# Periodicities of Dinucleotide Self-Information Values in ΦX174 DNA

N. Burr Furlong and C. F. Beckner

Department of Tumor Biochemistry, University of Texas System Cancer Center and Tumor Institute, M. D. Anderson Hospital, 6723 Bertner Ave.
The Department of Biochemistry and Molecular Biology, University of Texas Medical School, Houston, Texas 77030

Z. Naturforsch. **37 c**, 321 – 325 (1982); received September 22, 1981

DNA, Self-Information, Bacteriophage, Autocorrelation, Periodicity

The natural DNA sequence of bacteriophage, ΦX174, when analyzed as a "text" of dinucleotides, is shown to display an easily detectable degree of non-randomness by the distribution of values of dinucleotide self-information along the sequence. Self-information corresponding to occurrences of dinucleotides separated by a single nucleotide is found to be somewhat higher than the values which preceed or follow it for every third nucleotide position along the sequence. Consequently autocorrelation coefficients of these values display a strong periodicity and harmonic analysis of the values shows a spike at a value of 3. Self-information autocorrelation periodicity is used as a test of the effect of randomizing portions of the sequence. Any one or two of the three nucleotides in each triplet of the sequence can be chosen at random without losing dinucleotide self-information periodicity except when both the 1st and 3rd nucleotide of all of the triplets in the major ΦX174 protein reading frame are randomized. Periodicity is also lost when sequences are generated by randomizing triplets.

Autocorrelation and harmonic analysis also indicate other less marked periodic features of dinucleotide self-information values of the native sequence; non-random features are suggested at periods of 12, 20 and 24 nucleotides.

## Introduction

Structural, functional and, to some extent, behavioral information is apparently encoded in nucleic acid sequences of living organisms. Cellular decoding of this information is poorly understood except in the case of the specification of amino acids from nucleotide triplets by the translation apparatus. Recent approaches to analyzing more general informational properties of nucleotide sequences have utilized Gatlin's [1] indexes, Markov analysis and autocorrelation measurements. In the first approach Figueroa *et al.* [2, 3] have used base composition, nearest neighbor and higher order nucleotide occurrences to calculate informational indices for phage and viral genomes. Significant deviations from random behavior were found for triplet indices derived from ΦX174 gene sequences.

The second approach was exploited by Garden [4] who found that a third-order Markov chain best fit the sequence of ΦX174 DNA. Trifonov and Sussman [5] concluded from a correlation analysis of certain sequences of eukaryotic DNA's, mRNA's and viral DNA's that dinucleotides occurred non-randomly in chromatin-like DNA's at a regular interval coincident with the helical pitch, i.e. about 10.5 bases. They also observed a non-random distribution of dinucleotides at intervals of 3 bases. Prokaryotic nucleotide sequences manifested the latter periodicity but did not show the 10.5 base regularity. Darius and Grootaers [6] analyzed divergences from random expectation for a number of nucleic acid sequences and genes within these sequences finding that triplet divergences in the normal reading frame were high for most sequences.

In this report, we calculate dinucleotide self-information values for ΦX174 DNA and show that autocorrelation and Fourier transform analyses of these values reveal periodic, non-random dinucleotide occurrences in this sequence.

## Methods

We have followed Wong and Ghahraman [7] in applying self-information as a measure of independence in character sequences. The self-information of a given dinucleotide at a specific position in a sequence is a function of the number of occurrences of that dinucleotide in the neighborhood of the dinucleotides that precede and those that follow. In our analysis, following the sequence for ΦX174

Reprint requests to N. Burr Furlong.
0341-0382/82/0300-0321 $ 01.30/0

DNA given by Sanger *et al.* [8] we used up to four preceding and four following dinucleotides to calculate self-information values for each dinucleotide in a sequence. The specific function calculated is given by:

$$I_k^m = -\log_2 (P^m_{r_{k+m} r_k} \cdot P^m_{r_k r_{k-m}})$$

where $I_k^m$ = self-information of a dinucleotide at position k relative to dinucleotides m single nucleotide spaces to the left or right.

$P^m_{r_{k+m} r_k}$ = number of occurrences of the type of dinucleotide found at k with the type of dinucleotide found m single nucleotide spaces to the right of k divided by the total number of analyzable nucleotides in the sequence.

