# Spectrum and symbol distribution of nucleotide sequences

Vera Afreixo, Paulo J. S. G. Ferreira,* and Dorabella Santos
*Departamento de Electrónica e Telecomunicações/IEETA, Universidade de Aveiro, 3810-193 Aveiro, Portugal*

(Received 26 March 2004; published 23 September 2004)

This paper explores the connection between the size of the spectral coefficients of a nucleotide or any other symbolic sequence and the distribution of nucleotides along certain subsequences. It explains the connection between the nucleotide distribution and the size of the spectral coefficients, and gives a necessary and sufficient condition for a coefficient to have a prescribed magnitude. Furthermore, it gives a fast algorithm for computing the value of a given spectral coefficient of a nucleotide sequence, discussing periods 3 and 4 as examples. Finally, it shows that the spectrum of a symbolic sequence is redundant, in the sense that there exists a linear recursion that determines the values of all the coefficients from those of a subset.

DOI: 10.1103/PhysRevE.70.031910 PACS number(s): 87.10.+e, 02.90.+p

## I. INTRODUCTION

Consider a symbolic sequence $s$ with elements

$$s(0), s(1), \ldots, s(N-1),$$

where each symbol $s(i)$ belongs to a finite alphabet $\Gamma$. Although the results described in the paper apply for other finite alphabets, we will limit ourselves to the case $\Gamma = \{A, C, G, T\}$. This is of course motivated by DNA sequences, which consist of adenine ($A$), cytosine ($C$), guanine ($G$), and thymine ($T$).

There is a great deal of interest in methods for detecting and revealing structure (short and long range correlations, periodicities, and so on) in DNA sequences. The availability of DNA data, their volume, and the interest of the applications are some of the motivations to consider the problem. For a review of (spectral and other) methods, sequence modeling, and some related biological issues, see [1] and references therein. Among other things, [1] discusses strand symmetries, the power-law $1/f^{\alpha}$ decay of the spectrum, where $\alpha$ is close to unity, and the similarity between the base correlations ($A$ and $T$, as well as $G$ and $C$).

Fourier analysis is one of the natural tools to use, but the symbolic nature of the data poses a challenge: there is no underlying algebraic structure (such as group structure). However, this problem can be circumvented in several ways. One of the most common approaches depends on the indicator sequences of $s$, that is, numeric (binary) sequences that express the position of each symbol along $s$. For example, the indicator sequence for the symbol $A$ with respect to the symbolic sequence $s$ is

$$i_a(j) = \begin{cases} 1, & s(j) = A, \\ 0 & \text{otherwise}. \end{cases}$$

The Fourier spectrum of the sequence $s$ is defined in terms of the four individual spectra of the indicator sequences [2–6]. More precisely, the $N$ Fourier coefficients $S(k)$ $(0 \leqslant k < N)$ are

$$S(k) = |I_a(k)|^2 + |I_c(k)|^2 + |I_g(k)|^2 + |I_t(k)|^2, \qquad (1)$$

where $I_a$, $I_c$, $I_g$, and $I_t$ are the discrete Fourier transforms (DFTs) of the indicator sequences, that is,

$$I_a(k) = \sum_{j=0}^{N-1} i_a(j) e^{-i2\pi jk/N} \qquad (2)$$

(the $i$ in the exponent denotes the imaginary unit).

The purpose of this paper is to clarify the connection between the distribution of the nucleotides in the sequence, at certain positions, and the size of the spectral coefficients $S(k)$. Contributions toward this goal have been made before [7], but our results differ from those in several aspects. We give a much simpler way of obtaining necessary and sufficient conditions for a specific spectral coefficient to have a prescribed magnitude, not necessarily zero. The method also leads to insight regarding what makes the spectral coefficients large. We give a very fast algorithm for computing a selected coefficient $S(k)$, such as $S(N/3)$. This is of interest since the size of $S(N/3)$ is important to distinguish coding regions from noncoding ones. Finally, we find that the spectral coefficients are redundant (linearly dependent). For example, if there are $n$ symbols $A$ in a sequence of total length $N$, then the $N$ elements of $I_a$ can be determined from any contiguous subset of cardinal $n$ through a linear recursion.

