Available online at www.sciencedirect.com

SCIENCE DIRECT®

Digital Signal Processing 14 (2004) 523–530

Digital Signal Processing

www.elsevier.com/locate/dsp

Fourier analysis of symbolic data: A brief review [☆]

Vera Afreixo, Paulo J.S.G. Ferreira [*], Dorabella Santos

Departamento Electrónica e Telecomunicações/IEETA, Universidade de Aveiro, 3810-193 Aveiro, Portugal

Available online 11 September 2004

Abstract

We overview and discuss several methods for the Fourier analysis of symbolic data, such as DNA sequences, emphasizing their mutual connections. We consider the indicator sequence approach, the vector and the symbolic autocorrelation methods, and methods such as the spectral envelope, that for each frequency optimize the symbolic-no-numeric mapping to emphasize any periodic data features. We discuss the equivalence or connections between these methods. We show that it is possible to define the autocorrelation function of symbolic data, assuming only that we can compare any two symbols and decide if they are equal or distinct. The autocorrelation is a numeric sequence, and its Fourier transform can also be obtained by summing the squares of the Fourier transform of indicator sequences (zero/one sequences indicating the position of the symbols). Another interpretation of the spectrum is given, borrowing from the spectral envelope concept: among all symbolic-to-numeric mappings there is one that maximizes the spectral energy at each frequency, and leads to the spectrum.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Symbolic data; DNA; Fourier analysis; Correlation; Spectrum

1. Introduction

Protein and DNA data can be written as strings of symbols $s = (s_k)_{0 \leqslant k < n}$ taken from a finite alphabet. For simplicity, we will tacitly consider the alphabet $\{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$, although

[☆] Partially supported by the FCT, grants POCTI/BME/39030/2001 and POSI/CPS/38057/2001.

[*] Corresponding author.
E-mail address: pjf@ieeta.pt (P.J.S.G. Ferreira).
URL: http://www.ieeta.pt/~pjf.

1051-2004/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.dsp.2004.08.001

the results can also be applied to other finite alphabets. The wide availability of such data, their volume and the interest of the applications have lead to a great deal of interest in methods for detecting and revealing structure such as short and long range correlations, periodicities, and so on.

Fourier or linear transform analysis are among the natural tools for this task, but the symbolic nature of the data poses new challenges. It is true that Fourier analysis is possible and quite useful in finite fields and even in groups, Abelian or noncommutative. But even in these cases, the group structure provides an underlying algebraic framework that can be totally absent in the case of symbolic data.

The computation of Fourier and other linear transforms of symbolic data, without assuming any underlying algebraic structure, is a problem that has already been faced in the context of DNA data, and several apparently distinct and unrelated approaches exist.

The simplest solution is to map each symbol to a number. The difficulty with this approach is the dependence on the particular labeling adopted. Consider, for example, the following symbolic periodic sequence:

$$s = (\mathcal{A}\mathcal{T}\mathcal{A}\mathcal{G}\mathcal{A}\mathcal{C}\mathcal{A}\mathcal{T}\mathcal{A}\mathcal{G}\mathcal{A}\mathcal{C}\ldots).$$

The mapping

$$\mathcal{A} \mapsto 1, \quad \mathcal{T} \mapsto 0, \quad \mathcal{G} \mapsto 0, \quad \mathcal{C} \mapsto 0,$$

leads to a numeric sequence of period two, whereas the mapping

$$\mathcal{A} \mapsto 1, \quad \mathcal{T} \mapsto 2, \quad \mathcal{G} \mapsto 3, \quad \mathcal{C} \mapsto 4,$$

yields period six. This clearly shows that some of the relevant harmonic structure can be hidden (or exposed) by the symbolic-to-numeric labelling.

To achieve the required label invariance, each symbol s_k can be assigned a vector v_k pointing in one of four different directions [7]. Each vector can be expressed using an orthonormal basis in three-dimensional space. For each frequency, the sum of the squared modulus of the three corresponding Fourier coefficients provides a measure of spectral content with the required invariance properties [7]. It turns out that this can also be regarded as the Fourier transform of the inner-product autocorrelation of the vectors v_k.

