Sankhyā : The Indian Journal of Statistics
Special issue on Wavelets
2001, Volume 63, Series B, Pt. 2, pp. 218–233

NON-DECIMATED WAVELET ANALYSIS OF BIOLOGICAL SEQUENCES: APPLICATIONS TO PROTEIN STRUCTURE AND GENOMICS

By MARINA VANNUCCI
Texas A & M University, U.S.A
and
PIETRO LIÒ
University of Cambridge, U.K

SUMMARY. Here we investigate the potential of wavelet methods in the analysis of biological sequences as a complement method to those currently available. Specifically, we show how the non-decimated wavelet transforms and the wavelet variance and correlation scale-by-scale decompositions can be used to extract relevant structural features from proteins, such as helices in membrane proteins, to highlight similarities among different amino acid sequences and to detect genomic regions that have different composition characteristics from the nearby regions.

1. Introduction

There has been a long and productive history of interdisciplinary efforts between molecular biology and statistics. Recent genome projects have provided the field with an abundance of data. Nowadays there is evidence that biological data have generally a multi-scale structure, in that even the DNA sequence of the most simple micro-organism has revealed patchiness at different sequence scales, (see Karlin and Brendel (1993) and Liò *et al.* (1996)). Protein structure data, heart-beat patterns, bronchial tree and capillaries networks are other examples of relevant biological signals.

Paper received July 2000.
AMS (1991) *subject classification.* 42C40, 65T60, 92D10, 92D20.
Key words and phrases. Non-decimated wavelet transform, wavelet variance and correlation, membrane proteins, genome analysis, G+C plot.

Two main problems are currently challenging biostatisticians: the determination of the structure of a protein directly from the analysis of its sequence of amino acids, and the detection of patterns in genome sequences that may lead to the discovery of novel functions. Here we focus on two aspects of these problems: the detection of the positions of the helices in proteins inserted into the membrane (membrane proteins) and the extraction of genomic regions with a different base composition that may carry virulence factors in pathogenetic bacteria.

In order to fill the gap between the explosion of available biological data and the relatively slow speed at which experiments can reveal protein structures and untangle structure/function relationships of the different genome portions, hidden Markov models and neural networks have become increasingly popular in secondary structure prediction, (see Rost *et al.* (1996) and Sonnhammer *et al.* (1998)), among others. The analysis of entire genome sequences is usually based on the frequency distribution of DNA 'words' of different lengths. Karlin *et al.* (1998) have analyzed genome sequences in distant species, detecting alien portions of DNA, i.e. portions of genomes transferred from another specie, on the basis of the DNA composition similarities.

Given the multi-scale structure of most biological data, wavelet methods appear to be a natural way to achieve vast improvements in the quality of statistical analyses of such data, Aldroubi and Unser (1996). (Arneodo *et al.* (1996)) used the continuous wavelet transform to analyze long-range correlations associated to G+C patterns in DNA sequences. Discrete wavelet transforms and smoothing techniques were used by Liò and Vannucci (2000a, 2000b) and Hirakawa *et al.* (1999). Here we investigate the potential of non-decimated wavelet transforms as complementary methods. Non-decimated transforms provide a greater flexibility, in that, when decomposing a sequence, features at different scales will be aligned with those of the original data. More importantly, the newly defined wavelet variance and wavelet correlation, (Percival (1995)), allow powerful scale-by-scale decompositions that help the location and extraction of important features of the sequences. We show that these methods allow the detection of the modular units of a protein and provide information on similarities among proteins at different sequence length-scales. In genomics analysis we use wavelet methods to compare entire genomes and to detect DNA segments with different G+C content. We remark here that the novelty of the work described here is not in the methods, although recent and, are well known in the wavelet community, but in showing how the flexibility of these methods can be successfully used in interdisciplinary research to greatly help the understanding of complicated

structures, like proteins and genomes in molecular biology.

The paper is organized as follows. Section 2 provides a description of the biological problems addressed here. Section 3 is a brief introduction to the non-decimated wavelet transforms and to the wavelet variance and correlation scale-by-scale decompositions. Section 4 first describes the analysis of amino acid sequences. Specific task is the location of helices in a human membrane protein and in a bovine cytochrome membrane protein. Results are compared with those obtained with other currently used methods. Then the entire genome sequence of the recently sequenced bacterium *Ureaplasma urealyticum*, (Genbank accession NC_002162, University of Alabama at Birmingham, Glass *et al.* (2000)), is compared with the genome of two other mycoplasma, *Mycoplasma genitalium*, (Genbank accession NC_000908, Fraser *et al.* (1995)), and *Mycoplasma pneumoniae*, Genbank accession NC_000912, Himmelreich *et al.* (1996), and with the genome of the bacterium *Neisseria meningitidis*, serogroup A strain Z2491, (Genbank accession 002203, Parkhill *et al.* (2000)), that contains several pathogenicity regions. Regions that have different G+C content with respect to the nearby regions are located.