## II. RESULTS

It follows from Eq. (2) that $I_a(0)$ is the total number of occurrences of the symbol $A$ in the sequence. Also, $I_a(N/2)$ is the difference between the number of ocurrences at even positions and at odd positions. This suggests a study of the relations between the size of a spectral coefficient such as $I_a(N/n)$ and the distribution of the corresponding symbol in $n$ evenly spaced subsequences. Assume therefore that $N = nm$, and consider

$$y_a(v) = \sum_{u=0}^{n-1} i_a(um+v) \ (0 \leqslant v < m).$$

In words, split the original sequence of length $N$ into $n$ sequences of length $m$, and then add them together to obtain a

*Electronic address: pjf@det.ua.pt

length $m$ sequence. Note that $y_a(v)$ represents the number of occurrences of the symbol $A$ at the $n$ locations

$$v, v+m, v+2m, \ldots, v+(n-1)m.$$

For this reason we call $y_a$, and the sequences $y_c$, $y_g$, and $y_t$ that are similarly defined, symbol counts. The next step is to evaluate the DFT of the length $m$ symbol count $y_a$:

$$\begin{aligned}
Y_a(k) &= \sum_{j=0}^{m-1} y_a(j) e^{-i2\pi jk/m} \\
&= \sum_{j=0}^{m-1} \sum_{u=0}^{n-1} i_a(um+j) e^{-i2\pi jk/m} \\
&= \sum_{j=0}^{m-1} \sum_{u=0}^{n-1} i_a(um+j) e^{-i2\pi(um+j)k/m} \\
&= \sum_{p=0}^{N-1} i_a(p) e^{-i2\pi pk/m} \\
&= \sum_{p=0}^{N-1} i_a(p) e^{-i2\pi pnk/N} \\
&= I_a(nk).
\end{aligned}$$

The idea of reducing the computation of a length $N=mn$ DFT to the computation of a block sum and a length $m$ DFT is due to [8]. It was rediscovered by [9] (see also [10]).

The following theorem, which links the length $N$ DFT $S$ and the DFTs of the symbol counts $y_a$, $y_c$, $y_g$, and $y_t$ (of length $N/n$) can now be easily shown.

*Theorem 1*. Let $N=nm$. Consider any fixed integer $k$, satisfying $0 \leqslant k < m$. The spectral coefficient $S(nk)$ is given by

$$S(nk) = |Y_a(k)|^2 + |Y_c(k)|^2 + |Y_g(k)|^2 + |Y_t(k)|^2, \tag{3}$$

and as a result it vanishes if and only if $Y_a(k)$, $Y_c(k)$, $Y_g(k)$, and $Y_t(k)$ all vanish.

For the proof, note that Eq. (3) follows immediately from Eq. (1) and the transformation described above.

The computation of $S(nk)$ is much more efficient when done via the symbol counts (and DFTs of length $N/n$) than when done directly (with a DFT of length $N$).

The special case $N=3n$ turns out to be particularly interesting in connection with DNA sequences. It leads to a very fast algorithm for computing $S(N/3)$.

*Theorem 2*. Let $N=3n$. The spectral coefficient $S(N/3)$ is given by

$$S(N/3) = F(y_a) + F(y_c) + F(y_g) + F(y_t), \tag{4}$$

where

$$F(y_a) = \left[ y_a(0) - \frac{y_a(1)+y_a(2)}{2} \right]^2 + \frac{3}{4}[y_a(1) - y_a(2)]^2 \tag{5}$$

and $F(y_c)$, $F(y_g)$, and $F(y_t)$ are similarly defined.

To see that this is true, set $k=1$ and $n=N/3$ in Eq. (3). This leads to

$$S(N/3) = |Y_a(1)|^2 + |Y_c(1)|^2 + |Y_g(1)|^2 + |Y_t(1)|^2. \tag{6}$$

But

$$Y_a(1) = y_a(0) + y_a(1)w + y_a(2)w^2,$$

where $w = e^{-i2\pi/3}$. An elementary computation confirms that $|Y_a(1)|^2$ is $F(y_a)$. The three remaining cases are of course similar, and the result follows.

*Theorem 3*. Let $N=3n$. The spectral coefficient $S(N/3)$ is a symmetric function of the symbol counts.

To simplify the notation, let $x=y_a(0)$, $y=y_a(1)$, and $z=y_a(2)$. Note that $F(y_a)=x^2+y^2+z^2-xy-xz-yz$, and that this is invariant under permutations of the $x$, $y$, and $z$. The remaining three cases are identical, and the theorem follows.

The number of arithmetic operations needed to evaluate Eq. (4) is $O(1)$, that is, independent of $N$. The computation of the symbol counts can be done as the data are read, as it merely requires incrementing the appropriate counters. The overall result is a computational procedure that is $O(N)$, but dominated by the time needed to read the data. Recall that the computation of a fast Fourier transform of length $N$ is an $O(N \log N)$ process, and that the computation of one spectral coefficient requires $O(N)$ arithmetic operations.

Still in reference to the special case $N=3n$, note that it easily yields a necessary and sufficient condition for the vanishing of $S(N/3)$, contained in [7]. The following terminology is convenient: a symbol count, such as $y_a$, is *equidistributed* when all the $y_a(i)$ are equal.