Another approach is based on the symbolic autocorrelation concept. It can be defined in a very natural way and leads to a numerical sequence, the Fourier transform of which is the spectrum of the symbolic data. No hypothesis are necessary regarding the symbols. No algebraic structure or ordering needs to be imposed. Given two symbols, all that is necessary to know is whether they are equal or distinct.

A symbolic sequence can also be represented using indicator sequences, that is, binary (zero/one) sequences that indicate the positions of the symbols in the symbolic sequence. The Fourier spectrum of each indicator sequence can be numerically computed, and the total Fourier spectrum of the symbolic sequence is defined as the sum of the squared modulus of the individual indicator sequence spectra [9].

The concept of spectral envelope [8] provides yet another approach to the Fourier analysis of symbolic data. It is a symbolic to numeric mapping, optimized to emphasize any periodic feature present in the data. A modification of this concept to nonstationary data can be found in [10].

The purpose of this paper is to give an overview of methods for the Fourier analysis of symbolic data, such as DNA sequences, emphasizing their mutual connections. We show, for example, that the spectrum obtained using the symbolic autocorrelation is the sum of the squared Fourier transforms of the binary indicator sequences, which is computationally easier to obtain. We also show that precisely the same spectrum is obtained by adjusting a symbolic-to-numeric mapping, in such a way that the spectral energy at each frequency is maximized. This gives a unifying perspective on these Fourier analysis tools.

2. Methodologies

We begin by introducing the notation and terminology used. Let $s = (s_k)_{0 \leqslant k < n}$ be a symbolic sequence over the alphabet $\{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$. Extension modulo n is implied whenever necessary (hence, $s_{k+n} \equiv s_k$). All sequences of length n, symbolic or numeric, are identified with column vectors of dimension n. The discrete Fourier transform (DFT) $X = (X_i)_{0 \leqslant i < n}$ of the numeric sequence $x = (x_k)_{0 \leqslant k < n}$ is denoted by

$$X = Fx,$$

where F is the $n \times n$ Fourier matrix with elements

$$F_{ab} = e^{-j\frac{2\pi}{n}ab}, \quad a, b = 0, 1, \ldots, n-1.$$

One has

$$\langle x, x\rangle = \|x\|^2 = \frac{1}{n}\|X\|^2, \tag{1}$$

where

$$\langle x, y\rangle = \sum_{k=0}^{n-1} \bar{x}_k y_k = x'y.$$

Here, $\bar{x}_k$ is the conjugate of x_k and x' the conjugate transpose of x. With any symbolic sequence $(s_k)_{0 \leqslant k < n}$ over the alphabet $\{\mathcal{A}, \mathcal{C}, \mathcal{G}, \mathcal{T}\}$ we associate four numeric (binary) indicator sequences u_k^a, u_k^c, u_k^g, and u_k^t, where $0 \leqslant k < n$. These sequences identify the positions of each symbol, that is

$$u_k^a = \begin{cases} 1, & s_k = \mathcal{A}, \\ 0, & s_k \neq \mathcal{A} \end{cases}$$

and similarly in the remaining three cases. We now turn to specific methods that have been used to perform Fourier analysis of symbolic data.

2.1. Indicator sequences

The total spectrum of a symbolic sequence is often defined as the squared modulus of the DFTs of the indicator sequences, that is

$$R_i = \left|U_i^a\right|^2 + \left|U_i^c\right|^2 + \left|U_i^g\right|^2 + \left|U_i^t\right|^2 \tag{2}$$

with $i \in \mathbb{Z}_N$. In the literature, the spectrum is sometimes identified, with little or no explanation, with this sum. Intuitively, the solution seems reasonable. No algebraic operations need to be defined on the symbols, and no symbolic-to-numeric mapping is needed. However, the theoretical interpretation and meaning of this solution seems, at first glance, obscure.

We now turn to the symbolic autocorrelation method, which allows a more satisfactory view of the indicator sequence approach. As we will see, under that formulation the spectrum (2) emerges as the Fourier transform of the symbolic autocorrelation.