2. **Biological Sequences**

A DNA molecule can be described as a string of symbols (bases) from a four letters alphabet: A (adenine), T (thymine), C (cytosine) and G (guanine). The particular order of symbols is called DNA sequence and the overall DNA sequence in a bacterial cell is its genome. The most important portions of the DNA are small strings of variable length, termed genes, scattered all over the genome. Each gene defines the order of concatenation of 20 different small molecules, the amino acids, to form a single protein. While the DNA is the basis for heredity, the proteins are the basic building blocks of the organisms. The amino acid sequence of a protein is called its primary structure. Different regions of the sequence of amino acids form local regular protein secondary structures, such as alpha helices, beta strands or random coils. The tertiary structure is formed by the packing of such structural elements, (Branden and Tooze (1991)).

2.1 *Membrane proteins.* A large portion (10−35%) of proteins in all genome sequences encodes membrane proteins, currently the most biomedically important family of proteins, serving as targets for the majority of pharmaceutical agents. The determination of the structure of membrane

proteins is a very difficult task for nuclear magnetic resonance spectroscopy and for X-ray spectroscopy because these proteins are generally very large and do not crystallize. The amino acid sequences of the membrane proteins are mainly composed of stretches of about 20 predominantly hydrophobic amino acids, i.e. amino acids that have low affinity for water molecules. These stretches form helical segments in the membrane tridimensional space and are mostly connected by polar amino acids forming loops.

A number of algorithms have been designed to identify putative transmembrane helices in the primary amino acid sequence. With the majority of the prediction algorithms the candidate transmembrane segments are identified by some kind of hydrophobicity plot or profile, i.e. a graph showing the local hydrophobicity of the amino acid sequence as a function of the position. To generate a hydrophobicity profile, one has to choose a hydrophobicity scale and an averaging procedure. Profiles are obtained by coding each residue of the amino acid sequence according to a hydrophobicity scale. There are several hydrophobicity scales, some of them based on chemical properties of the amino acids, others on statistical analysis of a data set of membrane proteins. Here we will use the hydropathy scale proposed by Kyte and Doolittle (1982) that is based on a thermodynamic measure (the free energy) of transfer of each amino acid between organic solvent and water. The amino acid profile is subsequently scanned to locate segments of high average hydrophobicity or propensity that can be considered as transmembrane segments. Kyte and Doolittle used a sliding window of around 20 amino acids. Von Heijne (1992) proposed a trapezoid sliding window. Recently neural networks, (Rost *et al.* (1996)) and hidden Markov models, (Sonnhammer *et al.* (1998)) have been used. Liò and Vannucci (2000a) used wavelet shrinkage methods to smooth the profile. Propensity profiles based on hydrophobic or other propensity scales may in fact show sharp changes in correspondence of transmembrane regions.

2.2. *G+C contents and pathogenicity islands in genomes.* In bacteria, genomic regions with a different base composition, in particular differences in the content of guanine and cytosine (G+C content of the sequence) with respect to the neighbouring regions, are generally the result of the transfer of a portion of DNA from other bacterial genomes. The density of guanine and cytosine, the G+C content, along the DNA molecule is an important parameter for both the understanding of the evolution of genomes and the comparison of different genomes, Muto and Osawa (1987). In pathogenic strains, some of these G+C islands carry virulence genes which code for toxins and virulence factors, (Hacker *et al.* (1997)). In order to detect G+C patterns,

profiles are obtained by coding the DNA sequence as C,G=1; A,T=0, for example. Plots of G+C content are extensively used in comparative analysis of complete genomes. Current methods for genome analysis choose a window size and compute χ^2 statistics of the average value for each window with respect to the whole genome (Parkhill *et al.* (2000)). Recently, G+C plots have revealed that, at a scale of $10^5 - 10^6$ bases, genomes of warm-blooded vertebrates and plants are mosaics of long DNA segments, the isochores, with different G+C content, (see for example Hattori *et al.* (2000)). It is worth noticing that a correlation has been found in human genome between G+C rich genomic regions and their replication timing, Tenzen *et al.* (1997).

In Liò and Vannucci (2000b) wavelet methods are used to detect G+C patterns in bacterial genomes. Wavelet shrinkage is first applied to eliminate very small variations in the G+C content. Then, possible peaks in the scalogram of the wavelet representation, (Flandrin (1988)), are used to locate genomic regions with large variation in G+C content. See also Hirakawa *et al.* (1999) for similar analyses.

3. Methods

Wavelets seem to be well suitable for the analysis of biological signals that present a multi-scale nature. In Liò and Vannucci (2000a) the analysis of a 48 proteins blind test set, i.e. containing proteins not used to generate the propensity scales, resulted in accuracy per segment comparable with other currently used prediction methods. In this paper we investigate non-decimated wavelet methods as complementary tools with a greater flexibility. Components of sequences at different scales of interest can be easily extracted, whose features are aligned with those of the original data sequence. This unables us to detect transmembrane segments of proteins without having to apply first a smoothing procedure that would usually involve the choice of threshold parameters. Moreover, scale-by-scale wavelet decompositions of variance and correlation help us in highlighting hidden structures of single sequences and similarities among different sequences.

Let us first briefly summarise the main qualitative features of the wavelet transforms.

