Periodicity in DNA primary structure is defined by secondary structure of the coded protein

Victor B.Zhurkin

Institute of Molecular Biology, Academy of Sciences of the USSR, 117984 Moscow B-334, Vavilova 32, USSR

Received 23 February 1981

ABSTRACT

A 10.5-base periodicity found earlier is inherent in both eu- and prokaryotic coding nucleotide sequences. In the case of noncoding eukaryotic sequences no periodicity is found, so the 10.5-base oscillation seemingly does not correlate with the nucleosomal organization of DNA. It is shown that the DNA fragments, coding the α-helical protein segments, manifest the pronounced 10.5-base periodicity, while those regions of DNA which code the β-structure have a 6-base oscillation. The repeating pattern of nucleotide sequences can be used for comparison of the DNA segments with low degree of homology.

INTRODUCTION

There is now a strong evidence suggesting that disposition of nucleosomes on DNA is nonrandom (1-3). The nature of such a specificity is still unclear. An interesting attempt to explain this phenomenom has been made recently by Trifonov and Sussman (4,5). They found, that in the chromatin DNA nucleotide sequences some of the dinucleotides are repeated with the period of 10.5 bases, which coincides with the pitch of DNA double helix. It means that identical dinucleotides are separated by an integer number of the helical turns, so that when bending the double helix into a smooth loop, these dinucleotides would be oriented in the same way relative to the direction of bending. Proceeding from this fact and the postulated sequence-dependent deformational anisotropy of DNA (4,5) the authors concluded that the above periodicity promotes the packing of eukaryotic DNA in chromatin. Moreover, it follows from their assumpion that the nucleotide sequence determines also the outer and inner sides of nucleosomal DNA (4).

Nevertheless, theconception cited has at least two painful disadvantages. Firstly, the periodicity under discussion is too weak to influence noticeably "bendability" of the double helix, since less than 10% of all dinucleotides are repeated at a distance of 10-11 base pairs (4). Secondly, an assumption that deformational anisotropy is sequence-dependent does not agree with both experimental and theoretical data. As to the supposed equilibrium wedge-like structure of some of the dinucleotides (4), this hypothesis is inconsistent with the x-ray data on the B-form in crystal (6) according to which nine of eleven dimeric duplexes in dodecamer CGCGAATTCGCG have practically ideal regular conformation. In connection with the dynamic aspect of anisotropy (5) note that as was shown earlier by our computations (7), flexibility of the DNA double helix in direction of major and minor grooves markedly (by more than order of magnitude) exceeds flexibility in other directions. These findings are in accord with the structure obtained by Wing et al.(6) where the two CG dinucleotides flanking the AATT tetramer are bent by a 10° angle towards the grooves. It is of interest that bending of the DNA in direction perpendicular to the dyad axis increases the sugar-phosphate interaction energy mainly, i.e. the sequence-independent term. Therefore it is difficult to expect that in one site the DNA molecule would be most flexible along the dyad axis while in another site - in perpendicular direction.

These considerations made us search for another explanation of the periodicity, discovered by Trifonov and Sussman (4).

METHODS

The following positional autocorrelation function, P(n), is considered, defined for the nucleotide sequence $(a_1\ ,\ a_2\ ,\ a_3\ \ldots\ a_N\ )$:

$$P(n) = \sum_{i=1}^{N-n-1} p_i(n)\ /\ (N-n-1)\quad ,\ \text{where } p_i(n)=1 \text{ if } a_i=a_{i+n}\ ,\ a_{i+1}=a_{i+n+1}$$

and $p_i(n)=0$ otherwise (it is supposed that $n < N-1$). In other words, P(n) measures the frequency of occurence of the same dinucleotides at a distance n from each other along the given sequence. For the circular DNA it is assumed that $a_{N+i}= a_i$ and $P(n)=\sum_{i=1}^{N} p_i(n)\ /\ N$. In order to analyse several sequences the numerators and denominators must be summed up separately.

In this study as well as in the paper (4) only one strand of DNA is examined. The only feature, which differs our procedure from that of Trifonov and Sussman (4), is an extra multiplier $1/(N-n-1)$. Thus,"the frequency of occurence" is evaluated rather than "the number of occurences", $\sum p_i(n)$ (see ref. 4). It allows to deal properly with DNA segments of a relatively short length, N.