2.2. Symbolic autocorrelation

The simplest way of performing Fourier or any other transform analysis on a symbolic sequence is to map the symbols to numbers, and then process the sequence obtained. For example, one could start by finding the autocorrelation of the numeric sequence, and its Fourier transform [3,5]. This has disadvantages, as mentioned above. The mapping may either expose or hide some of the frequency information. Furthermore, there might be no biochemical meaning for the ordering and arithmetic structure that result from the symbolic to numeric mapping.

A better approach is to derive the autocorrelation function directly from the symbols. By autocorrelation of the symbolic sequence $(s_k)_{0\leqslant k<n}$ we mean the numeric sequence

$$r_k = \sum_{i=0}^{n-1} d(s_i, s_{i+k}),$$

where for any two symbols x and y

$$d(x, y) = \begin{cases} 1, & x = y, \\ 0, & x \neq y. \end{cases} \tag{3}$$

This is related to the equal-symbol correlation measure introduced in [9]. The autocorrelation r_k is a numeric sequence that can be rewritten in terms of the four indicator sequences, since $u_k^a = d(s_k, \mathcal{A})$, $u_k^c = d(s_k, \mathcal{C})$, and so on. Hence,

$$r_k = \langle u^a, S_k\, u^a\rangle + \langle u^c, S_k\, u^c\rangle + \langle u^g, S_k\, u^g\rangle + \langle u^t, S_k\, u^t\rangle,$$

where the operator S_k denotes a cyclic shift by k, that is, $S_k u^a = (u^a_{m+k})_{0\leqslant m<n}$, and similarly in the remaining three cases. By (1), we have

$$\langle u^a, S_k\, u^a\rangle = \frac{1}{n}\sum_{i=0}^{n-1} \left|U_i^a\right|^2 e^{-j\frac{2\pi}{n}ki}$$

and so

$$r_k = \frac{1}{n}\sum_{i=0}^{n-1} \left(\left|U_i^a\right|^2 + \left|U_i^c\right|^2 + \left|U_i^g\right|^2 + \left|U_i^t\right|^2\right) e^{-j\frac{2\pi}{n}ki}.$$

We conclude that the DFT of the symbolic autocorrelation is the sum of the squared modulus of the DFTs of the indicator sequences. In other words, we obtain (2), the total spectrum using the indicator sequences.

2.3. Spectral envelope

The concept of spectral envelope was introduced in [8], and a variant for nonstationary data was discussed in [10]. For simplicity the presentation that follows is closer to [10], but considers the stationary case. The relations between the two approaches (and the stationary/nonstationary case) will be discussed later.

Consider the $n \times 4$ matrix

$$u = [u^a \ u^c \ u^g \ u^t]$$

and the vector of real weights

$$w = [a \ c \ g \ t]^T.$$

The sequence $z = uw$ then corresponds to the mapping $\mathcal{A} \mapsto a$, $\mathcal{C} \mapsto c$, and so on. The DFT of z is

$$Z = Fz = Fuw = Uw,$$

where U is the $4 \times n$ matrix obtained by concatenating the DFTs of the indicator sequences,

$$U = Fu = [Fu^a \ Fu^c \ Fu^g \ Fu^t] = [U^a \ U^c \ U^g \ U^t].$$

Denoting by U_i the ith line of U, we may write $Z_i = U_i w$, and so

$$|Z_i|^2 = w'U_i'U_i w = \left|aU_i^a + cU_i^c + gU_i^g + tU_i^t\right|^2. \tag{4}$$

The idea underlying [8] and [10] is to adjust the symbolic-to-numeric mapping in such a way that the Z_i become in some sense extremal. For each frequency i, select the vector w of unit norm that maximizes $|Z_i|^2$. That is, consider the problem

$$\max_{\|w\|=1} |Z_i|^2 = \max_{\|w\|=1} w'U_i'U_i w.$$

The maximum of this Rayleigh quotient is $\lambda_{\max}(U_i'U_i)$, the maximum eigenvalue of the Hermitian matrix $U_i'U_i$. Furthermore, the weights w for which the maximum is achieved are given by

$$w = \frac{U_i'}{\|U_i\|}.$$

As a result,

$$\max_{\|w\|=1} |U_i w|^2 = \max_{\|w\|=1} w'U_i'U_i w = \left|U_i^a\right|^2 + \left|U_i^c\right|^2 + \left|U_i^g\right|^2 + \left|U_i^t\right|^2 = R_i$$

and so we obtain

$$\lambda_{\max}\bigl(U_i'U_i\bigr) = \left|U_i^a\right|^2 + \left|U_i^c\right|^2 + \left|U_i^g\right|^2 + \left|U_i^t\right|^2 = R_i.$$