3.1. *A qualitative description of the DWT and non-decimated DWT.* Let there be n data points collected at $t\Delta t$, with Δt as the time interval between consecutive observations (1 aminoacid or 1 bp). In matrix form we can represent the discrete wavelet transform (DWT), (Mallat (1989)),

through an orthogonal matrix

$$W = [W_1, W_2, ..., W_J, V_J] \tag{1}$$

where J is the coarsest level of the transform. A DWT is applied to the vector X of observations as $d = WX$ and decomposes the data into sets of wavelet coefficients

$$d = [d_1^T, d_2^T, ..., d_J^T, c_J^T]^T \tag{2}$$

with $d_j = W_j^T X$, $c_J = V_J^T X$. At scale $\tau_j = 2^{j-1}$, or level j, the $n/2^j$ coefficients d_j are associated with changes in averages of the data on a scale $\tau_j \Delta t$ at a set of location times. This means that each wavelet coefficient at that level tells us how much a weighted average of the data changes from a particular time period of effective length $\tau_j \Delta t$ to the next one. Scaling coefficients c_J of the wavelet transform are instead associated with averages of the data on scales $\tau_{J+1}\Delta t$ and higher. Each scaling coefficient is therefore a weighted average with bandwidth $\tau_{J+1}\Delta t$ over a particular time period of effective length $\tau_{J+1}\Delta t$.

In frequency terms, scale $\tau_j \Delta t = 2^{j-1}\Delta t$ is associated to a bandpass filter with pass-band $(\frac{1}{2^{j+1}\Delta t}, \frac{1}{2^j \Delta t})$. For example, level $j = 1$, that is scale $\tau_1 = 1$, is associated with a highpass filter with pass-band the frequency interval $(\frac{1}{4\Delta t}, \frac{1}{2\Delta t})$ that is to oscillations with period length from $2\Delta t$ to $4\Delta t$. Scaling coefficients are instead associated to a lowpass filter with pass-band $(0, \frac{1}{2^{J+1}\Delta t})$. The wavelet transform is therefore a cumulative measure of the variations in the data over regions proportional to the wavelet scales, with coefficients at coarser and coarser levels, i.e. for increasing values of j, describing features at lower frequency ranges and larger time periods.

Let us now consider the non-decimated version of the DWT, a modified transform where coefficients at each level are not subsampled. Shensa (1992) first discussed this transform in the literature under the name of "undecimated" DWT. Several other names have been used since then. Here we adopt the notation of the *maximal overlap* DWT (MODWT) of Percival and Guttorp (1994). The qualitative description of the transform given above still holds with the difference that now we have n coefficients at every scale, i.e. the number of coefficients does not decrease with the level. A disadvantage is then that the transformation is not orthogonal anymore. On the other hand, other useful features are gained. Coefficients are in fact translation-invariant, meaning that circularly shifting of the data is reflected in the same shifting of the coefficients. Moreover, the non-decimated DWT is capable of handling data with arbitrary size, i.e. it does not require the sample size n to be a power of two. Finally, the MODWT is associated with

zero-phase filters, meaning that it operates by circularly filtering the data allowing features at different scales to be aligned with those of the original data sequence. This last property is particularly important for the purposes of this paper.

3.2. *Multiresolution analysis.* A wavelet transform leads to an additive decomposition of a signal into a series of different components describing smooth and rough features of the signal. In fact we have

$$X = W^T d = \sum_{j=1}^{J} W_j d_j + V_J c_J = \sum_{j=1}^{J} D_j + C_J \tag{3}$$

with D_j the detail of the signal describing changes at the scale τ_j and C_J the smooth component associated with variations at scales τ_{J+1} and higher. These components are particularly meaningful when using the MODWT, in that the zero-phase filter properties allow smooth and details to be perfectly aligned with events in the original series.

3.3. *Wavelet variance and correlation.* The wavelet variance, Percival (1995) and Percival and Mofjeld (1997), is a scale-by-scale decomposition of the variance of a signal. An estimate of the wavelet variance at a given scale is obtained by summing the squares of the wavelet coefficients (usually only those not affected by boundary conditions) and dividing by the number of them. When a bivariate signal is available, summing, at a given level, cross-products of coefficients with the same location will instead lead to an estimate of the wavelet covariance at that level. In the case of Gaussian processes generating the data, approximate confidence intervals for the MODWT estimate of the wavelet covariance can be defined, (Lindsay, Percival and Rothrock (1996)). See Serroukh and Walden (2000) for the non-Gaussian case. The wavelet cross-covariance at a given level and lag τ can be estimated by summing cross-products of coefficients at locations whose distance from each other is equal to the given lag. Estimates of the wavelet correlation and cross-correlation are obtained by dividing the wavelet covariance and cross-covariance by the product of the wavelet standard deviations. An approximate $100(1-2p)\%$ confidence interval for the MODWT wavelet correlation at level j can be constructed, (see Whitcher, Guttorp and Percival (2000)).

4. **Analyses**

Here we employ non-decimated wavelet methods in the analysis of bio-

logical signals. We first detect transmembrane regions of proteins and investigate similarities among proteins at different sequence lengths. We then locate regions of genomes with different G+C content. In all analyses we use Daubechies (1992) wavelets. A limited exploratory analysis we performed, with both minimum phase and least asymmetric wavelets, resulted in the choice of minimum phase with 2 vanishing moments.

