33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Insight into DNA periodicity by a single-channel sequence data approach

Mariusz Zoltowski

Abstract—**It has not been obvious how to map a genomic sequence into the numbers to elucidate its periodicities by digital signal processing (DSP) in accord with the underlying biology [1]. The well known DNA spectra and their extensions appear the A-T-C-G – base-wise by Fourier (FT), wavelet (WT) or related transforms of the indicatory functions (IF-s) of these bases. The IF assumes either *1* -in the presence or *0* - in the absence of the indicated base in sequence. The IF's spectra can be next combined in a different way including the optimal one to provide the net spectrum [2].**

In this contribution, it is attempted, and not limited to; showing that single channel numeric DNA also turns out to be sufficient for biologically meaningful results by DSP with accompanying merits. Plausibility is possible considering any RNA message as a single-channel coded waveform; by the triplets of the codon bases which code for 20 different amino acids. This in turn enables a clear justification for the coding rhythm in terms of the codon usage frequency (CUF) and the gene series autocorrelation. The latter simply assesses a self-similarity of the message. Along with appending well established communication insight to biological perspectives, the answer to how the genetic code is becoming specific, inducing the self-similarity of the coded sequences under the three-base-shift case is addressed.

Supporting the focus, there are some findings in vertebrates' genes data elucidated by the EMD of Huang-Hilbert transform (H-HT) [3]; these are long-term spectra relevant to the coding, the content of dicodons and the structural properties of coded proteins [4].

Also a new finding in the coding rhythm - the one which is attributed to the coding DNA, is included. This is the net coding rhythms in Homo sapiens, Homo sapiens house-keeping and vertebrates' genes comparison by histograms of adaptively tracked amplitudes case.

It is intriguing how spectral features of genomic sequences correspond to related physical phenomena [5-8].

I. Introduction

GENOMIC signal processing (GSP) by DSP of genomic data emerged to handle important problems of genomics and proteomics in a fast and robust way [9]. The fields, capabilities and limits of GSP are subjects of continuing research.

Single-data-channel periodicities have mainly been involved in describing proteins. This has been so for some degree of symmetry or periodicity is often related to their function. Several researchers used Fourier transform to search for periodicities in residue properties or with hydrophobicity scales to reveal amphiphilic structures of proteins and also looked for improved tools to quantify periodicities in the symbolic sequences; extensive references are included in [1] and also in [5]. Another approach has either been related to the structural factors or function [1, 5], or applied across sets of proteins to elucidate common periodicities in amino acids sequences. The ligand-receptor resonance idea at sequence abstraction level was coined [5].

Assigning a statistical significance to the periodicities was also addressed by signal to noise ratio (S/N) for each peak in cross-spectral function, and more rigorously by Monte Carlo envelopes and the significance-level-tables [1].

Fourier's approach with spectral features either has limits or brings about some controversy [1]. Hence it is well suited for considerations while possibly awaiting some rectifications, completing or integrating.

Structured into material, methodology, results and discussion, this paper is to consider spectral properties of the DNA-RNA data mapped into single numeric series with an impact on the coding rhythm. The classical spectral approach and autocorrelation [10] with the EMD of Huang-Hilbert transform (HH-T) [3]; histograms of adaptively tracked periodicities and figures on the source of CUF are used in pursuit of better understanding of the "work" of genetic code in relation to DSP. With primitives related to the genomic sequence abstraction level, to what extent can any theory of intermolecular interactions in biological cell give plausible results?

II. Material and Methods

A. The EMD of Huang-Hilbert transform (H-HT [3]) and spectra across a set of genes

Assigning [5]; A-0.1260, T-0.1335, U-0.0289; C-0.1340, G-0.0806 [Ry] a gene which is composed of A, T/U, C, G - bases becomes a real valued coded sequence *x(n)*:

$x(n)=\sum_k a_k \delta(n-k)$ (1a), where $\delta()$ is Kronecker delta and a_k a real, quaternary valued image of mapped base.

In turn, the autocorrelation of *x(n)* is given by:

$$R_{xx}(\tau)=\sum_n R_n \delta(\tau-n);\; R_n=E\{a_{k+n}a_k\} \quad (1b).$$

Thus *x(n)* admits its long-term spectrum (Wiener-Khintchin[10]):

Manuscript received April 8, 2011. M. Zoltowski is with CM UMK, Bydgoszcz, Poland; phone 48 691 833 860; e-mail: mzoltowski@cm.umk.pl.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

$$S_{xx}(\omega) = \sum_{\tau} R_{xx}(\tau)\exp(-j\omega\tau) = R_0 + \sum_{n} R_n \cos n\omega \quad \text{(1c).}$$

Finding out if the (1a) coding implies biologically plausible results could be forwarded by looking for some special features of the coded sequences. The H-HT has proved to result in physically acceptable findings so it was expected to be good choice.

