BIOINFORMATICS

Vol. 16 no. 10 2000
Pages 932–940

Finding pathogenicity islands and gene transfer events in genome data

Pietro Liò [1,*] and Marina Vannucci [2]

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK and [2]Department of Statistics, Texas A&M University, USA

Received on January 25, 2000; revised on April 26, 2000; accepted on May 15, 2000

Abstract

Motivation: *There is a growing literature on wavelet theory and wavelet methods showing improvements on more classical techniques, especially in the contexts of smoothing and extraction of fundamental components of signals.* G+C *patterns occur at different lengths (scales) and, for this reason,* G+C *plots are usually difficult to interpret. Current methods for genome analysis choose a window size and compute a* χ^2 *statistics of the average value for each window with respect to the whole genome.*

Results: *Firstly, wavelets are used to smooth* G+C *profiles to locate characteristic patterns in genome sequences. The method we use is based on performing a* χ^2 *statistics on the wavelet coefficients of a profile; thus we do not need to choose a fixed window size, in that the smoothing occurs at a set of different scales. Secondly, a wavelet scalogram is used as a measure for sequence profile comparison; this tool is very general and can be applied to other sequence profiles commonly used in genome analysis. We show applications to the analysis of* Deinococcus radiodurans *chromosome I, of two strains of* Helicobacter pylori *(26 695, J99) and two of* Neisseria meningitidis *(serogroup B strain MC58 and serogroup A strain Z2491). We report a list of loci that have different* G+C *content with respect to the nearby regions; the analysis of* N. meningitidis *serogroup B shows two new large regions with low* G+C *content that are putative pathogenicity islands.*

Availability: *Software and numerical results (profiles, scalograms, high and low frequency components) for all the genome sequences analyzed are available upon request from the authors.*

Contact: *p.lio@zoo.cam.ac.uk*

Introduction

The G+C content is an interesting property of genomes, in that the content of different parts of the genome (protein genes, RNA genes and spacers, promoter and regulatory regions, micro- and mini-satellites) reveals different positive linear correlations with the G+C content of their genomic DNA (Mouchiroud *et al.*, 1991). In coding regions the correlation slopes, with respect to the average genomic G+C content, differ among the first, second, and third positions of the codons, depending on their functional importance. Differences in average genomic G+C content among different bacterial species reflect a directional mutational pressure due to misincorporation errors during DNA repair or replication (Sueoka, 1962, 1992). In bacteria, genomic islands with different G+C content with respect to the neighboring regions are generally the result of lateral gene transfer and recombination events. In pathogenic strains, some of the G+C islands carry virulence genes which code for toxins, adhesins, invasins or other virulence factors (Hacker *et al.*, 1997).

Plots of G+C content are extensively used in comparative analysis of complete genomes. Other typical analyses at genome sequence level include the GC skew, i.e. (G+C)/(G-C), the assessment of dinucleotide (δ-difference) and tetranucleotide relative abundance values, the identification of rare and frequent oligonucleotides and the evaluation of codon usage biases (Karlin *et al.*, 1998; Grigoriev, 1998).

Arneodo and collaborators used the continuous wavelet transform (CWT) to analyze long-range correlations associated to G+C patterns in DNA sequences, see Arneodo *et al.* (1998) and references therein. Here we use discrete wavelet methods to detect G+C patterns in genomes. Current methods of detection are based on averages of G+C content in a fixed window. There the window size is chosen based on the sequence length and on the scale of the features to be detected and, consequently, the localization accuracy of the methods is of the order of the chosen window-length (scale). Wavelets, on the contrary, provide a multi-scale representation of signals allowing efficient smoothing and/or extraction of basic components at the different scales. The basic idea of wavelets is to decompose a sequence profile in several groups of coefficients, each group containing information about features of the profile at a scale of sequence length. Coefficients at coarse scales capture gross and global

*To whom correspondence should be addressed.

features. Coefficients at fine scales contain the local details of the profile.

We first apply wavelet shrinkage to eliminate very small variations in the G+C content of regions of genomic sequences. The shrinkage consists in eliminating some of the coefficients at fine scales and reconstructing the profile using mainly the coarser scale coefficients. Donoho and Johnstone (1994) and Donoho *et al.* (1995) have shown that wavelet shrinkage performs better than linear methods such as splines, Fourier series and kernel-based smoothers. The improvement is particularly relevant when a function shows jumps or spikes (Wang, 1995; Raimondo, 1998). Here we use an approach to the wavelet shrinkage proposed by Ogden and Parzen (1996) that uses a change-point statistic.

Having obtained a smoothed profile we then extract and interpret characteristic components of the profile by looking at the scalogram of the wavelet representation. The scalogram, Flandrin (1988) and Chiann and Morettin (1998), is a wavelet analogue of the well-known periodogram for time series analysis, and can be used to locate high and low frequency components. From the point of view of our sequences, a low frequency component can be associated to large genomic regions with different G+C content with respect to the nearby regions. A high frequency component is instead associated to rapid variations in the G+C content that involve a single gene or few genes.

We present results on G+C patterns analysis of *D. radiodurans* chromosome I and compare the genome sequences of two different strains of *H. pylori* (26 695 and J99) and two of *N. meningitidis* (serogroup B, strain MC58, and serogroup A strain Z2491). We detect a set of loci that have very different G+C content and have not previously described. We discuss the presence of these genes under the hypothesis of lateral gene transfer events. Our findings show that wavelet smoothing and scalogram are powerful tools to detect differences within and between genomes and to separate small (gene level) and large (putative pathogenicity islands) genomic regions that have different composition characteristics.

Theory

Wavelet theory and wavelet transforms

Wavelets are well established in the mathematical sciences, with successful applications in signal and image processing, numerical analysis and statistics. For a complete introduction to the mathematical theory of wavelets see Daubechies (1992), among others. See Hirakawa *et al.* (1999) for a biologists' oriented introduction.