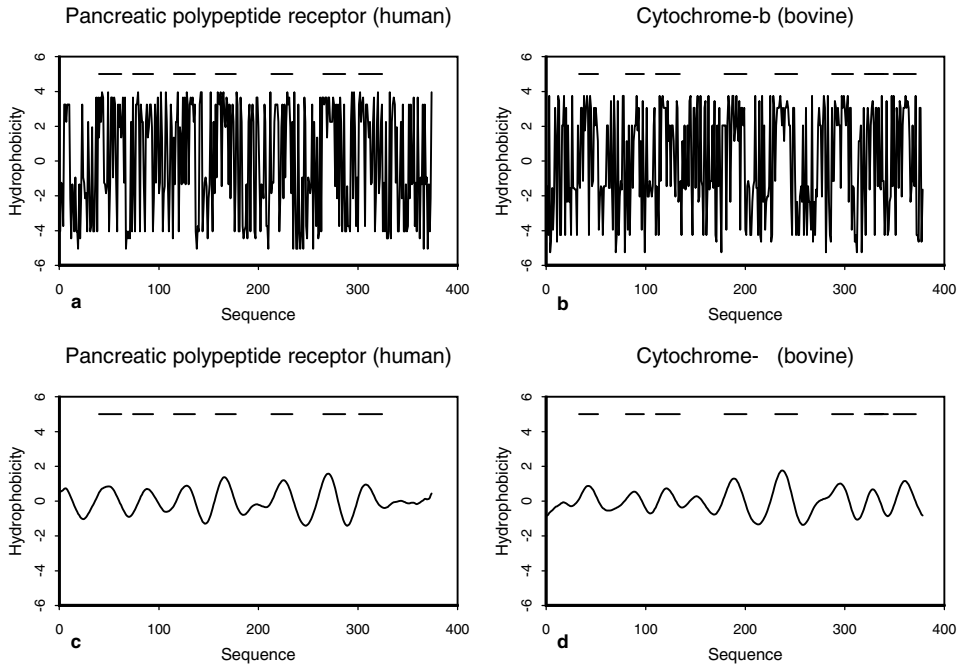

Figure 1. *(a)* profile of the human pancreatic polypeptide receptor; *(b)* profile of the bovine cytochrome b; *(c)* reconstructed smoothed profile of the human pancreatic polypeptide at level 5; *(d)* reconstructed smoothed profile of the bovine cytochrome b at level 5. The intensity of the signals is indicated on y-axis, the x-axis shows the position along the sequence. The top row of the figures indicates the positions of actual transmembrane helices (bars).

4.1. *Membrane proteins.* We coded each amino acid of the proteins according to the hydrophobic scale generated by Kyte and Doolittle (1982) to obtain signals that may show changes in correspondence of the (hydrophobic) transmembrane segments. Figure 1 shows the hydrophobicity profiles generated from the amino acid sequences of two α-helix transmembrane (HTM) proteins, *(a)* the human pancreatic polypeptide receptor, Genbank accession

AAC50280, and *(b)* the bovine cytochrome b, Genbank accession NP_008107. The intensity of the signal is indicated on the y-axis, while the x-axis shows the position along the sequence. Reported at the top of each figure are the positions of actual transmembrane helices (bars) taken from the SwissProt database. Plots *(a)* and *(b)* of Figure 2 show the wavelet variance decompositions of the two hydrophobic profiles. Both proteins have the largest wavelet variance at level 5, corresponding to changes in averages of the data on a scale of 16 aminoacids. Indeed, the amino acid sequences of a membrane proteins are mainly composed of stretches of about 20 predominantly hydrophobic residues. We studied variance plots for a large set of α-helix transmembrane proteins. Findings confirmed that these proteins always have very large wavelet variance at levels 5 and/or 6.