RESULTS AND DISCUSSION

First a comparative analysis of pro- and eukaryotic sequences enriched by coding regions was made. It was found that dependence of P on n for the

SV40 (8) and ØX174 (9) nucleotide sequences has practically the same pattern (see Fig. 1). The autocorrelation function has maxima at n = 3, 9, 12, 21, 30, 33 ... and the 10.5-base periodicity does reveal itself, at least for n ⩽ 45. Interestingly, it is the ØX174 sequence which is characterized by a more regular oscillation of P(n) : local maximum corresponds to n = 42 instead of n = 39 for SV40. Note that the values of P(n), presented here for the SV40 sequence, reproduce the data of Trifonov and Sussman (4). As to the statement of these authors that for the prokaryotic sequences on the whole "the 10.5-base periodicity seems not to be present" or to be "nonconvincing statistically" (4), we confine ourselves to an example, where it apparently exists (Fig. 1).

At the second stage the eukaryotic noncoding segments (introns and spacers) were investigated. As it follows from the Fig.2, these DNA fragments do not show any periodicity, though they cannot be regarded as completely random, since the mean value of P(n) is higher than 1/16 = 6.25% , especially for the small n. The latter can be explained by the presence of the (AA...A): :(TT...T) and, more seldom, (GG...G):(CC...C) clusters in the examined sequences. Quite naturally, not only the 10.5-base periodocity is absent in this case, but a 3-base one (4) either.

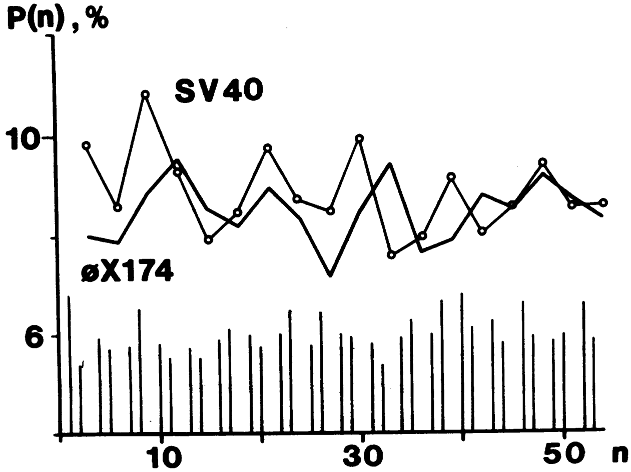

Figure 1. Frequency of occurence of the same dinucleotides, P , as a function of distance between them, n , for SV40 (solid line) and ØX174 DNA sequence (line with circles and vertical bars). From here on the envelope curves correspond to the distances of multiples of 3 bases; for the other n the value of P(n) is shown by vertical bars.

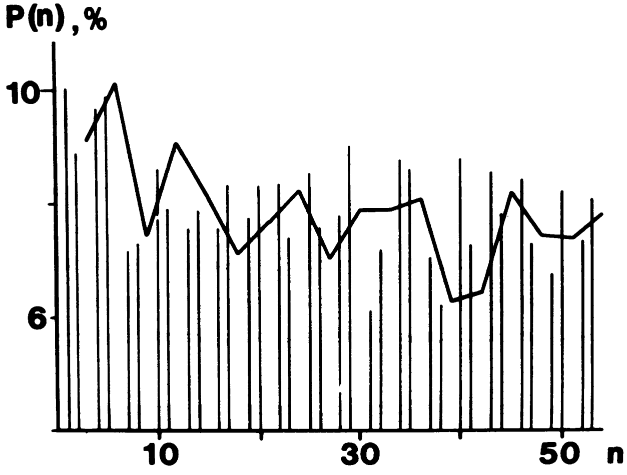

Figure 2. Autocorrelation function for noncoding sequences (introns and spacers) from the genes of rabbit β -globin (12), mouse Ig λ light chain (19) and cluster of histone genes of P. miliaris (20). The sequences analysed have a total length of about 2300 nucleotides.

So we have seen that the 10.5-nucleotide oscillation is inherent in both eu- and prokaryotic coding sequences and does not express itself in noncoding DNA fragments. Therefore it seems reasonable that this periodicity is connected with the structure of the coded protein.