In wavelet theory a function is represented by an infinite series expansion in terms of a dilated and translated version of a basic function ψ, called *mother wavelet*, each multiplied by an appropriate coefficient. The wavelet family is defined as $\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$, with j a dilation parameter and k a translation parameter, and the wavelet series representation of a function f is therefore

$$f(x) = \sum_{jk} f_{jk}\psi_{j,k}(x)$$

with wavelet coefficients

$$f_{jk} = \int_{-\infty}^{+\infty} f(x)\psi_{j,k}(x)\,\mathrm{d}x.$$

Coefficients $f_{j,k}$ describe features of f at the spatial locations $2^{-j}k$ and frequency proportional to 2^j (or scale j). Thus, the fundamental idea in wavelet theory is to break a signal down into its components and follow their evolution in the time or space domain. Unlike Fourier bases, wavelets are localized both in time/space and in frequency. This property enables the wavelet series to describe local characteristics of a function in a more efficient way. When representing a signal in a wavelet basis, narrow wavelets will detect sharp features and broader ones more global features.

The discrete wavelet transform (DWT; Mallat, 1989) decomposes a function into its wavelet coefficients. From a computational point of view, it proceeds by recursively applying two convolution functions, known as quadrature mirror filters, each producing an output stream that is half length of the original input, until the resolution level zero is reached. For practical purposes the DWT is often represented in matrix form as Wy with W an orthogonal matrix and y a vector of observations of the signal. An inverse wavelet transform can be also defined. The standard discrete wavelet transform, as the fast Fourier transform, operates on datasets with length 2^N, N integer. When required, data can be padded with zeros. These zeroes do not affect the results.

Wavelet shrinkage

Wavelet shrinkage, Donoho and Johnstone (1994) and Donoho *et al.* (1995) has found useful applications in many different scientific fields, one of the most recent being the analysis of protein structure, see Hirakawa *et al.* (1999) and Lio' and Vannucci (2000). There a signal is observed with an additive Gaussian component

$$\tilde{f}_i = f_i + \sigma \cdot z_i; \qquad i = 1, \ldots, n$$

with $\tilde{f}$ the vector of 'noisy' observations of f and σ the standard deviation of the noise component. The purpose is to recover f. A DWT is applied to the data obtaining a set of empirical coefficients

$$\begin{aligned} W_\psi \tilde{f}_i &= (W_\psi)(f_i + \sigma \cdot z_i) = W_\psi f_i + \sigma(W_\psi z_i) \\ &= W_\psi f_i + \sigma w_i. \end{aligned}$$

Given to the linearity and orthogonality properties of the DWT, the component $w_i = W_\psi z_i$ is again normally distributed and with constant variance. This component can now be removed from the empirical wavelet coefficients by applying a 'hard' or 'soft' threshold technique. Hard threshold functions apply a 'keep or kill' policy, i.e. each coefficient is removed if less, in absolute value, than a threshold value. Soft threshold functions, in addition, shrink the absolute value of coefficients larger than the threshold. An estimate of f is finally reconstructed from the smoothed coefficients by applying the inverse wavelet transform.

Thresholding can be done globally or locally, according to whether one chooses the same threshold value for all coefficients or different values for different groups of coefficients. Here we follow the approach to the wavelet shrinkage of Ogden and Parzen (1996), who, at each scale of resolution, use a hypothesis test procedure to determine if the set of coefficients at that scale behaves as noise or if a significant signal component is present. The maximum of each set of squared wavelet coefficients is tested to see if it behaves as the nth order statistic of a set of independent χ^2. If not, it is kept, and the maximum of the remaining subset is tested, continuing in this fashion until the maximum of the subset is judged not to be significant. The set of 'large' coefficients that 'pass the test' of significance describe significant features of the signal. The level of the hypothesis test, α, controls the smoothness of the resulting wavelet estimate, with small values associated to more smoothed estimates.

Wavelet scalogram

Here we describe the *scalogram*, a wavelet tool that can provide a useful interpretation of the wavelet representation of a signal. The scalogram is defined as a plot of the sums of the squares of the wavelet coefficients at the different levels. It was first introduced by Flandrin (1988), in the context of the continuous wavelet transform, and subsequently by Chiann and Morettin (1998) and Ariño *et al.* (2000), in the context of the discrete transform, as a decomposition of the 'energy' of a function in the time-frequency (scale) plane.

The scalogram is the analogue of the periodogram used in Fourier analysis to detect periodic components of a signal. Different components will result in well visible peaks in the scalograms. These components can be extracted from the signal by spliting the wavelet coefficients into different sets, each set containing those coefficients that belong to the same peak. Low and high frequency components of the signal can then be reconstructed by applying the inverse wavelet transform to the separate sets. Ariño *et al.* (2000) provide helpful details about implementation aspects and also suggest improvements of the spliting of the coefficients for the case of levels that fall in between peaks. To provide the readers with an illustrative example, and following Ariño *et al.* (2000), Figure 1a shows an artificial periodic signal with two components, Figure 1b its wavelet decomposition, Figure 1c the wavelet scalogram, showing two peaks, and Figure 1d the extracted components. Later we will show how the scalogram can be used as a useful measure of comparison of sequence profiles, allowing detection of periodic patterns of the different length-scale.