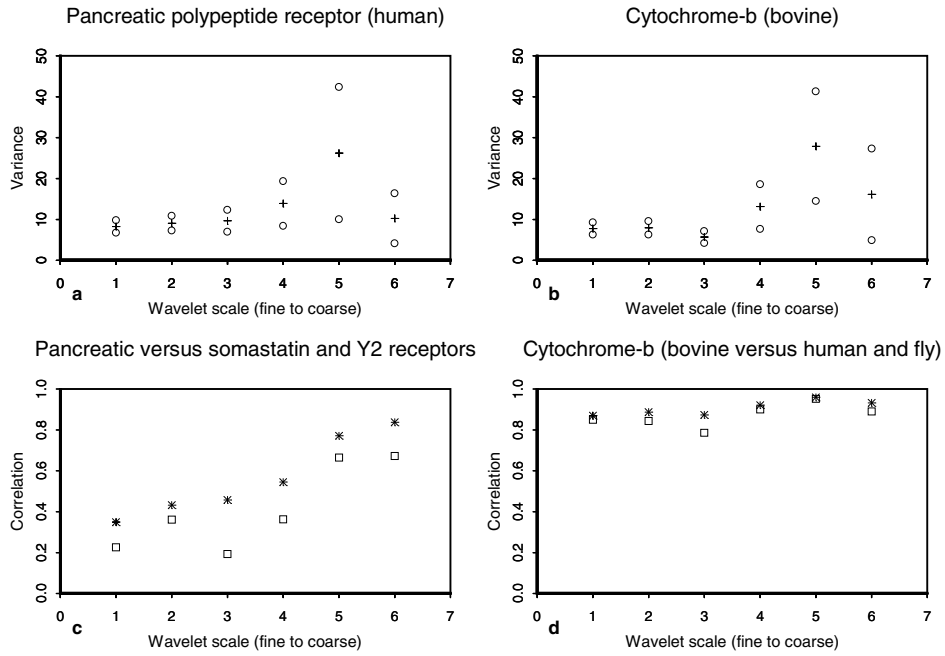

Figure 2. *(a)* Wavelet variance decomposition of the hydrophobic profiles of the human pancreatic polypeptide receptor with a 95% confidence interval; *(b)* wavelet variance decomposition of the hydrophobic profiles of the cytochrome-b sequence; *(c)* wavelet correlation decomposition of the human pancreatic polypeptide versus the somastatin receptor (*) and the neuropeptide Y2 receptor (square); *(d)* wavelet correlation decomposition of the bovine cytochrome-b versus rat (*) and insect (square) cytochrome-b amino acid sequences.

Plots *(c)* and *(d)* of Figure 1 show the reconstructed smoothed profiles at level 5 of the human pancreatic polypeptide receptor and the bovine cytochrome b, respectively. In Tables 1 and 2 we compare predictions on the smoothed profiles of the α-helix transmembrane segments of the proteins with the observed locations and also with predictions from other methods currently used: TMpred, Hofmann and Stoffel (1993), TMHMM, Sonnhammer *et al.* (1998), TopPred2, von Heijne (1992), MEMSAT, Jones (1994). Although current methods can identify around $90 - 95\%$, with an overprediction rate of only a few percent, in many membrane proteins there are candidate segments that have intermediate average hydrophobicity and cannot be confidently predicted as transmembrane segments. Moreover, $\alpha-$ helices not inserted in the membrane can be still enriched of hydrophobic amino acids because they are in contact with the membrane surface. These helices result in false positives, like the smaller peak in figure 1, plot *(d)*, predicted as a true helix by all methods in Table 2.

Table 1. COMPARISON OF (A) OBSERVED LOCATIONS OF THE 7 α-HELIX TRANSMEMBRANE SEGMENTS OF THE HUMAN PANCREATIC POLYPEPTIDE, CODED ACCORDING TO THE KYTE-DOOLITTLE SCALE, WITH (B) PREDICTIONS USING THE RECONSTRUCTION AT THE SCALE WITH LARGEST WAVELET VARIANCE, AND WITH OTHER PREDICTION METHODS CURRENTLY USED: (C) TMPRED, HOFMANN AND STOFFEL (1993), (D) TMHMM, SONNHAMMER *et al.* (1998), (E) TOPPRED2, VON HEIJNE (1992), (F) MEMSAT, JONES (1994).

	TM1	TM2	TM3	TM4	TM5	TM6	TM7
A	41-63	75-95	116–137	158–178	214-235	266-288	302-325
B	37-63	82-99	119-136	156-177	214-235	261-282	300-321
C	50-67	78-104	116-137	156-176	217-235	266-284	302-325
D	40-62	82-104	119-137	158-180	213-235	265-283	302-324
E	48-68	84-104	117-137	156-176	215-235	264-284	300-320
F	42-66	72-102	117-136	156-179	212-233	262–278	301-324

We also analysed bivariate sequences using the wavelet scale-by-scale correlation decomposition, hoping to find similarities among proteins at different length-scales. Wavelet correlation decomposition provides an important tool to investigate whether distantly related proteins maintain the same structural organization. Also, correlation values of homologous proteins belonging to different species can reveal differences among species. Figure 2, plot *(c)*, shows the wavelet correlation decomposition of the hydrophobic profile of the pancreatic polypeptide with respect to the somastatin receptor (*; Genbank accession NP_001044) and the neuropeptide Y2 receptor (square;

Table 2. Comparison of (A) observed locations of the 8 α-helix transmembrane segments of the bovine cytocrome-b, coded according to the Kyte-Doolittle scale, with (B) predictions using the reconstruction at the scale with largest wavelet variance, and with other prediction methods currently used: (C) TMpred, Hofmann and Stoffel (1993), (D) TMHMM, Sonnhammer *et al.* (1998), (E) TopPred2, von Heijne (1992), (F) MEMSAT, Jones (1994).

	TM1	TM2	TM3	-	TM4	TM5	TM6	TM7	TM8
A	33-52	80-98	110-134	-	179-201	230-252	287-308	325-343	349-371
B	30-55	85-95	114-133	149-153	177-199	226-249	281-306	325-335	350-370
C	32-51	81-99	115-132	140-161	178-200	229-247	288-307	322-340	348-372
D	33-55	76-98	113-135	148-166	178-200	229-251	288-310	323-341	350-372
E	33-53	81-101	114-134	138-158	178-198	228-248	288-308	320-340	352-372
F	29-52	-	114-132	139-157	177-199	228-245	287-306	322-338	347-371