$P^m_{r_k r_{k-m}}$ = number of occurrences of the type of dinucleotide found at k with the type of dinucleotide found m single nucleotide spaces to the left of k divided by the total number of analyzable nucleotides in the sequence.

This analysis is equivalent to evaluating the "surprisal" values of individual letters in a language text. Letters that are used in a neighborhood of other letters (including periods, commas and spaces) and are not common for the text analyzed have higher self-information values than do letters occurring in normal contexts. Wong and Ghahraman show that such letters have a high syntactic significance in recognizing the words in which they occur. Another way to express this idea is to say that redundant letter associations are more easily guessed than are rare ones.

The "textual" significance of dinucleotides is not yet possible to predict, but non-random features of dinucleotide occurrences might well have functional implications. To identify possibly periodic features associated with dinucleotide self-information measures, autocorrelation values for these sets of numbers were calculated as the sets were shifted from 1 to 100 positions along the sequence. Graphs of correlations at each of these intervals are presented in the next section. In order to identify features of a particular sequence that may have contributed to the periodicities found in the autocorrelation analyses, we generated sequences in which nucleotides at various positions were randomized by standard computer techniques.

An additional analysis for periodic occurrences of self-information values was calculated using the harmonic function (Fourier transform):

$$H_j^m = \left( \sum_{k=1}^{N} I_k^m \cdot \cos \frac{2\pi k}{j} \right)^2 \cdot \left( \sum_{k=1}^{N} I_k^m \cdot \sin \frac{2\pi k}{j} \right)^2 \cdot N^{-2}$$

where $H_j^m$ is the harmonic coefficient corresponding to $I_k^m$ values for a period of $j$.

The harmonic coefficients will be maximal for those $j$'s corresponding to regular intervals along the sequence having self-information values of about the same magnitude. Since the reinforcements or cancellations encountered in calculating these functions vary over many orders of magnitude, they are plotted as the logarithm of $H$ *vs* $j$.

The computer programs to accomplish these analyses were developed by the authors in BASIC, tested on a Wang 2200 minicomputer and then modified as necessary for running on the University of Texas Education and Research Computer which is a Control Data Corporation CYBER 174. The Wang was also used as a telecommunications terminal to the central facility. For quick visualization the output plots of these values were generated on the printer by scanning the sets of values and, where these numbers fell between increments on the vertical scale, asterisks were printed above the corresponding position shift values on the horizontial scale. Listings for these programs may be requested from the Senior author.

## Results

ΦX174 is a DNA sequence containing a total of 5375 nucleotides [8]: 1286 A, 1158 C, 1251 G and 1680 T. The 16 dinucleotide occurrences (in alphabetical order) are 387, 262, 253, 380, 257, 225, 268, 403, 327, 345, 252, 326, 312, 321, 476 and 569. The occurrences of A, C, G and T in positions 1, 4, 7, 10 ... are 381, 343, 474 and 594; in positions 2, 5, 8, 11 ... are 515, 417, 388 and 472 and in positions 3, 6, 9, 12 ... are 390, 398, 389 and 614. Note the high proportion of T (31.2% *vs* 25% random), TT (10.6% *vs* 9.7% from T occurrences) and of T in "3rd" positions (34.2% *vs* 25% if random).

Dinucleotides in the sequence for ΦX174 DNA were analyzed for self-information regularities; the autocorrelation graphs for immediate ($m = 1$), second ($m = 2$), third ($m = 3$) and fourth ($m = 4$) dinucleotide neighbors are shown in Fig. 1. It is im-