Implementation

Wavelet smoothing of DNA sequences

We first apply wavelet shrinkage to eliminate variations in the G+C regions of the genomic sequences that are too small to be of interest. Methods based on wavelet transforms generally require a powerful visualization tool. We have used S-Plus (Mathcad) and the S-Plus wavelet module implemented by Nason and Silverman (1994). In order to detect G+C patterns we have coded the DNA sequences in a simple way: C, G = 1; A, T = 0. The signal is padded with zeroes to achieve a total length as a power of two. The analysis of sequences of entire chromosomes may require large amounts of memory. A remedy can be to consider a spliting of the genome in several subsequences using windows of 1024 kb, 2048 kb or more. The ends of the denoised profiles can be joined using interpolation techniques. When using this option special care should be given to the interpretation of the extremes of the smoothed subsequences in that they may be affected by the well known boundary problems of the wavelet transform. Alternatively, the G+C can be preliminary averaged using a very small window of 4–10 bp. From a practical point of view, averaging does not change the accuracy of the localization of pattern boundaries. With this option, DNA sequences of 1 Mb can be processed in less than 5 s on a portable computer. Results of this paper were obtained on profiles averaged over 10 bp.

When applying the discrete wavelet transform (DWT) to G+C profiles we used Daubechies (1992) wavelets with N vanishing moments[†]. These wavelets are the most commonly used in applications. Performances of the different bases vary among different classes of signals, depending on characteristics such as ruggedness, intermittency and variance. It is known that more regular wavelets lead to high compressibility, since the fine scale wavelet coefficients are zero where the function is smooth. On the other hand, the support of the wavelets increases with the degree of smoothness and a trade off with the localization properties is necessary. Limited exploratory analyses (we tried Haar and Daubechies $N = 2, 4, 8, 10$) showed that the Daubechies' basis with $N = 10$ vanishing moments is a good compromise for G+C pattern analysis.

[†] A function f has N vanishing moments if $\int t^q f(t)\,\mathrm{d}t = 0$, $q = 0, 1, \ldots, N-1$.

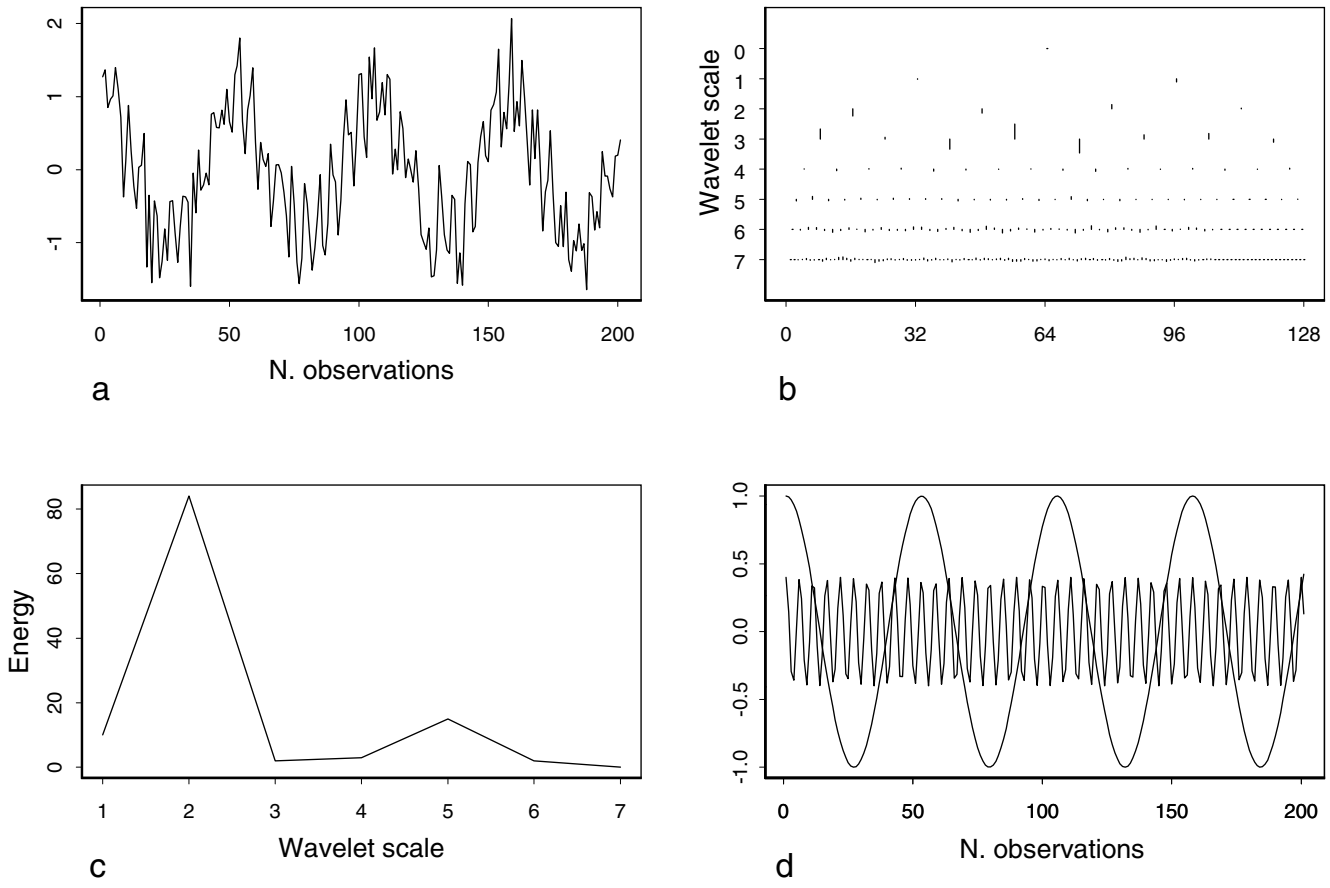

Fig. 1. An artificial periodic signal with two components is generated (**a**); its wavelet decomposition coefficients (**b**), the wavelet scalogram (**c**). The two peaks can be clearly seen. (**d**) shows the smoothed function.