Genbank accession I39163). These two latter receptors share about 30% sequence similarities with the pancreatic polypeptide receptor sequences. Correlations are higher at levels 5 and 6. These large values suggest that the somastatin and neuropeptide Y2 receptors may be α-helix transmembrane proteins with locations of their α-helix transmembrane segments very close to the corresponding segments in the pancreatic polypeptide protein. Figure 2, plot *(d)*, shows the wavelet correlation decomposition of the hydrophobic profile of the bovine versus rat (*; *rat norvegicus*, Genbank accession BAA85626) and insect (square; *drosophila simulans* Genbank accession AAF77575) cytochrome-b sequences. The correlation plots suggest similarities at all sequence-length scales. The correlation values are lower for the more distantly related species (bovine and insect). The large correlation values at all scales indicate a high level of sequence and structural similarities among the bovine, rat and insect cytochrome b proteins, i.e. this protein has not changed very much during evolution.

4.2. *Bacterial genomes.* We now turn to the analysis of G+C patterns of the genome of the bacterium *meningitidis* serogroup A (strain Z2491), (Parkhill *et al.* (2000)), the causative agent of meningitis, responsible for considerable morbidity and mortality throughout the world, and of three mycoplasma genomes: *U. urealyticum, M. genitalium, M. pneumoniae. U. urealyticum* is a human pathogen of the urogenital tract causing adverse pregnancy outcome, neonatal disease and supperative arthritis. *M. genitalium* is

a parasite of the primate genital and respiratory tracts. *M. pneumoniae* is a human pathogen causing tracheobronchitis and primary atypical pneumonia.

Figures 3 and 4 show the wavelet variance decompositions of the G+C profiles of the four genome sequences and corresponding reconstructions at sets of levels with high variance, levels [11-13] for the *M. genitalium*, levels [11-14] for the *U. urealyticum*, levels [11-14] for the *M. pneumoniae* and levels [11-15] for the *N. meningitidis* serogroup A. Large variance values at high scales are revelatory of large genomic regions with different G+C content with respect to the nearby regions. In particular, the genome of *N. meningitidis* has a large variance at high scales in that it contains several large pathogenic regions, (Parkhill *et al.* (2000) and Liò and Vannucci (2000b)). These regions, together with other genes transferred from other bacterial genomes give a patchiness-like landscape to *N. meningitidis* G+C plots.

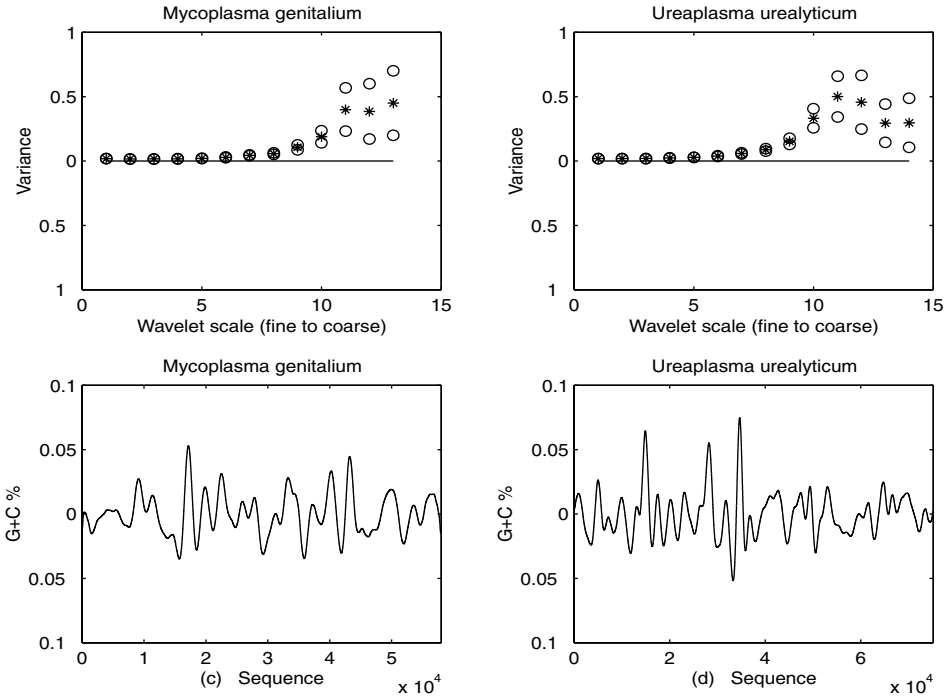

Figure 3. *(a)* wavelet variance decomposition of the G+C profile of the *M. genitalium* genome sequence; *(b)* wavelet variance decomposition of the G+C profile of the *U. urealyticum* genome; *(c)* reconstructed smoothed profile of the *M. genitalium* at levels [11-13]; *(d)* reconstructed smoothed profile of the *U. urealyticum* at levels [11-14].