It is well known that one of the basic elements of the three-dimentional structure of proteins is an α -helix. As a rule, the hydrophobic residues are collected in a continuous domain on its surface which is favorable for formation of the protein globule (10,11). The α-helix has a period of 3.6 residues, therefore the residues i+1, i+3, i+4, i+7, i+10, i+11, i+14 are placed on the same side of the helix with the i-th one. The coding triplets for these amino acids are separated from the i-th triplet by n = 3, 9, 12, 21, 30, 33, 42 nucleotides. If all mentioned amino acids were nonpolar (see above) then the corresponding codons would have much in common. For instance, most of them would have thymine in the second position. So one can expect that for the "α -helical DNA" (i.e. a DNA fragment coding an α -helical segment of a protein) autocorrelation function P(n) has its maxima just for the mentioned magnitudes of n, that is for the same n as in the case of SV40 and ØX174 DNA.

In an oversimplified way this reasoning can be formulated as follows . the length of periodicity in the nucleotide sequence, coding the given protein elementary regular structure, equals the period of this structure, multiplied by 3. Thus, the "α-helical DNA" would have the length of oscillation $3.6 \cdot 3 = 10.8$, which is close to 10.5. (The last two values are evidently indistinguishable in our case, see Fig.1). Similarly, "β-structural DNA" would have 6-nucleotide oscillation since the period of the β-structure itself equals 2.

In order to check the validity of the previous consideration the genes of the proteins, enriched by α-helices (β-globin) and β-structures (immunoglobulin) were thoroughly examined. The result for the nucleotide sequences from the rabbit β-globin gene (12) , coding five α-helical segments (A,B, E,G,H, see ref. 13) with the length from 15 to 21 residues, is presented in Fig.3a. Firstly, note that P(n) has its maxima at the predicted values of n : 3, 12, 21, 42; the only exception is n = 36 instead of 30 or 33. Secondly, the autocorrelation function reaches twice the SV40 DNA level and increases up to 20% . In other words, "α-helical DNA" does have the period of 10.5 (more precisely, 10.8) which is more pronounced than for the DNA sequence on the whole.

The question arises, however, as to whether the found periodicity is dictated by a special choice of coding triplets (4,5) or by amino acid sequence itself. To answer this question the procedure of "reverse translation" was

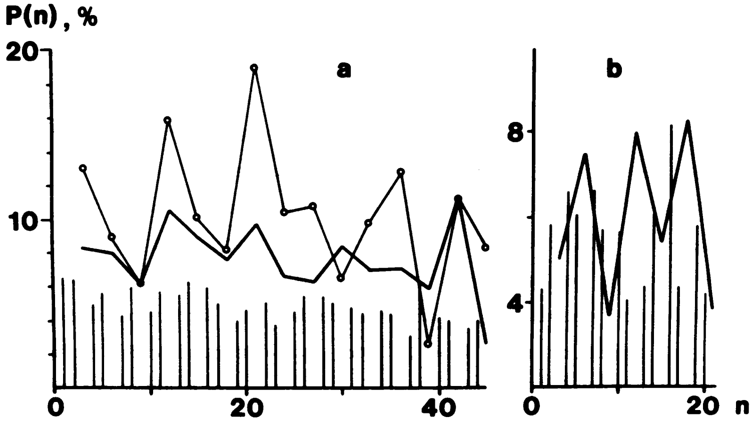

Figure 3. Autocorrelation function for the nucleotide sequences coding α-helical (a) and β-structural (b) regions of proteins. The line with circles corresponds to the real DNA sequences, the solid lines and vertical bars are for the sequences generated by the random "reverse translation" procedure (see the text)

used. The nucleotide sequence was generated after the amino acid sequence so that the corresponding codons were chosen randomly. Application of this method to the five α-helices mentioned earlier has proved that the function P(n) retains its periodicity after randomization (Fig.3a). Near n = 30 "the random curve" is even more similar to an ideal 10.5-base oscillation pattern, since it has a local maximum at n = 30 . The diminishing of the P(n) values in this case is quite natural due to the known preferential usage of some triplets in living systems (8,20).

As to the "β-structural DNA", it is also found to have the predicted 6-base periodicity (see Fig.3b). This result was obtained for the "random sequences", corresponding to the six β-structural regions (14) from the constant domain of the human Ig $\varkappa$ light chain (15), having from 7 to 15 amino acid residues. Of note is, however, a less regular oscillation of P(n) and a decrease of the P(n) values in comparison with α-helices (cf. Figs. 3a and 3b).It is probably explained by a less ordering of the β-structures at an average (16,17). For instance, consideration of the mouse Ig $\varkappa$ (18) or Ig λ (19) light chain genes and their contiguous "β-structural" fragments shows violation of the 6-base periodicity at $n \geqslant 15$. On the contrary, analysis of the entire β-globin gene (12) and the genes of histones H2A and H3 (20), also enriched by the α-helices, reveals a clearcut periodicity up to n = 50.