The denoising method we use depends upon the choice of the parameter α. Changing the value of this parameter corresponds to tuning the amount of signal denoising. Following Ogden and Parzen (1996) we have focused on two different values of the parameter. A value $\alpha \sim 0.5$ leads to severe smoothing and to reconstructed profiles that barely show the main features of the original sequences. A value $\alpha \sim 0.999$ leads instead to a more conservative smoothing and to profiles that preserve a large number of local features. We have therefore decided to concentrate on this latter value and used in all analyses $\alpha = 0.999$. Variations of α in the range $0.9-0.999$ do not alter the features of the smoothed sequences. Profiles were zero-averaged before applying the DWT and denoised profiles were rescaled to the average G+C content of the sequence.

Wavelet-based measure of sequence profile comparison

Further analysis of patterns of smoothed profiles can be done by looking at the scalogram of the wavelet representation. If the scalogram shows isolated peaks then we can extract the different components of the signal. A low frequency component will be associated to large genomic regions with different G+C content with respect to the nearby regions. A high frequency component will instead be associated to rapid variations in the G+C content that involve a single gene or few genes. The low and high frequency components can be reconstructed by applying the inverse wavelet transform to wavelet coefficients that belong to different peaks, as previously described.

Results and discussion

Genomes comparison: a G+C *perspective*

We have analyzed the G+C patterns of a set of bacterial genomic sequences with different average G+C content and for which the G+C patterns have not been previously analyzed or have been described at low resolution. This set includes *D. radiodurans* ($G+C\% = 0.67$; White *et al.*, 1999), *H. pylori* strain 26695 ($G+C\% = 0.38$; Tomb *et al.*, 1997) and strain J99 ($G+C\% = 0.38$; Alm *et al.*, 1999), and *N. meningitidis* serogroup A strain Z2491

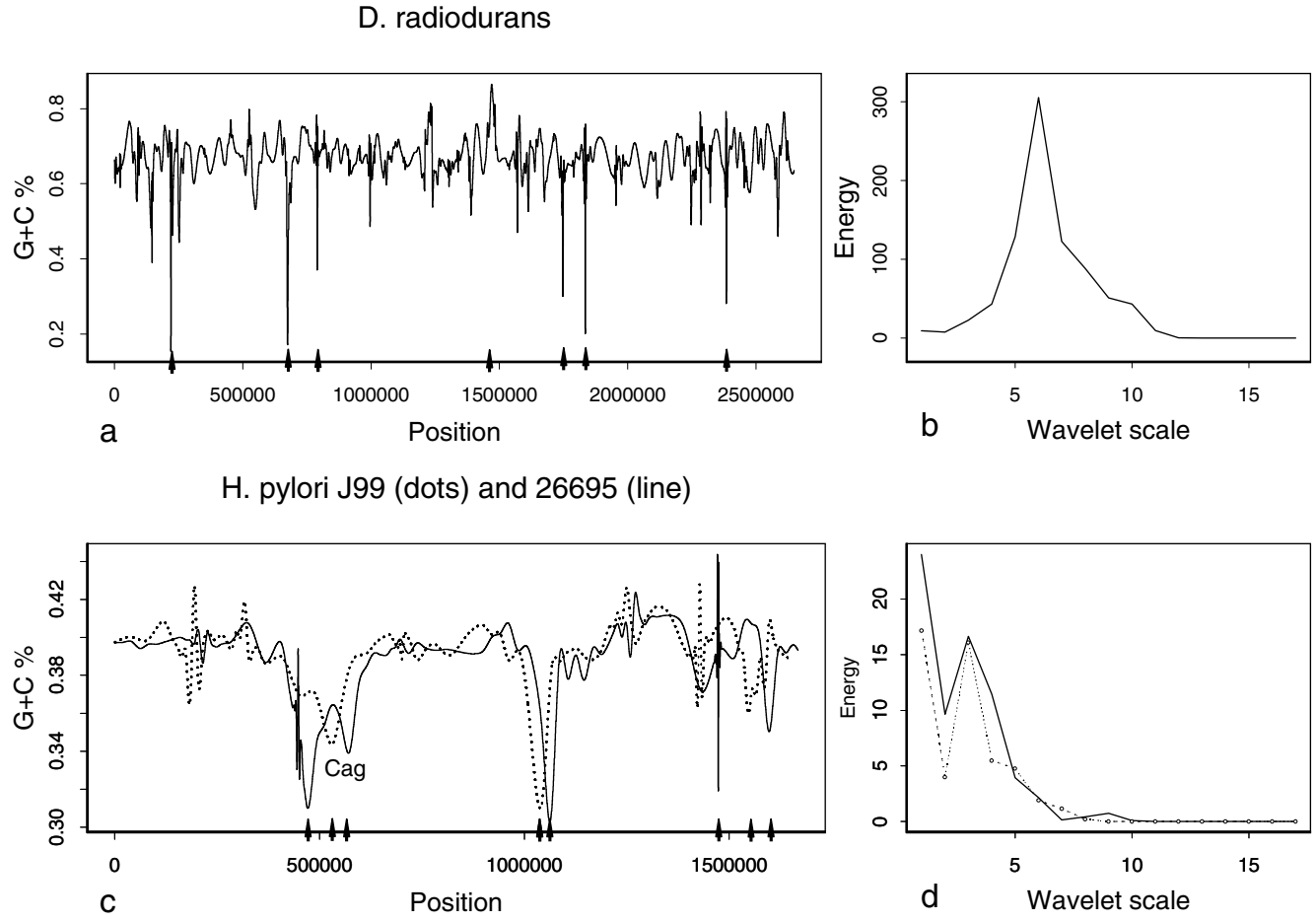

Fig. 2. Wavelet analysis of the G+C profiles of the genome sequences of *D.radiodurans* chromosome I (**a**) and relative scalogram (b), of *H. pylori* (strains 26695 and J99; (**c**) and relative scalograms (**d**). We have used $\alpha = 0.999$; the small arrows on the bottom border show the maxima and minima reported in Table 1; the arrow with empty head is referred in the text.

