Further Observations on Periodicities of Nucleotide Occurrences in Natural DNA's

N. Burr Furlong and Koenraad Marien

The Graduate School of Biomedical Sciences and the School of Public Health of the University of Texas Health Science Center, 6901 Bertner Dr., Houston, TX 77030

Z. Naturforsch. **40c,** 854–857 (1985); received May 20, 1985

DNA, Sequences, Autocorrelation, Periodicity, Bacteriophage

There are non-random features in the occurrences of nucleotides in the DNA's of certain organisms which are detectable by statistical analyses of the entire sequence. Earlier, using the bacteriophage Phi-X 174 DNA sequence, we had reported that the self-information values for one type of dinucleotide association showed a marked periodicity when their autocorrelation coefficients were graphed. A similar, but computationally simpler, analysis has been developed which gives a comparable indication of periodicity. The difference, in average autocorrelation coefficients obtained with this analysis, between the peak values and all others has been used as an index to compare the extent of periodic non-randomness for a series of natural DNA sequences and for various artificial sequences. Calculations show that triplet periodicity, the relationship between dinucleotides separated by a single nucleotide, is characteristic only of the natural sequences of certain filamentous phages and is not found prominently in any other DNA analyzed (including sequences of similar length from plasmids, yeast, bacteria and higher animals). By shuffling nucleotides in a given sequence or by substituting selected nucleotides to alter various positions in both periodic and aperiodic sequences, we have found that an excess or deficiency of a given nucleotide at one of the three positions in a triplet reading frame can simulate the periodic characteristic. Thus, it appears that this global statistical analysis detects the tendency for single-strand phages to utilize a specific nucleotide, rather than one randomly selected, to constitute codons.

Introduction

In an earlier paper, the self-information values of certain dinucleotide associations in the DNA of the coliphage Phi-X 174 were shown to be non-randomly distributed [1]. Specifically, graphs of autocorrelation coefficients of the self-information values for dinucleotides separated by a single nucleotide revealed that every third value was significantly higher than the average. No periodicity was detected in similar analyses of the associations of adjacent dinucleotides or those separated by 2 or 3 nucleotides. The periodicity in the first case was destroyed by total randomization or randomization within triplets but some degree of periodicity was observed if only every third nucleotide was randomized.

We have found that a similar ternary autocorrelation periodicity is also clearly evident in Phi-X 174 DNA when a unique but arbitrary index number is assigned to each dinucleotide type rather than the calculated self-information value used in the previous paper. This substitution greatly facilitates the computations required. The availability of libraries of DNA sequences has permitted the extension of these analyses to other natural DNA's. Thus, we have surveyed a number of DNA sequences for their triplet periodicity and find that this property is present to some extent in the filamentous phages, which are all single-stranded in their infective form, but apparently does not occur to the same degree in most other natural DNA sequences of similar length. The origins of this characteristic are examined and a possible cause is suggested.

Methods

DNA sequences for analysis were obtained from the genetic sequence data bank, Computer Systems Division, Bolt, Beranek and Newman, Cambridge, MA maintained at the University of Texas Educational and Research Computer Center. The analyses reported in this paper were run on an IBM PC computer using programs written in BASIC by the authors. Programs were developed and tested in the interpretive mode on short sequences and then compiled for analysis of the natural sequences and their randomized versions. The first 50 autocorrelation coefficients were calculated, displayed on the screen and printed. Analysis of the 5385 bases of Phi-X 174 for dinucleotide correlation required about 11 min-

Reprint requests to Dr. N. Burr Furlong.

Verlag der Zeitschrift für Naturforschung, D-7400 Tübingen
0341–0382/85/1100–0854 $ 01.30/0

utes of computation time. Other sequences analyzed had from 2767 to 7152 bases.

The sequence files used in this paper from the data bank consist of numbered lines containing 60 letters (A, C, G or T) per line. The computer program scans these files and the letters are associated with numerical values (1, 2, 3 or 4, respectively). These nucleotide values are used to create index numbers at each position in the sequence which uniquely correspond to one of the 256 possible combinations of the 16 different dinucleotides. As an example of the indexing calculation, the sequence AATCCGA ... would yield an index number of 81 for AA associated with CC, 56 for AT with CG and 185 for TC with GA. Dinucleotides are designated 1 through 16 for AA, AC, AG, AT, CA, CC, etc., through TT and the index for associations between dinucleotides is derived from the formula:

$$I = 16 \times (D2 - 1) + D1$$

where D1 and D2 are index values for the left and right dinucleotides. The associations indexed in this paper are for dinucleotides separated by a single nucleotide since similar analyses of dinucleotide associations of other types failed to show any obvious periodicity (data not shown). Because of the nature of the autocorrelation calculation, alternative ways of numbering nucleotides, dinucleotides or their associations have little effect on the appearance of periodicities in natural sequences. We made assignments of A, C, G, T as 4, 3, 2, 1 and as 4, 1, 2, 3 in the analysis of Phi-X 174 DNA without observing any major changes in the pattern of autocorrelation coefficients obtained or in the value of the index defined below (data not shown).

Autocorrelation coefficients for the first 50 shifts of the linear array of index numbers corresponding to dinucleotide associations in a given sequence are sufficient to identify short periodicities. In order to summarize the extent to which every third autocorrelation coefficient is elevated with respect to the intervening values, we calculated an interval parameter as follows: starting with the sixth, seventh and eighth coefficients (to avoid the first few which are non-representative), the next 14 values, counting by threes in each case, are averaged. The mean of the two lowest of these is subtracted from the highest to give a relative measure of the extent of the non-randomness associated with triplet periodicity. For randomized sequences that show no periodicity, this interval measure has been found to be less than 0.002; whereas, for native Phi-X DNA, the value is 0.065. To avoid decimals in the values given on the Tables, we arbitrarily multiply this parameter by 100 and define this as the index of periodicity.

Results

Table I lists values of the index of periodicity defined above for various partial randomizations of the Phi-X sequence. The results of substituting randomly selected nucleotides at every third position can be seen to depend on the specific frame chosen. The greatest reduction in the index occurs if positions 1, 4, 7, ... are randomized and the least for 3, 6, 9, ... In the latter case, a lowering of only 12% is seen; whereas, the former involves a reduction of more than 50%. In all of these cases, it should be noted that a periodicity in the graph of the correlation coefficients can be seen quite clearly, *cf.* Fig. 4: 1a, b, c in [1]. When, instead of substituting randomly generated bases, the bases naturally occurring in a given triplet frame are randomly shuffled and the sequence then analyzed, the results in rows 4, 5 and 6 are obtained. These values are essentially similar to the first three in order and magnitude.

The final values on Table I correspond to the results obtained after generation of random base substitutions in more than one triplet frame. In these cases the appearance of the graph does not clearly indicate the extent of residual periodic non-random-

Table I. Index of periodicity values for Phi-X DNA and variants.

Sequence modification	Index*
Natural sequence	6.53
3, 6, 9 ... randomized	5.76
1, 4, 7 ... randomized	3.83
2, 5, 8 ... randomized	3.13
All randomized	0.18
3, 6, 9 ... shuffled	5.57
1, 4, 7 ... shuffled	4.20
2, 5, 8 ... shuffled	3.50
All shuffled	0.06
1, 3, 4, 6, 7 ... randomized	2.79
2, 3, 5, 6, 8 ... randomized	2.34
1, 2, 4, 5, 7 ... randomized	1.45

* This index is a relative measure of the extent of periodicity in the occurrences of dinucleotide associations in the sequence analyzed. See the Methods section for a full definition.

ness in the distribution of the autocorrelation coefficients, *cf.* Fig. 4: 1e, f in [1].

Table II summarizes the results of applying the analysis for dinucleotide periodicity to a variety of other nucleotide sequences. Information on the nucleotide composition and length is included along with the index for these sequences. The symbols used for the sequences on this Table are those used to identify the files in the data bank. The sources of these sequences are as follows: 1)–4) *E. coli* bacteriophages Phi-X 174, M13, FD (strain 478) and G4; 5)–8) *E. coli* plasmids pBR322 (cloning vector), pBR327 (chimeric), pBR329 and PE194; 9)–11) *E. coli* genomic fragments of the alanyl-tRNA synthetase gene, the ATP synthetase operon and the LAC operon; 12) yeast plasmid (2 micron circle); 13) a human Alu type R DNA flanking the delta globin gene; and 14) a human messenger RNA (collagen type 1, pro-alpha-2).

Table II. Analyses of various DNA sequences.