*Theorem 4*. Let $N=3n$. The spectral coefficient $S(N/3)$ is zero if and only if the symbol counts $y_a$, $y_c$, $y_g$, and $y_t$ are equidistributed.

The simplest proof follows from Eq. (4). Its four terms are similar to Eq. (5), which vanishes if and only if $y_a(1)=y_a(2)$ and $y_a(0)=[y_a(1)+y_a(2)]/2$, that is, $y_a(0)=y_a(1)=y_a(2)$.

Alternatively, we see from Eq. (6) that $S(N/3)=0$ if and only if the DFTs of the symbol counts vanish for $k=1$. But

$$Y_a(1) = y_a(0) + y_a(1)w + y_a(2)w^2,$$

where $w=e^{-i2\pi/3}$. Equidistribution implies $Y_a(1)=0$, since $1+w+w^2=0$, because $w^3=1$. Conversely, assume that $Y_a(1)=0$. An elementary computation shows that $y_a(0)=y_a(1)=y_a(2)$, as required. The remaining three DFTs can be handled in a similar way.

The first part of this argument generalizes easily for $m$ not necessarily equal to 3, since $1+w+w^2+\cdots+w^{m-1}=0$ for $w^m=1$. We thus see that the equidistribution of the $m$ symbol counts implies the vanishing of the spectral coefficient $S(N/m)$.

The method used also characterizes the nucleotide distributions that lead to a coefficient $S(N/3)$ of a given size (not necessarily zero). Since $S(N/3)$ is given in Eq. (1), we will look at the component $|I_a(k)|^2$ only.

*Theorem 5*. Identify the symbol counts $y_a(0)$, $y_a(1)$, and $y_a(2)$ with three orthogonal axes $x$, $y$, $z$. Geometrically, $|I_a(N/3)|^2=v$ for some fixed $v$ if and only if $x$, $y$, and $z$ lie on a cylinder with axis $x=y=z$ and radius $r^2=2v/3$.

For the proof, rewrite $F(y_a)$ as

$$F(x,y,z) = \frac{1}{2}[(x-y)^2 + (y-z)^2 + (x-z)^2],$$

or, in matrix form,

$$F(x,y,z) = [xyz]\begin{bmatrix} 1 & -1/2 & -1/2 \\ -1/2 & 1 & -1/2 \\ -1/2 & -1/2 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

Upon reducing the quadratic form to diagonal form in the new variables $X$, $Y$, and $Z$, we get

$$\frac{3}{2}(X^2 + Y^2) = v,$$

a circular cylinder with radius $\sqrt{2v/3}$. An elementary computation shows that the axis is indeed the line $x=y=z$, completing the proof.

The following case is also of interest, and is also an easy consequence of Eq. (3) and Theorem 1.

*Theorem 6*. Let $N=4n$. The spectral coefficient $S(N/4)$ is given by

$$S(N/4) = F(y_a) + F(y_c) + F(y_g) + F(y_t),$$

where

$$F(y_a) = [y_a(0) - y_a(2)]^2 + [y_a(1) - y_a(3)]^2,$$

and $F(y_c)$, $F(y_g)$, and $F(y_t)$ are similarly defined.

For the proof, set $k=1$ and $n=N/4$ in Eq. (3), and evaluate the squared modulus of the length 4 DFTs.

*Theorem 7*. Let $N=4n$. The spectral coefficient $S(N/4)$ is zero if and only if the symbol counts satisfy $y_a(0)=y_a(2)$ and $y_a(1)=y_a(3)$, and similarly for $y_c$, $y_g$, and $y_t$.

The proof is obvious when we consider $F(y_a)=0$, say.

In [7], the case $S(N/4)=0$, with $y_a(0)=y_a(2)\neq y_a(1) =y_a(3)$, is called a "hidden periodicity" of period 4 ("hidden" since it cannot be detected by spectral analysis). We note that this is in fact a periodicity of period 2. Recall that $I_a(N/2)$ is the difference between the total number of occurrences of $A$ at the even and at the odd locations. In terms of the symbol counts for $N=4n$, this means that

$$I_a(N/2) = [y_a(0) + y_a(2)] - [y_a(1) + y_a(3)].$$

But since by hypothesis $I_a(N/4)=0$, the quantities $y_a(0) +y_a(2)$ and $y_a(1)+y_a(3)$ will be different, and $I_a(N/2)\neq 0$. Beware that the statistical significance of the magnitude of a given harmonic should be judged against that in a random sequence, and that the periodicities should in general be identified not with a single harmonic, but with the sum of the magnitudes of sets of equidistant harmonics [11].