The main part of H-HT is the empirical mode decomposition (EMD, see [3]) which is adaptively performed into so called intrinsic mode functions (IMFs).

As simple oscillatory modes-the IMFs of the EMD of *x(m)* (*m=1... L)* say c_j *(m)*, mostly resemble modulated in amplitude–and-frequency waveforms. Briefly *x* is the finite sum of *n* – IMF- c_j and a residual trend like term r_n i.e.:

$x(m) = \sum_{j=1}^{n} c_j(m) + r_n(m)$ (2a). IMFs $cj = h_{jk}$; j=1…n.

The latter holds for some *k>0* which comes from a stoppage criterion completing so called iterations of sifting [3]: $h_{jk} = h_{j(k-1)} - m_{jk}$, $h_{j0} = r_{j-1} - m_{j0}$ (2b); where r_j, j=1... n are given in (2a) and $r_0=x$ accordingly.

The *m*-subtrahend in (2b) is by the mean value of the upper and the lower envelopes of the minuend *h* or *r*. Those envelopes result by cubic splines which connect the local maxima and local minima in covering all data way [3].

Two stoppage criteria: S and SD have been postulated [3]; the S number of the 1st is the one of the consecutive siftings after which the both numbers are the same; these of zero-crossing and of extrema ones. The SD of the 2nd criterion should be smaller than a pre-set value, where

$$SD_{jk} = \sum_{m=1}^{L} \left| h_{j(k-1)}(m) - h_{jk}(m) \right|^2 / \sum_{m=1}^{L} h^2_{j(k-1)}(m)$$

and L is the length of decomposed sequence [3].

In turn, let's look for some net spectral properties across a whole set of genes. Then consistently with the (1a-c) it is to handle the expectation of (1b) as in the below, using:

1. Single-gene-g of length L_g-wise correlation:

$$R^g_n = \frac{1}{L_g - n} \sum_{k=1}^{L_g - n} a_k a_{k+n}; \quad \text{n=0, 1.... (3a);}$$

2. R^g_n of (3a) across-the-set-weighting by the length of g - L_g to yield $R_n \cong \frac{1}{L} \sum_g L_g R^g_n$; $L = \sum_g L_g$ (3b).

The estimation of (3a-b) applied to every IMF-c_j of (2a) results in spectra; these of (1c), windowed possibly to alleviate Gibbs's effect of the truncation [10]. The (3a) can be handled by FFT.

B. Histograms of adaptively tracked periodicities [11]

Let the frequency-range of a genomic signal of (1a) (*position- n* in sequence substitutes time-*n*) be spanned by a bank of filters. Band-pass amplitude - considered the response of the each filter, can be adaptively tracked. The histogram of such amplitudes across the coding exons serves as a feature of net periodicities e.g. in coding DNA. This is distinct from the position-in-sequence dependent amplitudes used to infer the coding regions [11].

To uncover the coding rhythm at *1/3* [Hz] from the sampling of the genomic series at *1* [Hz], a digital band-pass filtering (around *1/3* [Hz]) is performed by the filter which kernel is [11]

$$\mathbf{h}^*_{BP} = \{h^*_{BP}(n), -N_{FIR} \le n \le N_{FIR}\} = Z^{-1}\{H_{BP}(z)H_{BP}(z^{-1})\}$$

(5a). This is Z-inverse of the causal - $H_{BP}(z)$ and non-causal - $H_{BP}(z^{-1})$ filtering cascade case that is given in (5a) with $N_{FIR}+1$ taps of the each FIR filter. It is well known that such a symmetric-by-kernel filter is the delay-less i.e. the filter input data are not shifted versus the output ones. Finally, synchronously demodulated amplitude of a band-pass signal is obtained by the phase-locking technique in two steps [11]; *1.* Transforming a band-pass filtered genomic signal *x* of length *L* into the analytical at ~1/3 Hz with Hilbert transformed x_h:

$$x_a = x + jx_h = \{x_a(n) = |x_a(n)| \exp[j\varphi_a(n)] \times$$

$$\times \exp(j2\pi n/3) \quad n = 1,\ldots,L\} \quad (5b)$$

2. Constituting a synchronous reference: $\{r(n) = \exp[j2\pi n/3 + \varphi(n-1)]\, n = 1,\ldots,L\}$ (5*c*) at ~1/3 Hz by adaptive adjusting its phase φ to match φ_a of (5b).