DNA code name	Index	Length	Composition			
			A	C	G	T
1) PHIX174	6.53	5386	1291	1157	1254	1684
2) M13	3.62	6407	1575	1296	1315	2221
3) FD	1.37	6408	1578	1295	1325	2210
4) G4	0.85	5577	1519	1446	1102	1510
5) PBR322	0.68	4363	984	1210	1134	1035
6) PBR327	0.28	3273	721	911	835	806
7) PBR329	0.32	4150	970	1121	1021	1038
8) PE194	0.75	3728	1363	460	721	1184
9) ECOALA5YNB	0.19	2767	665	665	806	631
10) ECOATPOP	0.96	7152	1687	1800	1971	1694
11) ECOLAC	0.81	5804	1275	1514	1594	1421
12) YSTPLSM	0.10	6318	1876	1284	1179	1979
13) HUMALUR	0.17	3161	823	757	484	1097
14) HUMCG1PAT	1.78	2486	529	646	716	595

Table III. Sequences with forced periodicities.

DNA	Length	Composition					Index
		A	C	G	T	%T	
pBR322	4363	984	1210	1134	1035	23.7	0.68
pBR322	4363	932	1150	1072	1209	27.7	1.43
pBR322	4363	886	1088	1011	1378	31.6	4.07
PE 194	3728	1363	460	721	1184	31.7	0.75
PE 194	3728	1292	439	543	1313	35.2	2.20
PE 194	3728	1222	416	655	1434	38.5	4.91
ECOLAC	5804	1275	1514	1594	1421	24.5	0.81
ECOLAC	5804	1209	1437	1513	1644	28.3	2.36

The index values for three of these natural sequences are repeated in Table III along with values obtained when the bases in the second position of each triplet are replaced by bases generated to produce an excess representation of thymine to various extents. This forced non-randomness was designed to simulate a distribution of bases such as that found normally in the DNA of the filamentous phages.

Discussion

Values for the index of triplet periodicity listed in Table I for various permutations on the Phi-X sequence indicate the importance of the asymmetric distribution of bases in the triplet frames of this phage. The numbers of each of the bases in the three frames (listing 1, 4, 7, ...; 2, 5, 8 ... and then 3, 6, 9 ...) are as follows: A = 456, 363 and 471; C = 382, 395 and 380; G = 352, 405 and 497; T = 605, 632 and 447. As noted in our previous paper, thymine makes up 31% of the bases in Phi-X DNA and this excess over parity with the other bases may be a major factor in the appearance of non-random features seen in this and other single-stranded phages. Note that only in the third frame is the proportion of thymine about the same as that of other bases. This fact correlates with the relatively small effect on the index (12% *vs.* 50%) of randomizing this position. The similarity of the results obtained with random shuffling suggests that it is the number rather than the distribution of the bases that is primarily responsible for the triplet periodicity observed.

The index values for the cases in which only one of the three bases is left unaltered from the natural sequence indicate a distinct residue of periodicity still present. The smallest of these double substitution index values is still 8-fold greater than that associated with the completely randomized sequence. It is significant that this lowest value is associated with the case where the frame retained is that in which T is approximately evenly represented.

Values of the index of periodicity obtained from fully randomized sequences were found to vary from 0.00 to 0.20 for sequences having a wide range of compositions. Analyses of other natural DNA sequences show triplet index values significantly higher than the random range for the filamentous phages, two of the plasmids, two of the bacterial genomic fragments and the human mRNA sequence. In many of these sequences the composition information

shows that one of the bases is disproportionately represented; *e.g.*, high T in PHIX174, M13 and FD, low G in G4 or low C in PE194. But this is not true in all cases and other sequences can be seen to have low values despite compositional imbalance; *viz.*, high G in ECOALA or low G in HUMALUR. Thus, we conclude that the type of non-randomness this particular analysis is able to detect depends on the specific placement of bases along a sequence rather than merely their relative numbers.

That the disproportionate occurrence of a given base at a specific position in a triplet frame is capable of producing an elevated value of the periodicity index is evident from the data of Table III. In all cases where the proportion of T was increased by substitution into every third position along one of these sequences, the index value was found to increase. On the basis of such observations, then, we can conclude that the triplet periodicity we originally observed in the DNA sequence of Phi-X 174 most likely arises from an asymmetric distribution of thymine in specific positions within nucleotide triplets. Functionally this situation would correspond to a tendency for the codons of this phage to complete their ambiguous base positions with thymine bases rather than by an equal representation of all four of the possible bases.

[1] N. Burr Furlong and C. F. Beckner, Z. Naturforsch. **37c,** 321–325 (1982).