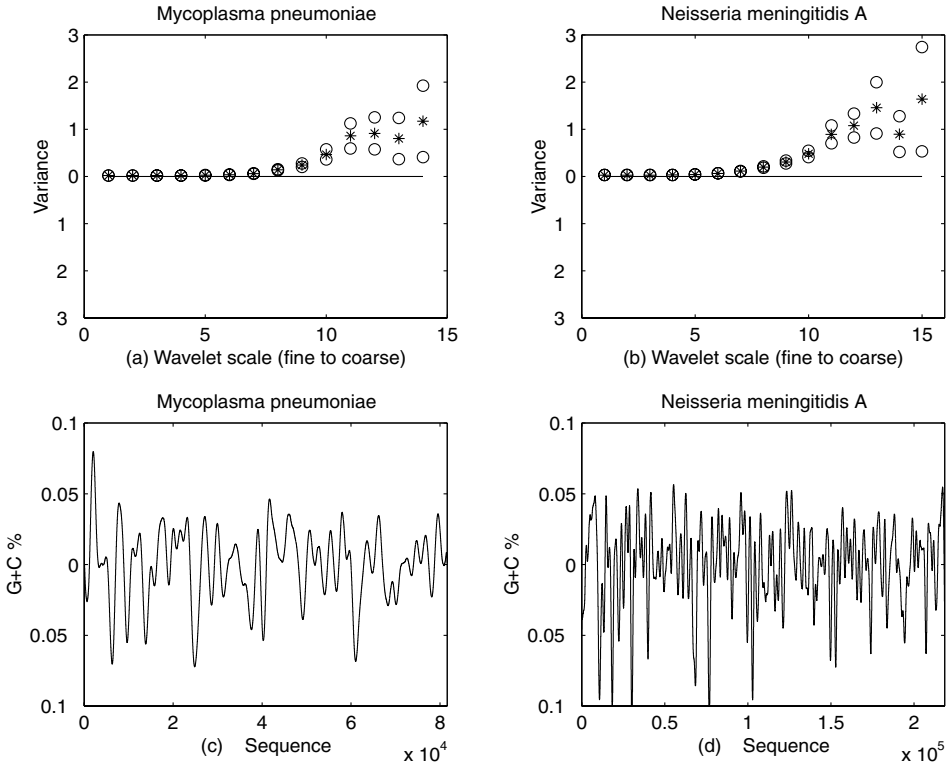

Figure 4. *(a)* wavelet variance decomposition of the G+C profile of the *M. pneumoniae* genome sequence; *(b)* wavelet variance decomposition of the G+C profile of the *N. meningitidis* serogroup A genome; *(c)* reconstructed smoothed profile of the *M. pneumoniae* at levels [11-14]; *(d)* reconstructed smoothed profile of the *N. meningitidis* serogroup A at levels [11-15].

The *U. urealyticum* variance decomposition shows values of the same order of magnitude of the other mycoplasmas. This may suggest that there are no large regions transferred from other kind of bacteria with different G+C content. Mycoplasmas comprise a large group of bacteria that represent a minimal life form, with very small genomes. Moreover, mycoplasmas lack of cell wall, unlike many pathogenic bacteria, e.g. *Neisseria meningitis*, that rely on exchanging genes involved in biosynthesis of cell wall, among other bacteria, in order to escape the immunitary defense of the host organism.

It is worth noticing that *M. pneumoniae* has larger variance than the other two mycoplasmas. This bacterium has a slightly higher average G+C content (41%) with respect to *U. urealyticum* (∼26%) and *Mycoplasma genitalium* (∼30%). Thus, it has a less biased codon usage and a larger popu-

lation of tRNAs, (Nakamura *et al.* (1997)), meaning that it can more easily exchange genetic material with the other bacteria than the other mycoplasmas.

5. Concluding Remarks

We have used here non-decimated wavelet transforms to analyse biological signals, specifically proteins and genomes. We have looked at the wavelet variance decomposition of membrane proteins and used the wavelet reconstructions of the sequences at selected scales to locate the positions of transmembrane helices. Scale-by-scale correlation decompositions have helped us to investigate similarities among different amino acid sequences. We have also detected DNA segments that have different composition characteristics and are putative pathogenicity regions. Our work has emphasised how newly defined wavelet methods can help in interpreting complicated structures, like the proteins and genomes analysed here. More work needs to be done in investigating the structures that variance and correlation decompositions seem to suggest. The application of wavelet methods in genetics and molecular biology is just at its beginning. Many contributions are to be expected in the analysis of the vast amount of data that biologists are accumulating on mutations and structural variations of genomes and proteomes from different species.

Acknowledgments. Marina Vannucci is supported by National Science Foundation, CAREER award number DMS-0093208 and partially by MURST, Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Italy. Pietro Liò is supported by an EPSRC/BBSRC Bioinformatics Initiative grant.

References

ALDROUBI, A. and UNSER, M. (eds) (1996). *Wavelets in Medicine and Biology.* CRC Press, Boca Raton, Florida.

ARNEODO, A., D'AUBENTON, CARAFA Y., BACRY, E., GRAVES, P.V., MUZY, J.F. and THERMES, C. (1996). Wavelet based fractal analysis of DNA sequences, *Physica D*, **1328**, 1–30.

BRANDEN, C. AND TOOZE, J. (eds) (1991). *Introduction to Protein Structure.* Garland Publishing, New York.

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

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

FRASER, C.M. ET AL. (1995). The minimal gene complement of Mycoplasma genitalium, *Science*, **270**, 397–403.

