H., Olins, A.L. and Olins, D.E. (1978) Nucleic Acids Res. 5, 3603–3617.
[27] Simpson, R.T. and Künzler, P. (1979) Nucleic Acids Res. 6, 1387–1415.