The latter is done by a diminishing of the phase error:

$$\varphi^{err}(n) = \varphi_a(n) - \varphi(n-1) = angle[x_a(n)r(n)^*] \qquad \text{(5c),}$$

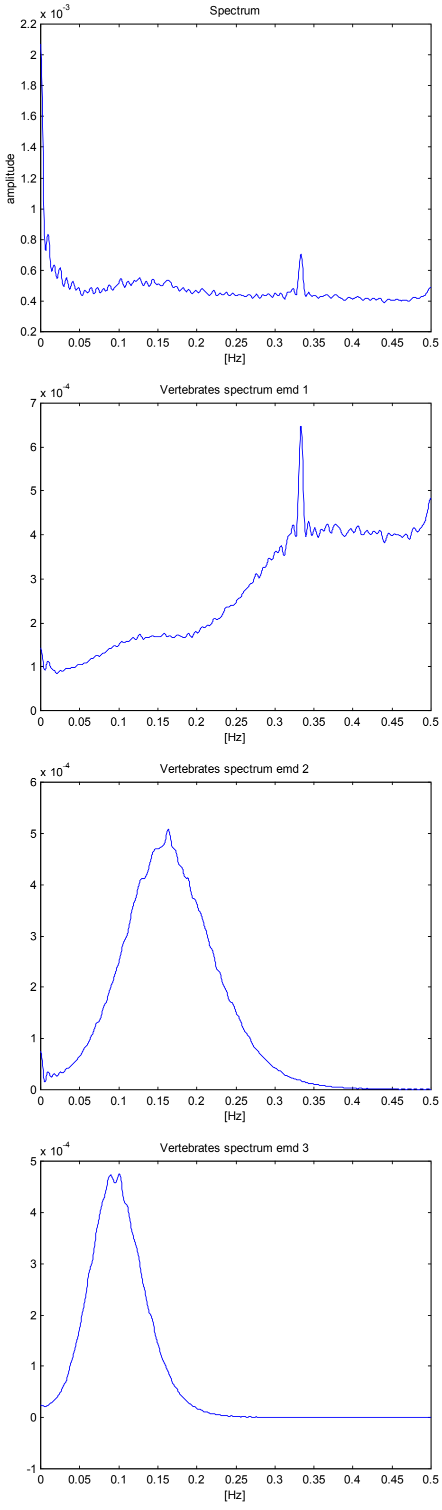
within the iterations: $\varphi(n) = \varphi(n-1) + K_1\varphi(n)^{err}$ (5d) n=*1*... (*-conjugation). The gain - K_1 of low-pass filter (5d) is responsible either for smoothing or tracking of phase-φ; the smaller K_1 the better smoothing but the worse tracking and vice versa. This is the product of (5c) in rectangular brackets that provides the amplitude entry to the histogram.

C. Genomic data

The data used include; the B-G set standing for the well known Bursét and Guigo collection of 570 vertebrates' genes and; the set of human house-keeping H-K genes derived from the article [12]. The net human codon usage was taken from statistics available on Web.

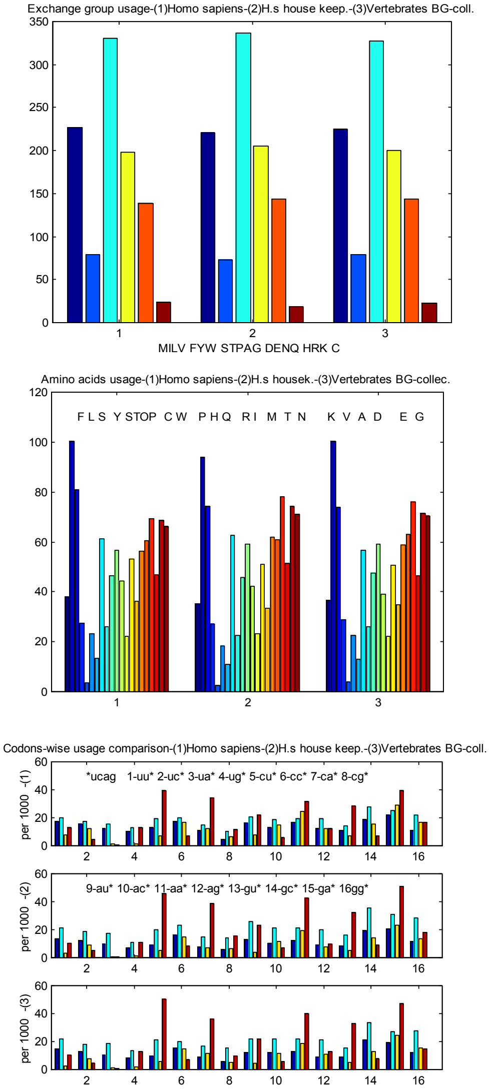
III. Results

The derivations of section IIA are followed by the illustrating Figs 1-4. They include in turn B-G set of vertebrates' genes by their net spectrum and the spectra of their first EMD-components of the H-HT (-sec. IIA). The next Figs are on the coding rhythm. They show a strategy engaged in coding for proteins in “closely” related cases.