Thus, we have shown that : (i) the periodicity of about 10.5 bases is characteristic of both pro- and eukaryotic coding sequences; (ii) the noncoding eukaryotic sequences do not reveal any periodicity; (iii) the "α-helical" DNA segments display oscillation with the period of 10.5 nucleotides, while the "β-structural DNA" has the period of 6 nucleotides. Since the α-helices are found in proteins much more frequently than the β-structures and the other elements, such as a collagen-type helix (17), it is tempting to suppose that the total oscillation in the nucleotide sequences is caused by the DNA fragments coding α-helices. If one assumes that the whole SV40 DNA sequence is a coding one and P(n) rises up to 20% at the "α-helical" sections (Fig.3a), and equals 6% at an average for the other part of DNA, then the α-helical fraction should be 30% to give the observed maximum value of P(n) 10% in the case of SV40 DNA (Fig.1). (We consider here the magnitude of 6% since it is the mean value of P(n) for the random sequence: 1 / 16 =6.25%). This estimation of α-helicity is quite realistic (17).

One more indirect confirmation of the link between 10.5-base oscillation and α-helices is a breakdown of the periodicity of P(n) at n greater than 50 (4). Indeed, as was mentioned above, the length of α-helices in β-globin does not exceed 15-20 residues; it is likely to be a general rule. The distance between the neighbouring α-helices can be arbitrary, therefore the 10.5-nucleotide oscillation, caused by particular sequence of amino acids in the α-helical fragments, should not spread over the distance of 45-60 nucleotides, that is in keeping with the data of Trifonov and Sussman (4).

Following the lead of Trifonov and Sussman, one might think on the basis of our findings that those sections of DNA, which code α-helices, could be bent more easily in nucleosomes, but, because of the arguments presented in Introduction, it seems highly improbable.

So, we come to a conclusion that the periodicity observed in the primary structure of DNA is connected not with the nucleosomal organization of the eukaryotic DNA, but rather with the specific arrangement of the amino acids in a coded protein, thereby with its secondary structure.

CONCLUSION

Such particular sites of DNA as promotors, origins of replication, etc. can be compared with the punctuation marks in a printed text- these are the elements, which determine the "syntax" of the DNA language. To understand DNA functioning one should learn the laws regulating short-range order in the nucleotide sequence as well ("orthographical" laws). The periodicity found by

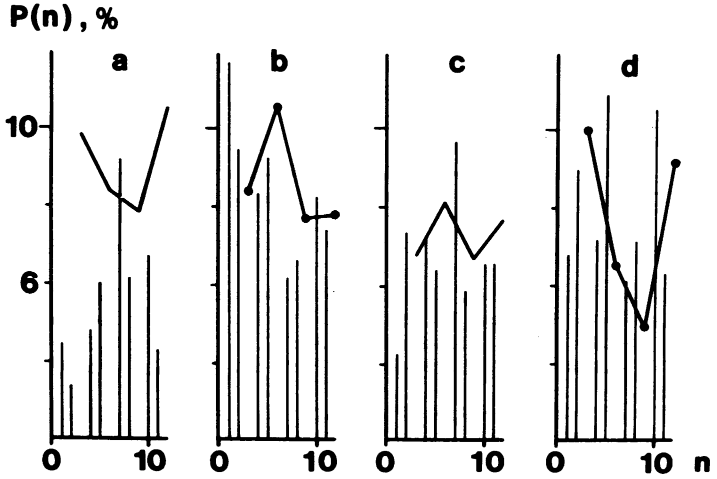

Figure 4. Comparison of autocorrelation functions for coding (a,c) and non-coding(b,d) sequences. Figures (a,b) are for the rabbit β-globin gene (12), (c,d) correspond to the mouse Ig λ light chain gene (19).

Trifonov and Sussman (4) is one of not numerous so far "orthographical" laws of the DNA language (see also refs. 21-23). In a conclusion I wish to show an example of application of the autocorrelation function P(n), characterizing this periodicity.