We now show that the $N$ spectral coefficients of a symbolic sequence are redundant (linearly dependent). More precisely, if there are $n$ symbols $A$ in a sequence of total length $N$, then the $N$ elements of $I_a$ can be determined from any contiguous subset of cardinal $n$ through a linear recursion. Similar assertions are valid for the three other symbols, but we consider the indicator sequence $i_a$ and its DFT $I_a$ only, since the remaining three indicators can be handled in the same way.

*Theorem 8*. Let $i_a$ be an indicator sequence of length $N$ with $n$ ones, located at $\{j_0, j_1, \ldots, j_{n-1}\}$. The $N$ spectral coefficients $I_a$ given by Eq. (2) satisfy the recursion

$$I_a(\ell + n) = -\sum_{k=0}^{n-1} h_k I_a(\ell + k),$$

where the $h_k$ are the coefficients of the polynomial

$$P(z) = \sum_{k=0}^{n} h_k z^k,$$

determined by $h_n=1$ and

$$P(e^{-i2\pi j_p/N}) = 0 \ (0 \leqslant p < n).$$

For the proof, consider the $n$ equations

$$P(e^{-i2\pi j_p/N}) = \sum_{k=0}^{n} h_k e^{-i2\pi j_p k/N} = 0 \ (0 \leqslant p < n).$$

Multiplication of each of them by

$$i_a(j_p)e^{-i2\pi j_p \ell/N},$$

followed by summation over $p$ and an interchange of the summation order, yields

$$\sum_{k=0}^{n} h_k \sum_{p=0}^{n-1} i_a(j_p)e^{-i2\pi j_p(\ell+k)/N} = \sum_{k=0}^{n} h_k \sum_{p=0}^{N-1} i_a(p)e^{-i2\pi p(\ell+k)/N}.$$

Combination of this with Eq. (2) shows that

$$\sum_{k=0}^{n} h_k I_a(\ell + k) = 0,$$

and since $h_n=1$ we finally get the result.

Note that, in general, any $N$ arbitrarily prescribed real or complex numbers are the spectral coefficients of a certain time series. Theorem 8 shows that this is not the case with the spectral coefficients of a symbolic sequence. If $A$ appears $n$ times out of $N$, then only $n$ contiguous spectral coefficients can be arbitrarily prescribed. The remaining $N-n$ are determined by the recursion given.

## III. CONCLUSION

We have clarified the connection between the size of the spectral coefficients $S(k)$ and the distribution of the nucleotides in certain equispaced subsequences. We gave a much simpler way of obtaining necessary and sufficient conditions for a specific spectral coefficient to have a prescribed size (including zero), using a block sum transformation. The method also leads to insight regarding what makes the spectral coefficients large. We gave a computational procedure for computing a selected coefficient $S(k)$, such as $S(N/3)$, which has been used to tell coding regions from noncoding ones. Finally, we showed that the spectral coefficients are linearly dependent and therefore redundant: If there are $n$ symbols in a sequence of total length $N$, then the $N$ elements of $I_a$ can be determined from, say, $I_a(0), I_a(1), \dots, I_a(n-1)$, through a linear recursion. This dependence could be used, for example, to check for and even to correct errors in the data corresponding to a given indicator sequence [12].

## ACKNOWLEDGMENTS

The present work was supported by the FCT, Grant Nos. POCTI/BME/39030/2001 and POSI/CPS/38057/2001.

[1] W. Li, Comput. Chem. (Oxford) **21**, 257 (1997).
[2] D. Anastassiou, IEEE Signal Process. Mag. **18**, 8 (2001).
[3] M. de Sousa Vieira, Phys. Rev. E **60**, 5932 (1999).
[4] S.-L. Nyeo, I.-C. Yang, and C.-H. Wu, J. Biol. Syst. **10**, 233 (2002).
[5] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, CABIOS, Comput. Appl. Biosci. **13**, 263 (1997).
[6] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).
[7] W. Lee, and L. Luo, Phys. Rev. E **56**, 848 (1997).
[8] C. G. Boncelet, Jr., IEEE Trans. Acoust., Speech, Signal Process. **34**, 1658 (1986).
[9] D. Bourchard, R. Schmidt, and H. Waller, Mech. Syst. Signal Process. **6**, 483 (1992).
[10] P. J. S. G. Ferreira, Mech. Syst. Signal Process. **7**, 191 (1993).
[11] V. R. Chechetkin, and A. Y. Turygin, J. Theor. Biol. **175**, 477 (1995).
[12] P.J.S.G. Ferreira, in *Signal Processing for Multimedia*, edited by J. S. Byrnes (IOS Press, Amsterdam, 1999), p. 35.