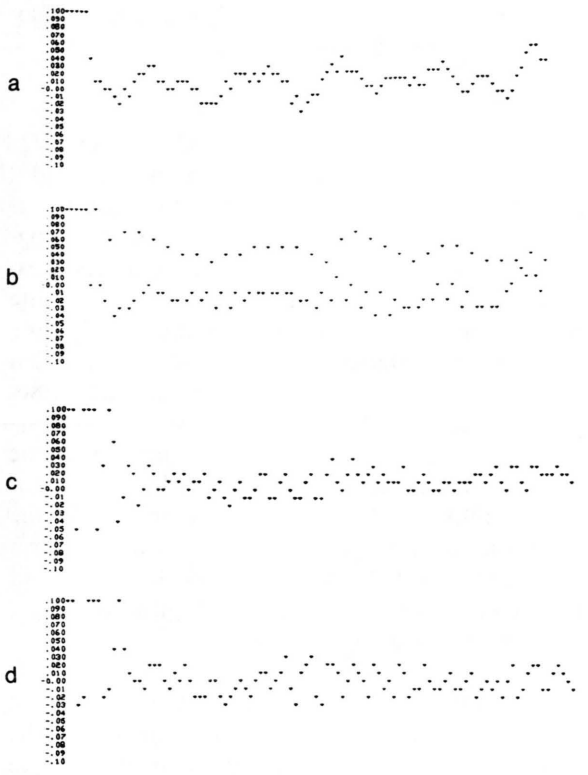


Fig. 1. Autocorrelation values of dinucleotide self-information for the sequence of ΦX174 DNA. Self-information values were calculated for dinucleotides relative to neighboring dinucleotides: a) immediately adjacent, b) separated by a single nucleotide, c) separated by two nucleotides and d) separated by three nucleotides. The ordinates are autocorrelation values; abcissas represent the shift of the sequence relative to itself. Data are given for shifts up to and including 100.

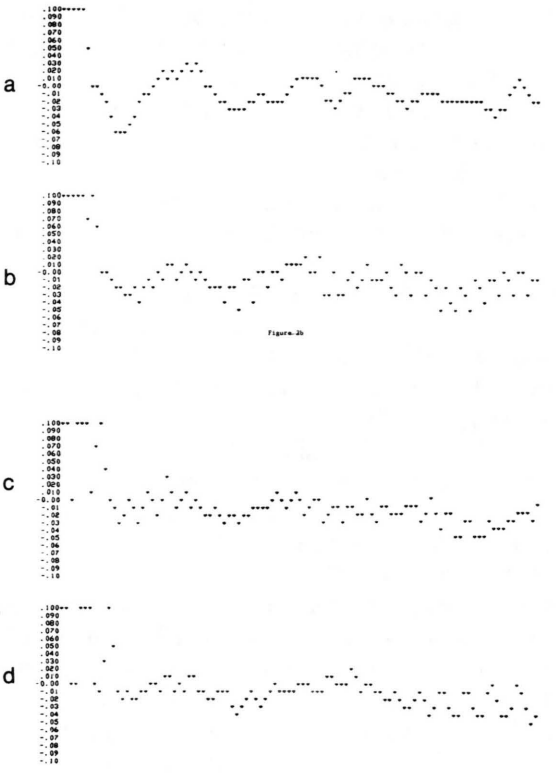


Fig. 2. The same analyses as those given in Fig. 1, but applied to a randomized sequence with base composition identical to that of ΦX174.

mediately obvious that when $m = 2$, *i.e.* for dinucleotides separated by a single nucleotide, higher values are found every third autocorrelation. For comparison, plots are presented in Fig. 2 for a randomized sequence identical in base composition to that of ΦX174. The "damped since wave" periodicity that appears in the graphs of Fig. 2 arises from the nature of the autocorrelation calculation and emphasizes the need for caution in the interpretation of data of this type. In the $m = 1$ and $m = 2$ plots of autocorrelation values from the natural sequence there can be seen a regularity with a period of about 20 nucleotides that is not damped out with successive autocorrelation shifts. Plots of harmonic coefficients for these self-information

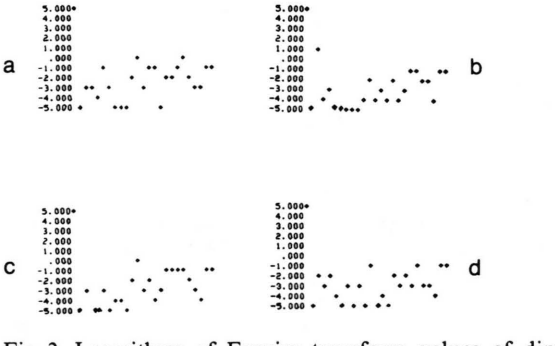


Fig. 3. Logarithms of Fourier transform values of dinucleotide self-information values as described for Fig. 1; *i.e.*, (a) immediate neighbors, (b) 1 nucleotide apart, (c) 2 nucleotides apart and (d) 3 nucleotides apart. Values are given only for periods up to and including 25.

values are given in Fig. 3. Again for $m = 2$, the high value for a period of 3 is obvious. In addition, as seen especially in the $m = 1$ plot, intervals of 6, 12, 20 and 24–25 have elevated harmonic function values and regions around 9, 16 and 22 have especially low values.