This reveals yet another way of looking at the total spectrum (2). We have seen that the sum of the squares of the DFTs of the four indicator sequences, at frequency i, is equal to the DFT of the symbolic autocorrelation, at frequency i. Now we see that it is also related to the value of the DFT of a certain numerical sequence, again at frequency i. The particular numerical sequence that leads to this spectrum corresponds to a symbolic-to-numeric mapping optimized to achieve the maximum squared magnitude for frequency i.

One should note that in the nonstationary case, considered in [10], the matrix $U_i'U_i$ is averaged over several data segments, and the solution to the eigenproblem cannot be found as easily. On the other hand, if the averaging is not performed or if the data are stationary, the approach should lead to results close to (2), which might be easier to compute.

We considered the Rayleigh quotient and the corresponding eigenvalue problem mainly for ease of comparison with the general approaches in [8,10]. Such an approach is not strictly necessary in the case discussed. To see this, apply the Cauchy inequality to (4),

$$\begin{aligned} |Z_i|^2 &= \left|aU_i^a + cU_i^c + gU_i^g + tU_i^t\right|^2 \\ &\leqslant \left(|a|^2+|c|^2+|g|^2+|t|^2\right)\left(\left|U_i^a\right|^2+\left|U_i^c\right|^2+\left|U_i^g\right|^2+\left|U_i^t\right|^2\right) \end{aligned}$$

and then note that the condition for equality readily leads to the results.

The spectral envelope was introduced by Stoffer [8] but depends on the concept of generalized eigenvalue. To continue the discussion we need the following notation: $\lambda(A,B)$ denotes a generalized eigenvalue for the problem $Ax=\lambda Bx$. We continue to use $\lambda(A)$ in reference to the eigenproblem $Ax=\lambda x$. Of course $\lambda(A,I)=\lambda(A)$. The idea in [8] is to find the weights w that maximize the power or variance for each frequency, relative to the total variance $V(w)$. This leads to a generalized eigenproblem, since the maximum is of the form

$$\max \frac{w'A'U_i'U_iAw}{w'A'VAw},$$

where, for alphabets of size 4, A is 4×3 and can be taken as

$$A=\begin{pmatrix} I_3 \\ 0 \end{pmatrix}$$

and I_3 is the 3×3 identity matrix. We thus seek $\lambda_{\max}(A'U_i'U_iA, A'VA)$. For a given frequency i, the spectral envelope represents the largest proportion of the total power that can be attributed to that frequency, among all possible symbolic-to-numeric mappings.

It is helpful to compare $\lambda_{\max}(A'U_i'U_iA, A'VA)$, which appears in [8], with $\lambda_{\max}(U_i'U_i)$, which is related to the total spectrum (2). To do that observe that, if D is a positive definite symmetric matrix, then $D=SS'$ with S non-singular, and

$$\lambda_k(C,D)=\lambda_k(S^{-1}CS^{-T},I)=\lambda_k(S^{-1}CS^{-T}).$$

It readily follows from Ostrowski's theorem [4] that

$$\lambda_{\max}(C)\lambda_{\min}(D^{-T})\leqslant\lambda_{\max}(C,D)\leqslant\lambda_{\max}(C)\lambda_{\max}(D^{-T}).$$

Setting $C=A'U_i'U_iA$ and $D=A'VA$, and noting that $\lambda_{\max}(A'U_i'U_iA)=\lambda_{\max}(U_i'U_i)-|U_i^t|^2$, we obtain upper and lower bounds for the required generalized eigenvalue, in terms of $\lambda_{\max}(U_i'U_i)$ and consequently the total spectrum.