(G + C% = 51.8%; Parkhill *et al.*, 2000) and serogroup B strain MC58 (G + C% = 51.5%; Tettelin *et al.*, 2000). Figure 2a shows the wavelet denoised profile of *D. radiodurans* chromosome I. All the sequence segments near each maximum and minimum (small arrows) were used as query in sequence database searching using Blast 2.0 (Altschul *et al.*, 1990). We report in Table 1 the most important minima and maxima, their positions, the corresponding locus (gene), the putative function and the G+C content. In the case of *D. radiodurans* we found that all minima and maxima but a dGTPase, DR1808 and a secretory protein, DR0774, are coding regions with unknown functions and with no close homologous in GenBank; thus, no assumption can be made on the selective advantage of these genes. The low G+C content of these loci and the very low G+C content of the minima suggest that a high G+C content is not a necessary requisite for the radiation-resistance and desiccation/starvation recovery of this bacterium. It is worth mentioning that, although there is no correlation between the genomic G+C content and the growth temperature, a correlation has been found between the G+C content of rRNA and tRNA stems and the optimal growth temperature (Garnier and Lobry, 1997). The high transformability of this bacterium (see White *et al.*, 1999, and references therein) and the low G+C content of these loci suggest that they may have been transferred from low G+C content species.

Figures 2c and 2d show the wavelet denoised G+C profile of the genome sequences of *H. pylori* strains 26 695 (line) and J99 (dotted line) and their scalograms. Karlin and collaborators reported the analysis of the G+C plot of *H. pylori* strain 26 695 (Karlin *et al.*, 1998) using a sliding window of 50 kb. The two genomes have very similar smoothed plots. The strain 26 695 contains a low G+C content region (first arrow from the left at the bottom of Figure 2c), located from 454–480 kb, that is not present

Table 1. Position of minima and maxima of Figures 2a, 2c, 3a, 3c (small arrows), the corresponding locus and putative function; the average G+C content; legend: DR, *D. radiodurans* chromosome I; HP26695, *H. pylori* strain 26695; HP99, *H. pylori J99*; NMB, *N. meningitidis* serogroup B strain MC58; NMA, *N. meningitidis* serogroup A strain Z2491

Figure	Position	Put. locus	Pred. funct.	G+C%
2a	219 970	DR0219	unknown	36.9
2a	673 830	DR0664	unknown	34.1
2a	789 690	DR0774	secretory pathway	56.7
2a	1 469 580	DR1461	unknown	76.1
2a	1 747 190	DR1721	unknown	36.2
2a	1 834 180	DR1808	dGTPase	40.0
2a	2 384 560	DR2382	unknown	41.5
2c	471 560	HP0453	unknown	33.8
2c	570 850	HP0539	cag18 (CAG)	35.0
2c	1 062 160	HP0998	transposase	33.5
2c	1 473 200	HP-	close to 23S rRNA	31.8
2c	1 597 550	HP1521	restriction/mod.	36.2
2c	529 480	JHP482	cag11 (CAG)	33.2
2c	1 038 810	JHP937	unknown	35.3
2c	1 547 140	JHP1409	restriction/mod.	36.3
3a	183 950	NMA0200	capsule biosyn.	29.8
3a	300 730	NMA0313	MafB adhesin	36.2
3a	682 160	NMA0690	FhaB adhesin	25.8
3a	769 780	NMA0784	putative membrane prot.	27.9
3a	1 028 820	NMA1076	unknown	24.6
3a	1 528 210	NMA1623	outer membrane prot.	27.5
3a	1 927 250	NMA1987	putative membrane prot.	34.3
3a	2 071 320	NMA2118	unknown	36.4
3c/4b	77 130	NMB0067	capsule biosyn (IHT-A1 in Figure 4a)	29.2
3c/4b	107 490	NMB0098	ABC transporter (IHT-A2 in Figure 4a)	34.4
3c	203 320	NMB0200	unknown	28.9
3c	526 660	NMB0498	pathogenesis (IHT-B in Figure 4a)	25.7
3c	757 550	NMB0726	restric/mod	31.0
3c	1 434 700	NMB1404	unknown (X2 in Figure 4a)	22.7
3c	1 839 620	NMB1761	unknown (IHT-C in Figure 4a)	40.8
3c	1 845 660	NMB1767	unknown (IHT-C in Figure 4a)	20.5
3c/4b	2 122 200	NMB2008	ABC transporter (Y3 in Figure 4a)	29.9
3c	2 230 630	NMB2118	unknown (X3 in Figure 4a)	25.4

in the J99 strain. This region contains two homologues of VirB4 (HP0441, HP0459) that is implicated in bacterial conjugation (Dang *et al.*, 1999), transposases (HP0438, HP0437) and several hypothetical proteins. There are three large low-G+C content regions that are very similar in both *H. pylori* genome sequences. The location of CAG pathogenicity islands in both genomes is shown ('cag' in Figure 2c). The regions 1050–1069 kb (strain 26 695) and 1041–1060 kb (strain J99) contain transposases (HP0988, HP989, HP997, HP998, HP1007), integrase/recombinase proteins (HP0995, JHP941, JHP951) and proteins regulating plasmid functions (HP1000, HP1006). The regions 1581–1613 kb (strain 26 695) and 1535–1574 kb (strain J99) contain loci that are putative restriction/modification enzymes (HP1517, HP1521 and JHP1409); these are often phase-variable genes; outer membrane proteins (HP1512). A peak at 1 473 200 in the strain 26 695 occur in proximity of 23S rRNA genes.

The *H. pylori* and *D. radiodurans* scalograms (Figures 2b and 2d) are rather different in energy and maximum positions. The scalograms of *D. radiodurans* (Figure 2b) has a peak more shifted towards high frequencies than the scalograms of *H. pylori* genomes (Figure 2d) because the signal has more wiggles. The scalograms of *H. pylori* sequences show a large low-frequency component while the high-frequency one seems negligible. Low frequency components are revelatory that there are large genomic regions with different G+C content with respect to the nearby regions.