We have been interested in examining the nucleotide occurrences that contribute to the strong "triplet" correlation, *i. e.* for dinucleotides separated by a single nucleotide (Fig. 1 b; Fig. 3 b). Nucleotides at specific positions in the ΦX174 sequence were randomized and the resulting altered sequences were analyzed for autocorrelations of the new sets of dinucleotide self-information values. Results (only for $m = 2$) are graphed in Fig. 4 and include the following randomizations: (a) of nucleotides 1, 4, 7, 10 ..., (b) of 2, 5, 8, 11 ..., (c) of 3, 6, 9, 12 ..., (d) of 1, 2, 4, 5, 7, 8 ..., (e) of 1, 3, 4, 6, 7, 9 ... and (f) of 2, 3, 5, 6, 8, 9 ... These graphs show that some extent of regular "triplet" autocorrelation is evident after any single nucleotide position is randomized and even after two positions are randomized unless both the initial and final nucleotides of each triplet are altered. Additional sequences were generated and analyzed in which randomizations by triplets in each of the three reading frames were carried out.

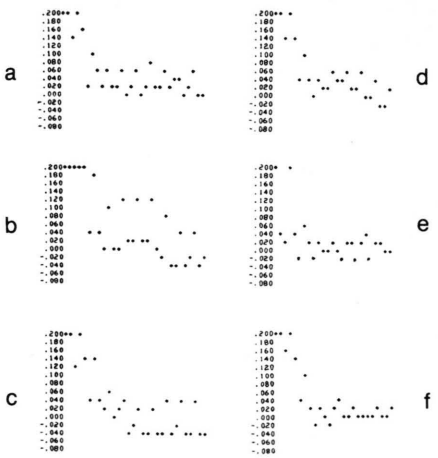


Fig. 4. Autocorrelation values of dinucleotide self-information values derived from ΦX174 DNA sequences in which nucleotides at various positions were randomized: a) every third nucleotide was randomly selected starting with position 1, b) same but starting with position 2, c) same but starting with position 3, d) every third nucleotide starting with position 3 is unchanged; all others are randomly selected, e) same as d but starting with position 2 and f) same as d but starting with position 1. Only the first 25 shift coefficients are graphed in each case.

Dinucleotide periodicities were not evident in any of these analyses (data not shown).

## Discussion

Autocorrelation analysis for periodic properties of a sequence of discrete values can provide valid identification of regular features in the sequence if intrinsic regularities of the analysis itself are recognized. The analysis of randomized sequences illustrates that autocorrelation values exhibit a "white noise" type of variation with a period, in this case, of 15–20 nucleotides. This phenomenon masks a variation of similar periodicity in the natural sequence over about the first 50 shift positions. However, since this artifact is rapidly damped, periodic behavior that is seen in the autocorrelation data for further shifts probably reflects a real property of the natural nucleotide sequence. The Fourier transform of the data also helps to identify regular features of natural sequences by displaying a "spike" at integer values corresponding to their periods.

We conclude from our analysis that the self-information of dinucleotides along the sequence of ΦX174 indicates a strong non-randomness in the occurrence of dinucleotides separated by a single nucleotide. We must remember that this self-information is a mutual measure, *i. e.*, it is normalized to the total number of occurrences of each dinucleotide in the neighborhood of other dinucleotides occurring at a distance of one nucleotide *on either side.* The probabilities involved in these self-information values are probabilities of occurrences of dinucleotide *pairs* (there are 256 different types of pair associations). A dinucleotide at a position with a high self-information value has fewer occurrences in a neighborhood of the type at that position when compared to all of the occurrences of that dinucleotide with its one-nucleotide-removed neighbors throughout the full sequence. A nucleic acid sequence that is designed primarily to be translated into proteins would be expected to have somewhat more restricted occurrences of the first two nucleotides in each triplet codon relative to the occurrence of such dinucleotides as the final nucleotides of a codon or as nucleotides overlapping two codons. Thus, it may not be too surprising that the autocorrelation graphs of dinucleotide self-information values indicate this relative restriction.