Figs.1 - 4 (top - bottom). F.1 - net B-G set spectrum; F.2- Coding spectrum by the 1st-EMD component of H-HT; F.3-Dicodons’ spectrum by the 2nd-EMD comp.; F.4- Protein structure related spectrum by the 3rd –EMD component.

Six Exchange groups [4] of similar amino acids are given in the comparison of Fig.5. Unequal demands for their pools are the first step the code becomes specific. The next two steps are due to amino acids recruitment-Fig.6 and codon usage-Fig.7 and also Fig.9.



Figs 5 - 7 (top - bottom). Comparison between; Homo sapiens (h.s.) and; h.s. H-K set and; B-G set; -that is left-right in Figs 5-6 and top-bottom in Fig.7. The usage; of Exchange groups-in Fig.5; of amino acids by standard code in-Fig.6; and of codons ordered the U C A G -wise; this ordering runs at the 3rd - then at the 2nd – and at the 1st -base position in 16 clusters in the each figure-row of Fig.7.

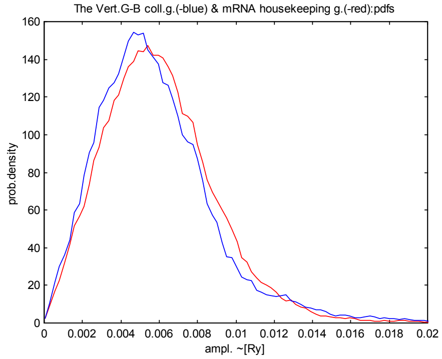


Fig. 8. Highly significant “a bit” difference (Kolmogorov-Smirnov test with p<<0.001) between the G-B set in blue and the house-keeping genes (H-K) in

red. Compared histograms of the amplitudes of the coding rhythm in the coding regions of genes from the both sets.

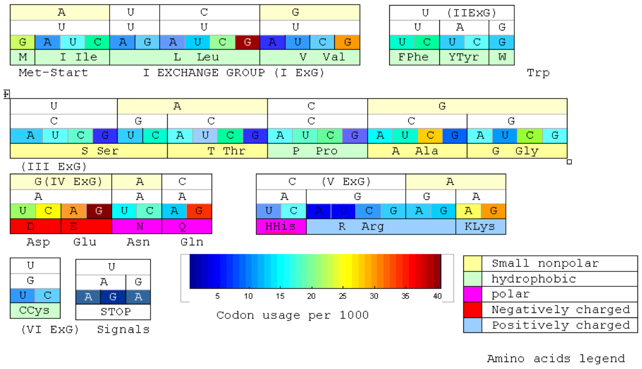


Fig.9. Homo sapiens net CUF by heat map colors assigned to the 3rd base of the each codon showing how the code is specific by deviating from the uniform usage case. Six Exchange groups and stop signals by subsequent left-right ordered rectangles. Coded amino acids and their properties by colored labels in the bottom row of every rectangle.

IV. Conclusion

What it is shown in this paper, and not limited to, is that single-channel data of genes by their nucleotide bases mapped by (1a) into the numbers are sufficient to provide biologically meaningful results and thus are also adequate for further considerations.

The H-HT EMD spectra besides the one of coding rhythm of Fig.2 (-period 3, 0.33 Hz) include other spectra that are not perceived at glimpse in Fig. 1 which is but H-HT EMD case. They are related to the rhythm of dicodons (0.17 Hz) - in Fig.3 and to the hydrophobic or hydrophilic periodicity - in Fig.4 (~10.5 period, 0.095Hz). The patterns of dicodons are important as features of coding DNA [4]. Hence H-HT has proved valid and its use for further exploration of genomic data may also be interesting.