Figures 3 show the comparison of the wavelet denoised G+C profiles of *N. meningitidis* serogroup A strain Z2491 (Figure 3a) and *N. meningitidis* serogroup B strain MC58 (Figure 3c). Although the two profiles have different minima and maxima they have similar G+C patterns of ribosomal genes (large arrows with empty heads). Table 1 shows that several G+C content minima in both plots

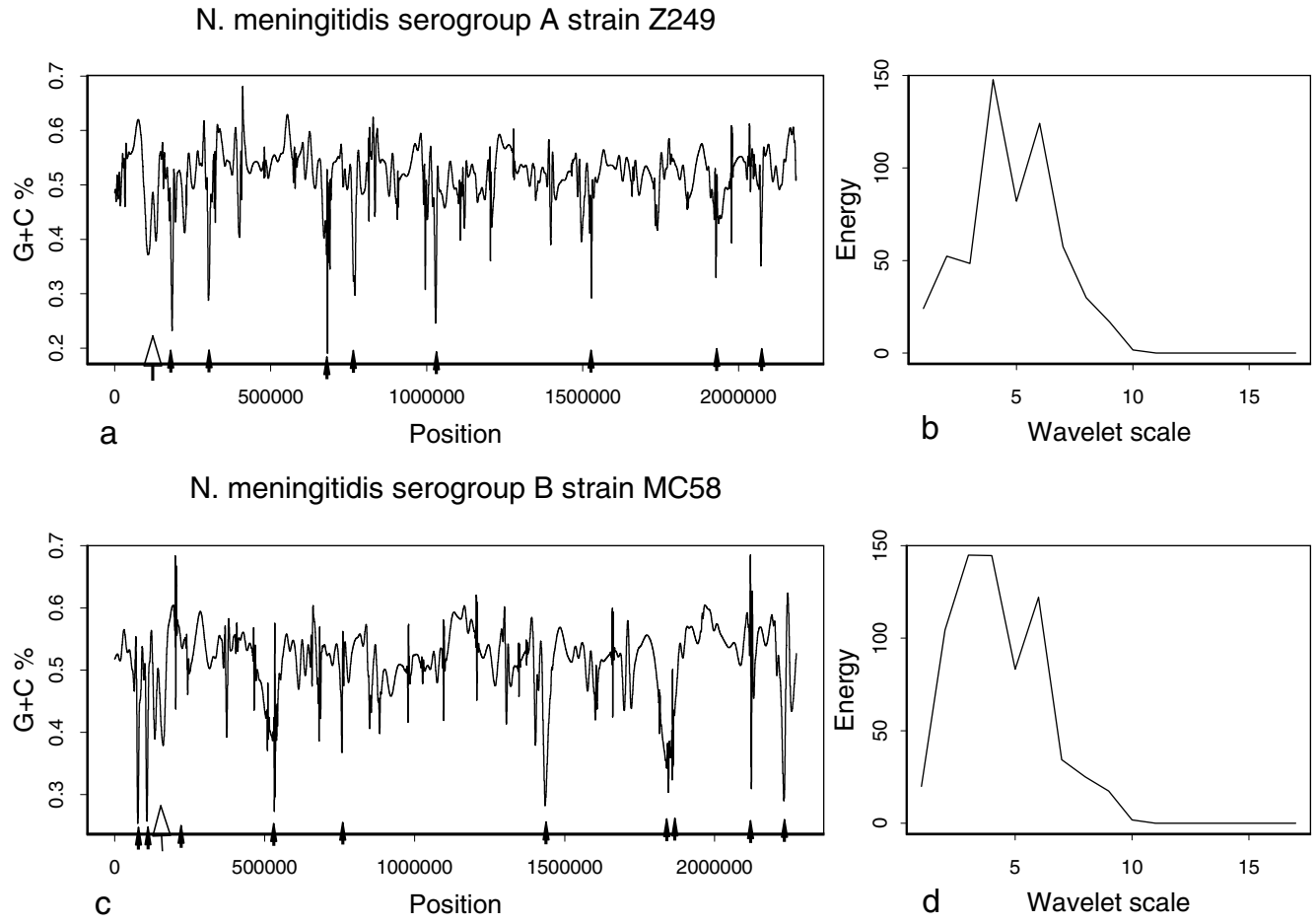

Fig. 3. Wavelet analysis of the G+C profiles of the genome sequences of *N. meningitidis* serogroup B strain MC58 (**a**) and relative scalogram (**b**); of *N. meningitidis* serogroup A strain Z2491 (**c**) and relative scalogram (**d**). We have used $\alpha = 0.999$; the small arrows on the bottom border show the maxima and minima reported in Table 1; the arrows with empty heads correspond to the positions of ribosomal proteins.

occur at loci that have a putative pathogenesis function, such as biosynthesis of capsule (NMB0067, NMA0200), adhesins (NMB0313, NMA0690), phase variability loci (NMB0098, NMB2008) and outer membrane protein (NMA1623).

The scalograms of the profiles of Figures 3a and 3c clearly show two peaks (Figures 3b and 3d, respectively). This may help us in extracting significant components of the smoothed profiles. The scalogram of *N. meningitidis* serogroup B (Figure 3d) shows a low-frequency peak that is bigger and spanning over a larger range of scales than the one of *N. meningitidis* serogroup A: this suggests that *N. meningitidis* serogroup B may have more pathogenicity-related genes than the other genome. Thus, we investigate in more details the profile of *N. meningitidis* B. We show in Figures 4 the plots of low- (Figure 4a) and high-frequency (Figure 4b) components of *N. meningitidis* serogroup B. Tettelin and collaborators (2000) described three main pathogenicity regions in the genome of *N. meningitidis* serogroup B: IHT-A (subdivided in A1, 75–84 kb and A2, 104–109 kb), IHT-B (526–544 kb), IHT-C (1827–1860 kb).