2.4. Reduction of the dimensionality

The four indicator sequences are of course redundant, since

$$u^a+u^c+u^g+u^t=1$$

and so

$$U_i^a + U_i^c + U_i^g + U_i^t = \begin{cases} N, & i = 0, \\ 0, & i \neq 0. \end{cases}$$

The total spectrum can therefore be obtained with three DFTs, rather than four. In fact, it is possible to work with three (x, y, z) nonredundant sequences, rather than with four redundant ones [1,6,7]. The assignments used in [1] are

$$\mathcal{A} \mapsto \begin{pmatrix} 0 \ 0 \ 1 \end{pmatrix}, \quad \mathcal{C} \mapsto \left(-\frac{\sqrt{2}}{3} \ \frac{\sqrt{6}}{3} \ -\frac{1}{3} \right), \quad G \mapsto \left(-\frac{\sqrt{2}}{3} \ -\frac{\sqrt{6}}{3} \ -\frac{1}{3} \right),$$
$$T \mapsto \left(\frac{2\sqrt{2}}{3} \ 0 \ -\frac{1}{3} \right).$$

The connection with the indicator sequences is

$$x = \frac{\sqrt{2}}{3}(2u^t - u^c - u^g), \qquad y = \frac{\sqrt{6}}{3}(u^c - u^g), \qquad z = \frac{1}{3}(3u^a - u^t - u^c - u^g).$$

The equivalence between the method that relies on the indicator sequences, and those of reduced dimensionality was shown in [2], for an arbitrary number of symbols. In the present case, we have

$$3\left(\left|U_i^a\right|^2 + \left|U_i^c\right|^2 + \left|U_i^g\right|^2 + \left|U_i^t\right|^2\right) = \begin{cases} 4\left(|X_i|^2 + |Y_i|^2 + |Z_i|^2\right), & i \neq 0, \\ 4\left(|X_i|^2 + |Y_i|^2 + |Z_i|^2\right) - 1, & i = 0. \end{cases}$$

3. Conclusion

We have discussed several methods for the Fourier analysis of symbolic data, emphasizing the case of DNA sequences (four-symbol alphabets). We considered the indicator sequence approach, the vector and the symbolic autocorrelation methods, and methods similar to the spectral envelope, that for each frequency optimize the symbolic-no-numeric mapping to emphasize any periodic data features.

We discussed the equivalence or connections between these methods. We have shown that it is possible to define the autocorrelation function of symbolic data, under weak assumptions: basically we need to be able to tell if two symbols are equal or distinct. The autocorrelation is a numeric sequence, and its Fourier transform leads to the spectrum of the symbolic data. But we have shown that this spectrum can also be obtained by summing the squares of the Fourier transform of indicator sequences (zero/one sequences indicating the position of the symbols). We also examined the spectral envelope concept, which provides yet another interpretation of the spectrum. Among all symbolic-to-numeric mappings there is one that maximizes the spectral energy at each frequency, and leads to the spectrum.

References

[1] D. Anastassiou, Genomic signal processing, IEEE Signal Process. Mag. 18 (4) (2001) 8–20.

[2] E. Coward, Equivalence of two Fourier methods for biological sequences, J. Math. Biol. 36 (1997) 64–70.

[3] M. de Sousa Vieira, Statistics of DNA sequences: A low-frequency analysis, Phys. Rev. E 60 (1) (1999) 5932–5937.

[4] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge Univ. Press, Cambridge, 1990.

[5] W. Lee, L. Luo, Periodicity of base correlation in nucleotide sequence, Phys. Rev. E 56 (1) (1997) 848–851.

[6] L. Luo, W. Lee, L. Jia, F. Ji, L. Tsai, Statistical correlation of nucleotides in a DNA sequence, Phys. Rev. E 58 (1) (1998) 861–871.

[7] B.D. Silverman, R. Linsker, A measure of DNA periodicity, J. Theor. Biol. 118 (1986) 295–300.

[8] D.S. Stoffer, D.E. Tyler, A.J. McDougall, Spectral analysis for categorical time-series: Scaling and the spectral envelope, Biometrika 80 (3) (1993) 611–622.

[9] R.F. Voss, Evolution of long-rang fractal correlations and $1/f$ noise in DNA base sequences, Phys. Rev. Lett. 68 (25) (1992) 3805–3808.

[10] W. Wang, D.H. Johnson, Computing linear transforms of symbolic signals, IEEE Trans. Signal Process. 50 (3) (2002) 628–634.