GLASS, J.I., LEFKOWITZ, E.J., GLASS, J.S., HEINER, C.R., CHEN, E.Y. and CASSELL, G.H. (2000). The complete sequence of the mucosal pathogen ureaplasma urealyticum, *Nature*, **407**, 757–762.

HACKER, J., BLUM-OEHLER, G., MUHLDORFER, I. and TSCHAPE, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution, *Molecular Microbiology*, **23**, 1089–1097.

HATTORI, M. ET AL. (2000). The DNA sequence of human chromosome 21, *Nature*, **405**, 311–319.

VON HEIJNE, G. (1992). Membrane Protein Structure prediction – hydrophobicity analysis and the positive-inside rule, *Journal of Molecular Biology*, **225**, 487–494.

HIMMELREICH, R. ET AL. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae, *Nucleic Acids Research*, **24**, 4420–4449.

HIRAKAWA, H., MUTA, S. and KUHARA, S. (1999). The hydrophobic core of proteins predicted by wavelet analysis, *Bioinformatics*, **2**, 141–148.

HOFMANN, K. and STOFFEL, W. (1993). TMbase a database of membrane spanning proteins segments, *Biological Chemistry Hoppe-Seyler*, **347**, 166–171.

JONES, D.T. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology, *Biochemistry*, **33**, 3038–3049.

KARLIN, S. and BRENDEL, V. (1993). Patchiness and correlations in DNA sequences, *Science*, **259**, 677–680.

KARLIN, S., CAMPBELL, A.M. and MRAZEK, J. (1998). Comparative DNA analysis across diverse genomes, *Annual Review of Genetics*, **32**, 185–225.

KYTE, J. and DOOLITTLE, R.F. (1982). A simple method for displaying the hydrophatic character of a protein, *Journal of Molecular Biology*, **157**, 105–132.

LINDSAY, R.W., PERCIVAL, D.B. and ROTHROCK, D.A. (1996). The discrete wavelet transform and the scale analysis of the surface properties of sea ice, *IEEE Transactions on Geoscience and Remote Sensing*, **34**, 771–787.

LIÒ, P., RUFFO, S., POLITI, A. and BUIATTI, M. (1996). Analysis of genomic patchiness of *Haemophilus influenzae* and *S. cerevisiae* chromosomes, *Journal of Theoretical Biology,* **183**, 455–469.

LIÒ, P. and VANNUCCI, M. (2000a). Wavelet change-point prediction of transmembrane proteins, *Bioinformatics*, **16**, 376–382.

LIÒ, P. and VANNUCCI, M. (2000b). Finding pathogenicity islands and gene transfer events in genome data, *Bioinformatics*, **16**, 932–940.

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

MUTO, A. and OSAWA, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution, *Proceedings National Academy Science U.S.A.* **84**, 166–169.

NAKAMURA, Y., GOJOBORI, T., and IKEMURA T. (1997). Codon usage tabulated from the international DNA sequence databases, *Nucleic Acids Research*, **25**, 244–245.

PARKHILL, J. ET AL. (2000). Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491, *Nature*, **404**, 502–506.

PERCIVAL, D.B. (1995). On estimation of the wavelet variance, *Biometrika*, **82**, 619–631.

PERCIVAL, D.B. and MOFJELD, P. (1997). Multiresolution and variance analysis using wavelets with applications to subtidal coastal seas, *Journal of the American Statistical Association*, **92**, 868–880.

PERCIVAL, D.B. and GUTTORP, P. (1994). Long-memory processes, the Allan variance and wavelets. In *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar (eds.) Academic Press, San Diego, 325–344.

ROST, B., FARISELLI, P. and CASADIO, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy, *Protein Science*, **5**, 1704–1718.

SHENSA, G. (1992). The discrete wavelet transform: Wedding the à trous and Mallat algorithms, *IEEE Transactions on Signal Processing*, **40**, 2464–2482.

SONNHAMMER, E.L.L., VON HEIJNE, G. and KROGH, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences, *Intelligent Systems for Molecular Biology*, **6**, 175–182.

SERROUKH, A. and WALDEN, A.T. (2000). Wavelet scale analysis of bivariate time series I: motivation and estimation, *Journal of Nonparametric Statistics*, **13**, 1–36.

TENZEN, T., YAMAGATA, T., FUKAGAWA, T., SUGAYA, K., ANDO, A., INOHO, H., GOJOBORI, T., FUJIYAMA, A. OKUMURA, K. and IKEMURA, T. (1997). Precise switching of DNA replication time in the GC content transition area in the human MHC, *Molecular Cell Biology*, **17**, 4043–4050.

WHITCHER, B., GUTTORP, P. and PERCIVAL, D.B. (2000). Wavelet analysis of covariance with application to atmospheric time series, *Journal of Geophysical Research – Atmospheres*, **105**, 14,941–14,962.

MARINA VANNUCCI
DEPARTMENT OF STATISTICS
436 BLOCKER BUILDING
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-3143
USA
E-mail: mvannucci@stat.tamu.edu

PIETRO LIÒ
DEPARTMENT OF ZOOLOGY
UNIVERSITY OF CAMBRIDGE
DOWNING STREET
CAMBRIDGE CB2 3EJ
UNITED KINGDOM
E-mail: P.Lio@zoo.cam.ac.uk