The plot of the low-frequency component of *N. meningitidis* B (Figure 4a) reveals three other large genomic regions with low G+C content that have not yet been described in the literature. Particularly, they do not appear in the low resolution G+C plot shown by Tettelin *et al.* (2000).

- The region X1 (Figure 4a) ranges from 159–163 kb and contains ribosomal genes; these genes have low G+C content also in *N. meningitidis* serogroup A (arrows with empty head in Figures 3a and 3c).
- The region X2 ranges from 1428–1455 kb (Figure 4a) and contains genes that may have pathogenicity functions, as for instance loci coding for an ABC

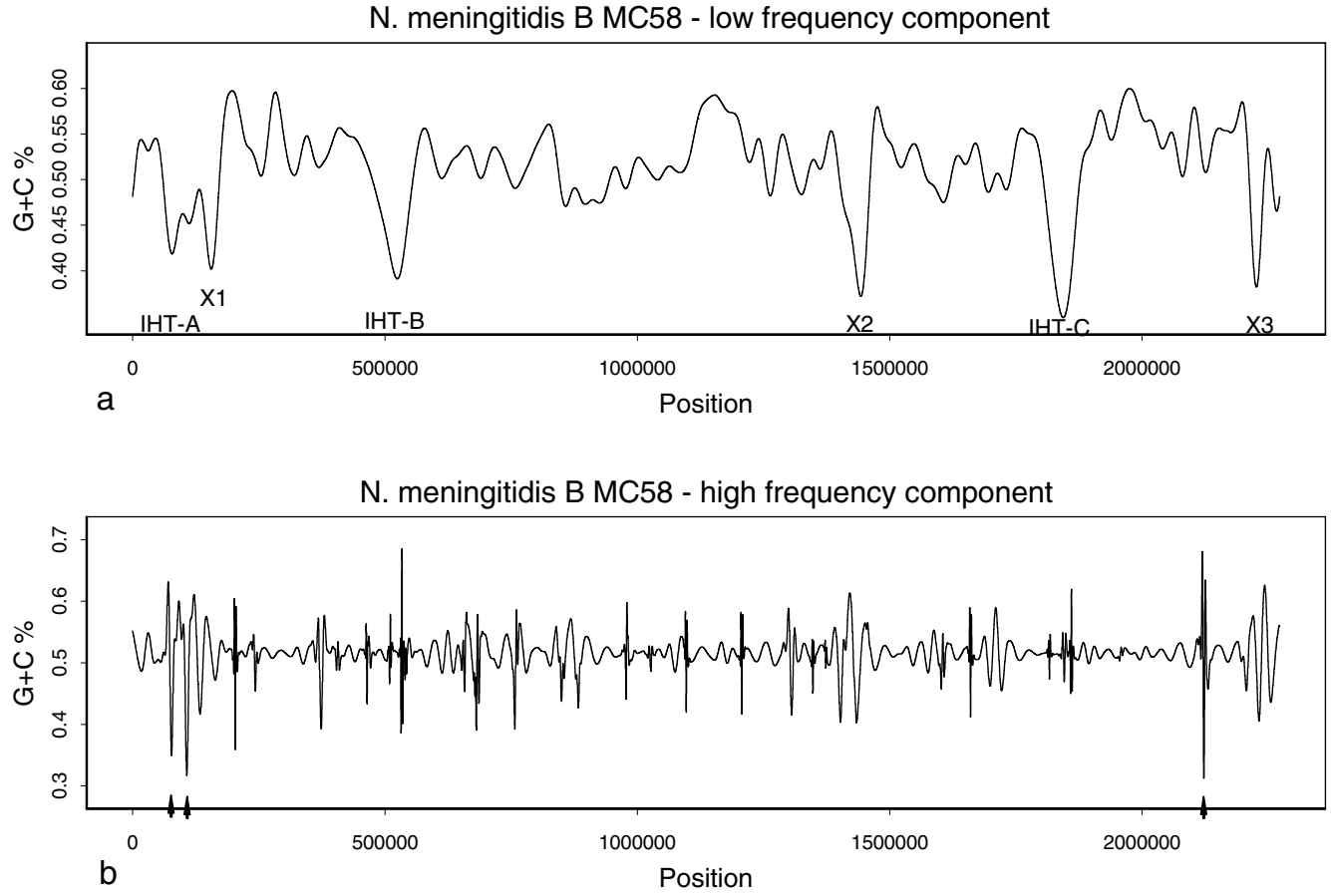

Fig. 4. Low (**a**) and high frequency component (**b**) of the scalogram of *N. meningitidis* serogroup B strain MC58. The small arrows on the bottom border show the maxima and minima reported in Table 1.

transporter family protein (NMB1400), IS1106 transposases (NMB1399, NMB1411), FrpA/C proteins (NMB1403, NMB1405, NMB1407, NMB1409, NMB1412, NMB1414, NMB1415) that are similar to RTX toxins (Thompson *et al.*, 1993); RuvC (NMB1419) that is part of the restriction/modification system (Handa *et al.*, 2000 and references there in) and htrB/msbB (NMB1418) that is involved in membrane lipid A synthesis and can inhibit the complement pathway activation (de Haas *et al.*, 2000). The other proteins of this region have unknown function.

- The region X3 ranges from 2222–2232 kb (Figure 4a) and contains 6 copies homologous to MafB adhesin (NMB2105, NMB2107, NMB2111, NMB2119, NMB2122,) and a IS1016 transposase (NMB2126); all the other proteins have unknown function.

