J. theor. Biol. (2000) **206,** 323–326
doi:10.1006/jtbi.2000.2127, available online at http://www.idealibrary.com on IDEAL®

Fourier and Wavelet Transform Analysis, a Tool for Visualizing Regular Patterns in DNA Sequences

GUY DODIN*†, PIERRE VANDERGHEYNST‡, PATRICK LEVOIR*, CHRISTINE CORDIER* AND LAURENCE MARCOURT*

**ITODYS, associé au CNRS, Université Denis Diderot, Paris, France*
‡Signal Processing Laboratory, Swiss Federal Institute of Technology, CH-Lausanne, Switzerland

Received on 14 *December* 1999, *Accepted in revised form on* 19 *June* 2000

A correlation function that compares each base in a DNA sequence to its various neighbours and which is subsequently processed by Fourier and wavelet transforms has been developed. The procedure has been applied to sequences from the human chromosome 22, to *nef* genes from various HIV clones and to myosin heavy chain DNA. It permits to readily visualize regular features in DNA which are related to the stability of heteroduplexes formed upon strand slippage.

We have designed a procedure that readily and evocatively displays regular patterns in DNA sequences. The search for regularities in nucleotide sequences starts with the construction of a correlation function according to the algorithm of Dodin (Dodin *et al.*, 1996). (Supplementary Material 1 and the formula below.) Briefly, base i is compared with its first upstream neighbour (base $i + 1$) along the sequence, scoring 1 when the two bases are identical and 0 otherwise. The sum of all scores gives $S(1)$. This process is reiterated by successively comparing base i with base $i + 2$ [leading to $S(2)$], then base i with base $i + 3$ and so forth. $S(k)$ appears as a function of the displacement, k. The overall procedure is summarized as follows:

$$S(k) = \sum_{i=F}^{L-1-k} \frac{g_{i,i+1+k}}{(L-F)-k} = \sum_{i=F}^{L-1-k} \overline{g_{i,i+1+k}} \quad \text{with } k = 0, (L-F)-1,$$

where F and L are the numberings of the first and last base in the sequence, respectively. $g_{i,i+1+k}$ is 1 when bases i and $i + 1 + k$ are identical, and is 0 otherwise.

$S(k)$ is then Fourier-transformed in order to sort out its regular components which appear as peaks in the frequency spectrum. However, classical Fourier analysis provides information in the "frequency" domain only, and, besides, it is not always appropriate to show low-frequency signal amplitude modulations (see an example of this in Supplementary Material 2). Valuable information is expected to be found in the "time" domain (meaning in this context the displacement along the signal) such as the distributions along the amplitudes of coefficients attached to the various frequencies in the Fourier spectrum. Indeed, the local values of these parameters are linked to an important macroscopic property of the DNA molecule, namely, the stability of the various heteroduplexes which may form upon strand slippage (see the equation above and the subsequent discussion). Gabor (short-time Fourier transform) or

†Author to who correspondence should be addressed.

0022–5193/00/190323 + 04 $35.00/0

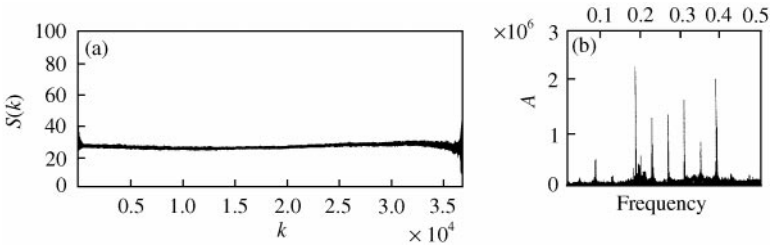

FIG. 1. (a) The correlation function $S(k)$ for clone c:20h12, in the CES region of chromosome 22 (GenBank accession AP000543, Dunham *et al.*, 1999). (b) Fourier spectrum of $S(k)$.

wavelet transforms (which have been successfully used to detect long-range correlations in DNA sequences) (Arneodo *et al.*, 1995, 1996) appeared to be also well adapted to analyse the frequency coefficients along the correlation function as they act as a local "microscope". The overall procedure is made available on the Internet as the DNAcorr. package (Supplementary Material 2).

Regular patterns in DNA sequences appear spectacularly from the correlation function and from its FT or wavelet transforms. A few specific examples are presented. At this stage, no attempt will be made to analyse the causes or to discuss possible biological implications of regular patterns in DNA sequences:

(i) Among the 27 sequences mapping the Cat Eye Syndrome genes in the centromeric region of chromosome 22 (GenBank accessions: AP000522 to AP000547, and AP000365) (Dunham *et al.*, 1999), one (AP000543) presents a highly regular pattern reflecting the numerous repeats in this sequence (Fig. 1). The Fourier spectrum shows frequencies not equal to the reverse of an integer and which, however, can be understood as 1/3, 1/4, 1/5 and 1/16 frequency peaks whose amplitudes are modulated by a low 1/48 frequency.