The extent to which this triplet regularity persists despite various randomizations was not anticipated,

however. Higher correlation values for dinucleotide self-information of one-nucleotide-removed neighbors were still evident after choosing random nucleotides for substituting in the 1st, 2nd or 3rd position of every triplet in the sequence and even after randomly substituting both the 1st and 2nd *or* 2nd and 3rd nucleotides in each triplet. Autocorrelation was lost when both the 1st and 3rd nucleotides were randomized and also when nucleotides were randomly permuted within triplets. As noted above, the number of occurrences of each nucleotide in the three positions of triplets throughout the sequence of ΦX174 shows greatest inequality for third positions and next for first positions, whereas, there is an almost equal likelihood of each of the four bases occurring in second positions. Randomizations in which these inequalities are lost are associated with a diminution or loss of periodicity in autocorrelation values. The shift of triplets of nucleotides (with unchanged order) among various places in the total sequence did not alter the inequality of base occurrences at positions among triplets; however, this alteration did destroy autocorrelation regularity. Thus, we conclude that the inequality of numbers of individual bases at fixed positions within triplets may be necessary but not sufficient for periodicity of dinucleotide self-information values. The specific dinucleotides found in the codons of the natural sequence must occur with non-random frequencies relative to occurrences of these dinucleotides in other associations in the entire sequence.

The third-order Markov chain best fit for ΦX174 DNA reported by Garden [4] probably arises from the same non-randomness responsible for self-information regularities, as does the 3-period correlation of dinucleotide occurrences seen in the analysis of viral sequences by Trifonov and Sussman [5] and the $D_3$ divergences from randomness reported for ΦX174 by Figueroa *et al.* [3], and by Darius and Groataers [6]. The "white noise" inherrent in autocorrelation analyses masks the demonstration of longer periods but our analyses of ΦX174 sequences, native and altered, have provided evidence for some regularity in the self-information values with a period from 15 to 20 nucleotides since the pattern persists up to shifts beyond which "noise" would have been expected to be "damped out". The Fourier transform results tend to support this conclusion with "spike" values beyond the region of white noise components. For example, dinucleotide self-information data at $m = 1, 2, 3$ and 4 all tend to display higher harmonic function values at 12, 20 and 24 shift positions. The significance of these indications has not been examined but we are now evaluating the effect of various randomizations on these patterns to identify critical dinucleotide associations. Of course there is no reason to assume that the textual significance of DNA sequences is restricted to periodically occuring features. A logical extension of our studies is the search for correlations between information measures and regions of sequences associated with biological or chemical functions. We are also applying our procedures to selected DNAs and other nucleotide sequences of various types.

*Acknowledgements*

This research was supported, in part, by Grant G-120 from the Robert A. Welch Foundation, by an NIH grant, CA 11430, to the Biomathematics Department of The University of Texas System Cancer Center and by the Graduate School of Biomedical Sciences of the University of Texas Health Science Center. Valuable contributions were made by Mark and Jim Pretorius during preceptorships in their medical programs and by Bart Sheinberg during a tutorial laboratory in the Graduate School of Biomedical Sciences. The authors are also grateful to Dr. Stuart Zimmerman and Dr. Dennis Johnson for many helpful suggestions.

[1] L. Gatlin, Information Theory and the Living System, 66–68 Columbia Univ., Press, New York 1972.
[2] R. Figueroa, A. Sepulveda, M. A. Soto, and J. Toha, J. Theoret. Biol. **74,** 203–207 (1978).
[3] R. Figuora, A. Sepulveda, M. A. Soto, and J. Toha, Z. Naturforsch. **32 C,** 850–854 (1977).
[4] W. Garden, J. Theor. Biol. **82,** 679–684 (1980).
[5] E. N. Trifonov and J. L. Sussman, Proc. Nat. Acad. Sci., USA **77,** 3816–3820 (1980).
[6] P. Darius and J. L. Grootaers, Arch. Int. de Physiol. Biochem. **87,** 1047 (1979).
[7] A. K. C. Wong and D. Ghahraman, Informat. Sci. **8,** 173–88 (1975).
[8] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Carlson, J. C. Fiddes, C. A. Hutchison, III, D. M. Slocombe, and M. Smith, Nature **265,** 687–695 (1977).