We hypothesize that the X2 and X3 regions may have originated from lateral transfer events and their gene composition suggests that they may be putative pathogenicity islands. Further investigations may require experimental evidences on some of the proteins with unknown function. The low G+C content of the X1 region may also suggest that ribosomal genes cluster can be transferred among close species. Asai and collaborators have recently shown that bacteria can exchange even rRNA operons (Asai *et al.*, 1998). The high frequency component of the *N. meningitidis* B, plotted in Figure 4b, is associated to rapid variations in the G+C content that involve a single gene or few genes. The most important maxima correspond to those of Figure 3c and are reported in Table 1.

Concluding remarks

Papers reporting results of genome sequencing projects usually display low resolution G+C plots and perform a χ^2 statistics of the average value for each window and the whole genome. The method described in this paper is based on performing a χ^2 statistics on the wavelet coefficients of a profile; thus we do not need to choose a

fixed window size, in that smoothing occurs at different scales. The wavelet scalogram seems a powerful tool to detect G+C pattern differences among genomes and to separate the different components of a profile. We have shown that low and high frequency components generated from a scalogram correspond to large (cluster of genes) or small (single gene) regions with different G+C patterns. This method can be applied to other measure of genome composition such as δ-difference, GC skew and purine/pyrimidine patterns.

Acknowledgements

P.L. is supported by an EPSRC/BBSRC Bioinformatics Initiative grant. We thank Julian Parkhill (Sanger Centre) for helpful suggestions.

References

Alm,R.A. *et al.* (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, **397**, 176–180.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ariño,M., Morettin,P. and Vidakovic,B. (2000) Wavelet scalograms and their application in economic time series. Submitted to Communications in statistics. Theory and Practice

Arneodo,A., d'Aubenton, CarafaY., Audit,B., Bacry,E., Muzy,J.F. and Thermes,C. (1998) What can we learn with wavelets about DNA sequences. *Physica A*, **249**, 439–448.

Asai,T., Zaporojets,D., Squires,C. and Squires,C.L. (1998) An *Escherichia coli* strain with all rRNA operons inactivated: complete exchange of rRNA genes between bacteria. *Proc. Natl. Acad. Sci. USA*, **96**, 1971–1976.

Chiann,C. and Morettin,P.A. (1998) A wavelet analysis for time-series. *J. Nonparametric statistics*, **10**, 1–46.

Dang,T.A., Zhou,X.R., Graf,B. and Christie,P.J. (1999) Dimerization of the *Agrobacterium tumefaciens* VirB4 ATP-binding cassette mutations on the assembly and function of the T-DNA transporter. *Mol. Microbiol.*, **32**, 1239–1253.

Daubechies,I. (1992) *Ten Lectures on Wavelets*. SIAM, Philadelphia.

Donoho,D. and Johnstone,I. (1994) Ideal spatial adaption via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho,D., Johnstone,I., Kerkyacharian,G. and Picard,D. (1995) Wavelet shrinkage: Asymptopia? (with discussion). *J. R. Statist. Soc.,* B, **57**, 301–369.

Flandrin,P. (1988) Time-frequency and time-scale, *IEEE Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling*. Minnesota, Minneapolis, pp. 77–80.

Garnier,N. and Lobry,J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.*, **44**, 632–636.

Grigoriev,A. (1998) Analyzing genomes with cumulative skew diagrams (1998). *Nucleic Acids Res.*, **26**, 2286–2290.

de Haas,C.J., van Leeuwen,E.M., van Bommel,T., Verhoef,J., van Kessel,K.P. and van Strijp,J.A. (2000) Serum amyloid P component bound to gram-negative bacteria prevents lipopolysaccharide-mediated classical pathway complement activation. *Infect Immun*, **68**, 1753–1759.

Hacker,J., Blum-Oehler,G., Muhldorfer,I. and Tschape,H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol. Microbiol.*, **23**, 1089–1097.

Handa,N., Ichige,A., Kusano,K. and Kobayashi, (2000) Cellular responses to postsegregational killing by restriction-modification genes. *J. Bacteriol.*, **182**, 2218–2229.

Hirakawa,H., Muta,S. and Kuhara,S. (1999) The hydrophobic core of proteins predicted by wavelet analysis. *Bioinformatics*, **2**, 141–148.

Karlin,S., Campbell,A.M. and Mrazek,J. (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.

Lio',P. and Vannucci,M. (2000) Wavelet change-point prediction of transmembrane proteins. *Bioinformatics* **16**, 376–382.

Mallat,S.G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transaction on Pattern and Machine Intelligence*, **11**, 674–693.

Mouchiroud,D., D'Onofrio,G., Aissani,B., Macaya,G., Gautier,C. and Bernardi,G. (1991) The distribution of genes in the human genome. *Gene*, **100**, 181–187.

Nason,G.P. and Silverman,B.W. (1994) The discrete wavelet transform in S. *J. Comput. Graph. Stat.*, **3**, 163–191.

Ogden,R.T. and Parzen,E. Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Comput. Stat. Data Anal.*, **22**, 53–70.

Parkhill,J. *et al.* (2000) Complete DNA sequence of a serogroup A strain of *Neisseria menigitidis* Z2491. *Nature*, **404**, 502–506.

Raimondo,M. (1998) Minimax estimation of sharp change points. *Ann. Stat.*, **26**, 1379–1397.

Sueoka,N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA*, **48**, 582–588.

Sueoka,N. (1992) Directional mutation pressure, selective constraints and genetic equilibria. *J. Mol. Evol.*, **34**, 95–114.

Tettelin,H. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* Serogroup B Strain MC58. *Science*, **287**, 1809–1815.

Tomb,J.-F. *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.

Thompson,S.A., Wang,L.L. and Sparling,P.F. (1993) Cloning and nucleotide sequence of frpC, a second gene from *Neisseria meningitidis* encoding a protein similar to RTX cytotoxins. *Mol. Microbiol.*, **9**, 85–96.

Wang,Y. (1995) Jump and sharp cusp detection by wavelets. *Biometrika*, **82**, 385–397.

White *et al.* (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*, **286**, 1571–1577.