(ii) The Fourier transform readily unveils the remarkable property of *protein-encoding* sequences of showing a triplet period displayed as a peak at frequency 1/3 (Mani, 1992; Voss, 1992; Chechetkin & Turygin, 1996; Dodin *et al.*, 1996; Lee & Luo, 1997). The 1/3 frequency is not observed in non-coding DNA nor in random synthetic sequences thus making this parameter a convenient criterion for recognizing coding sections in genes. The value of the 1/3 frequency coefficient, $C_{1/3}$, is strongly dependent on the sequence (it spans 2 orders of magnitude over the numerous (>200) sequences we have investigated (sequences from Dodin *et al.*, 1996; Karlin & Burge, 1996; Cheol-Koo *et al.*, 1999; Dunham *et al.*, 1999; Kirchhoff *et al.*, 1999; Toyota *et al.*, 1999; Weiss *et al.*, 1999). The interest of performing the wavelet analysis of the correlation function is to allow exploring the variation of the frequency coefficients in the Fourier transform along the displacement domain and to detect their low-frequency modulation [see, for example, the variation of $C_{1/3}$ along the displacement k (Fig. 2(c) and (d)]. Myosin heavy chain (MyHC) genes present the highest $C_{1/3}$ coefficients in the sequences we have so far investigated (see above), (see, the human β-cardiac MyHC gene, GenBank accession number M58018, open reading frame 87–5894; mouse α-cardiac MyHC, accession M76598; dictyostelium discoideum MyHC, accession M14628, open reading frame 70–6420). In some of the genes of human MyHC isoforms (Weiss *et al.*, 1999), the $C_{1/3}$ coefficient is likely to be slowly modulated by a 1/84 frequency not readily visible in the direct Fourier spectrum but observed in the Fourier transform [Fig. 2(d)] of the wavelet transform of $C_{1/3}$ [Fig. 2(c)] (see gene names and their GenBank accession numbers in Weiss *et al.*, 1999). The causes for the high $C_{1/3}$ coefficient in MyHC is not understood (possibly due to codon bias linked to high gene expression).

(iii) All sequences in a series of *nef* alleles from HIV-1-infected individuals show the typical triplet frequency (Kirchhoff *et al.*, 1999) (GenBank accession numbers AF129 333–AF129 395). A frequency corresponding to a displacement of 33 is also clearly identified in several other sequences. Remarkably, the significantly highest

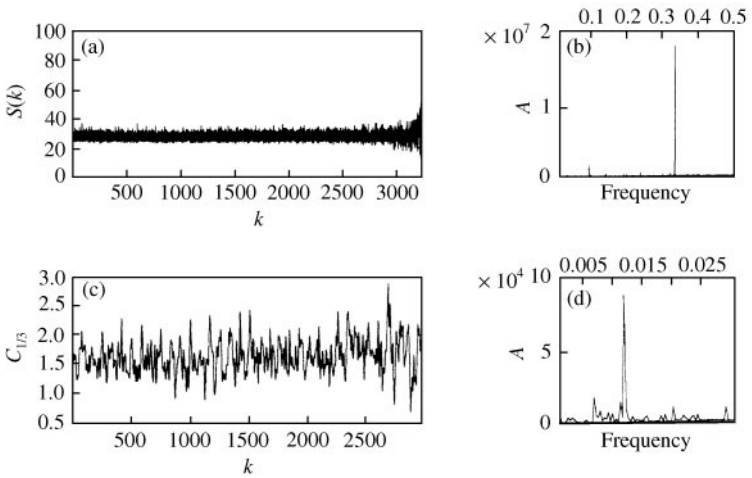

FIG. 2. (a) The correlation function $S(k)$ for the human MyHC-perinatal 3′gene (GenBank accession Y00821, coding sequence: 1–3237, see Weiss *et al*., 1999 for other sequences within the same gene cluster); (b) Fourier transform of $S(k)$ shows a large peak at frequency 1/3 and a smaller one at frequency $0.09 \cong 1/11$; in the cluster significant amplitude of the 0.09 peak is observed in MyHC-IIx/d3′ (GenBank X03740) and in MyHC-IIa 3′ (GenBank S73840) only (not shown); (c) Variation of the amplitude of the coefficient of the frequency 1/3, $C_{1/3}(k)$, with the displacement k analysed by the Morlet wavelet routine in DNAcorr (see Supplementary Material), (d) Fourier transform of $C_{1/3}(k)$ showing the frequency 0.012 corresponding to a displacement of 84.