Compare this function for the coding and noncoding segments in case of two genes: β -globin (12) and immunoglobulin (19). As was pointed out earlier, the noncoding regions as a whole do not reveal any periodicity (see Fig.2) , but consideration of the dependence P(n) only for the small $\underline{n}$, which are multiples of 3, shows that the introns and spacers, flanking the coding sequences, differ from them by the pattern of the peaks (Fig.4). Namely, noncoding sequences of the "α -helical" gene have a "β -structural" pattern of P(n) (cf. Figs. 4a and 4b) and vice versa (Figs. 4c and 4d). It probably indicates that these noncoding segments descend from some other distant coding regions.

At any rate it seems reasonable to apply the autocorrelation function for comparison of the highly diverged nucleotide sequences, i.e. in the case when direct search of homology is ineffective.

ACKNOWLEDGEMENTS

The author is thankful to Drs. V.I.Ivanov, L.V.Abaturov, S.A.Nedospasov and O.V.Rokhlin for fruitful discussions.

REFERENCES

1. Ponder,B.A.J. and Crawford,L.V.(1977) Cell 11, 35-49.
2. Nedospasov,S.A. and Georgiev,G.P.(1980) Biochem. Biophys. Res. Commun. 92, 532-539.
3. Levy,A. and Noll,M.(1980) Nucl. Acids Res. 8, 6059-6068.
4. Trifonov,E.N. and Sussman,J.L.(1980) Proc. Natl. Acad. Sci. USA 77, 3816-3820.
5. Trifonov,E.N.(1980) Nucl. Acids Res. 8, 4041-4053.
6. Wing,R., Drew,H., Takano,T., Broka,C., Tanaka,S., Itakura,K. and Dickerson, R.E.(1980) Nature 287, 755-758.
7. Zhurkin,V.B., Lysov,Yu.P. and Ivanov,V.I.(1979) Nucl. Acids Res. 6, 1081-1096.
8. Reddy,V.B., Thymmappaya,B., Dhar,R., Subramanian,K.M., Zain,B.S., Pan,J., Ghosh,P.K., Celma,M.L. and Weissman,S.M.(1978) Science 200, 494-502.
9. Sanger,F., Coulson,A.R., Friedman,R;, Air,G.H., Barrel,B.G., Brown,N.L., Fiddes,J.C., Hutchison,C.A., Slocombe,P.M. and Smith,M.(1978) J. Mol. Biol. 125, 225-246.
10. Schiffer,M. and Edmundson,A.B.(1967) Biophys. J. 7, 121-135.
11. Lim,V.I.(1974) J. Mol. Biol. 88, 857-872.
12. van Ooyen,A., van der Berg,J., Mantei,N. and Weissman,C.(1979) Science 206, 337-344.
13. Sack,.J.S., Andrews,L.C., Magnus,K.A., Hanson,J.C., Rubin,J. and Love,W.E. (1978) Hemoglobin 2, 153-169.
14. Edelman,G.M.(1970) Biochemistry 9, 3197-3205.

15. Poljak,R.F., Amzel,L.M. and Phizackerly,R.P.(1976) Progr. Biophys. Mol. Biol. 31, 67-93.
16. Schulz,G.E., Barry,C.D., Friedman,J., Chou,P.Y., Fasman,G.D., Finkelstein, A.V., Lim,V.I., Ptitsyn,O.B., Kabat,E.A., Wu,T.T., Levitt,M., Robson,B. and Nagano,K.(1974) Nature 250, 140-142.
17. Volkenstein,M.V.(1977) Molecular Biophysics, Academic Press, New York.
18. Hamlyn,P.H., Brownlee,G.G, Cheng,C.-C., Gait,M.J. and Milstein,C.(1978) Cell 15, 1067-1075.
19. Bernard,O., Hozumi,N. and Tonegawa,S.(1978) Cell 15,1133-1144.
20. Schaffner,W., Kunz,G., Daetwyler,H., Telford,J., Smith,H.O. and Birnstiel, M.L.(1978) Cell 14, 655-671.
21. Ivanov,V.I., Zhurkin,V.B., Zavriev,S.K., Lysov,Yu.P., Minchenkova,L.E., Minyat,E.E., Frank-Kamenetskii,M.D. and Schyolkina,A.K.(1979) Int. J. Quant. Chem. 16, 189-201.
22. Nussinov,R.(1980) Nucl. acids Res. 8, 4545-4562.
23. Grantham,R.(1980) FEBS Letters 121, 193-199.