coefficients associated with the 1/3 frequency are encountered in the three isolates (MB, 105 and, to a lesser extent, 161 in Kirchhoff's study) that present *additional PxxP motifs close to the N terminus* of the coding frame and which correspond to patients with a *progressive disease*. Other isolates with extra PxxP in the variable length region (isolates 027, 165, IP1, FA, RP2) show coefficients statistically similar to those of the other clones. From the analysis of 47 isolates from Kirchhoff's study it is observed that the coefficient of the 1/3 frequency is statistically 20% higher in progressive patients (where rapid progressors show coefficients 10% larger than slow progressors) with respect to non-progressive patients. High triplet power in progressors may reflect active expression of the *nef* protein.

(iv) Examination of the sequences containing CpG islands in a series of clones from colon cancer readily unveils regular patterns not previously detected in this study (besides clones MINT13, 18 and 28, clones MINT3 and 24 exhibit marked regularities) (Toyota *et al*., 1999). (See sequence GenBank accession numbers and sequence nomenclature in Toyota *et al*., 1999.)

It is worth stressing that Fourier and wavelet transforms not only lead to a pictorial representation of regular motifs in DNA sequences but also provide a quantitative link (through the values of the coefficients associated with the various frequencies in the Fourier spectrum) between the occurrence of regular motifs in DNA and the stability of DNA heteroduplexes which would be formed when the two strands in a duplex are stepwise shifted with respect to each other. Indeed, finding statistically the same base at position n and $n + p$ ($p = 3$ for triplet displacement) on one strand is similar to finding the complementary Watson–Crick base on the other strand. The repetitive peptides in surface antigenic proteins in Plasmodium have been tentatively linked to the stabilization of aberrant DNA structures which would lead to protein polymorphism offering the parasite a mechanism to escape the host immune response (Dodin, 1986).

The approach we presented offers a broad scope of application as it provides a tool for rapidly scanning the sequences for regular patterns. It may prove useful in analysing the relationship between repeating triplets and unusual secondary DNA structures likely to be at the origin of several genetic diseases.

REFERENCES

ARNEODO, A., BACRY, E., GRAVES, P. V. & MUZY, J. F. (1995). Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* **74,** 3293–3296.

ARNEODO, A., D'AUBENTON-CARAFA, Y., BACRY, E., GRAVES, P. V., MUZY, J. F. & THERMES, C. (1996). Wavelet based fractal analysis of DNA sequences. *Physica D* **96,** 291–320.

CHECHETKIN, V. R. & TURYGIN, A. Y. J. (1996). Study of correlation in DNA sequences. *J. theor. Biol.* **178,** 205–217.

CHEOL-KOO, L., KLOPP, R. G., WEINDRUCH, R. & PROLLA, T. A. (1999). Gene expression profile of aging and its retardation by caloric restriction. *Science* **285,** 1390–1393.

DODIN, G. (1986). Heteroduplex stability in highly repetitive DNA. *FEBS Lett.* **197,** 5–8.

DODIN, G., LEVOIR, P. & CORDIER, C. (1996). Triplet correlation in DNA sequences and stability of heteroduplexes. *J. theor. Biol.* **183,** 341–343.

DUNHAM, I., SHIMIZU, N., ROE, B. A. & CHISSOE, S. (1999). The DNA sequence of human chromosome 22. *Nature* **402,** 489–495.

KARLIN, S. & BURGE, C. (1996). Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 1560–1565.

KIRCHHOFF, F., EASTERBROOK, P. J., DOUGLAS, N., TROOP, M., GREENOUGH, T. C., WEBER, J., CARL, S., SULLIVAN, J. L. & DANIELS, R. S. (1999). Sequence variation in human immunodeficiency virus type 1 Nef are associated with different stages of disease. *J. Virol.* **73,** 5497–5508.

LEE, W. & LUO, L. (1997). Periodicity of base correlation in nucleotide sequence. *Phys. Rev. E* **56,** 848–851.

MANI, G. S. J. (1992). Long range doublet correlation in DNA and the coding regions. *J. theor. Biol.* **158,** 429–445, 447–464.

TOYOTA, M., HO, C., AHUJA, N., JAIR, K. W., LI, Q., OHE-TOYOTA, M., BAYLI, S. B. & ISSA, J. P. (1999). Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res.* **59,** 2307–2312.

VOSS, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.* **25,** 3805–3808.

WEISS, A., MCDONOUGH, D., WERTMAN, B., ACAKPO-SATCHIVI, L., MONTGOMERY, K., KUCHERPALATI, R., LEINWAND, L. & KRAUTER, K. (1999). Organization of human and mouse skeletal myosin heavy chain gene clusters is highly conserved. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 2958–2963.