Obviously this is the bias of codon's pattern (see Fig.9) that gives rise to the 3-base-wise-lagged correlations, introducing the coding triplet periodicity. If also other lags were pronounced the sequence would be too poor to code for all amino acids. This is also not random, since of the bias origin, the self-similarity under a codon-wise-shift case, due to which the spectrum of the coding rhythm and its autocorrelation are shaped on, given a long enough term. The latter can be easily assessed in a single series manageable case. How specific a transcript translated in cell's ribosomes is, can be considered by the codon usage frequency (CUF). The CUF in turn, has been shaped in response to many different evolutionary pressures. How the genetic code works is of great importance to applications ranging from synthetic biology for biotechnology to protein crystallography [13]. Addressing this aspect, Figs. 5-7 serve as a comparison (see the captions); similar demands on the membership to Exchange group in Fig.5 undergo a diversification with the concrete amino acid recruitment in Fig.6 and further on with a synonymous codon usage for each amino acid - in Fig.7. Since the self-similarity is increased by abundant similar-and-the-same-codons, the codon usage has a direct impact on the coding rhythm. Such a rhythm is compared by the section IIB histograms of amplitudes in Fig.8. The coding rhythm of human house-keeping genes looks "a bit stronger" than the one of vertebrates from the B-G set. Is it another major feature of the house-keeping (H-K) genes - which support the cell maintenance? Presumably it is, since H-K genes can be expected to translate efficiently and promptly [12]. An augmented abundance of similar-and-the-same-codons enforcing the coding rhythm seems to be correlated with the population size of tRNA isoacceptors. This in turn, should imply a prompt delivery of proteins by ribosomes [14].

The FT related methods to infer biological properties and function of the amino acids or gene nucleotides have either limitations or induce controversy [1]. Insight to organic chemistry by simplified models of Quantum Mechanics [6, 7] applied to quasi-periodic structures is consistent with the FT method of [1, 5] and observed absorption of light in cell's proteins can be explained. However, could the interactions of a ligand-receptor type [8] be justified in more depth at a genomic sequence abstraction level than by a "numerical homology" like approach possibly categorizing current attempts?

References

[1] C.J.R. Illingworth, K. E. Parkes, C. R. Snell, P.M. Mullineaux, C.A Reynolds, "Criteria for confirming sequence periodicity identified by Fourier transform analysis: Application to GCR2, a candidate plant GPCR?," *Biophysical Chemistry*, vol.133, Elsevier, Holland, 2008, pp 28-35.

[2] D. Anastassiou, "Genomic Signal Processing", *IEEE Signal Process. Mag.*, IEEE, NJ, July 2001, pp 8-20.

[3] N.E. Huang, S. S. Shen, Hilbert-Huang transform and its application, Interdisc. Math. Sciences, vol.5, World Scientific, London, U.K., 2005.

[4] C.H. Wu, J.W. McLarty, *Neural Networks and Genome Informatics, Methods in Comput. Biol. and Biochem.* , Elsevier, Oxford UK. 2000.

[5] E. Pirogova, M. Akay and I. Cosic, "Computational Analysis of interactions between Tumor and Tumor Suppressor proteins" *in Genomics and Proteomics Engineering in Medicine and Biology*, John Wiley & Sons, USA, 2007, pp 257-287.

[6] R.F. Feynman, R.B. Leighton, M. Sands, *The Feynman lectures on physics*, vol.3 , Addison-Wesley Publishing Company Inc., Reading Massachusetts, U.S.A, 1969.

[7] D. Stauffer, H.E. Stanley, *From Newton to Mandelbrot. A primer in Theoretical Physics with Fractals for the personal computer*, 2nd ed., Springer - Verlag Berlin, Heidelberg, Germany, 1995.

[8] J. R. Reimers, L.K. McKemmish, R. H. McKenzie, A.E. Mark, and N.S. Hush, "Weak, strong, and coherent regimes of Fröhlich condensation and their applications to terahertz medicine and quantum consciousness", *PNAS* , vol. 106, no. 11, USA, March 17, 2009, pp 4219-24.

[9] "Dissecting the Building Blocks, A Look at the Signal Processing in Genomics" – special issue: *IEEE Signal Process. Mag*., vol. 24, no 1, IEEE, USA, January 2007.

[10] J.P. Proakis, D.G. Manolakis, *Digital Signal Processing, Principles Algorithms and Applications*, 3rd ed., Prentice Hall, USA, 1996.

[11] M. Zoltowski, "Is DNA periodicity only due to CUF- codons usage frequency?", *in Proc. of the 29th Annual Int. Conf. of the IEEE EMBS*", Cité Int., Lyon, France, August 23-26, 2007, pp 1383-6.

[12] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes are compact", *Trends in Genetics*, vol.19, no 7, USA, July 2003, pp 362-5.

[13] "From genes to proteins. The impact of gene sequence on translation and expression", *Science webinar series*, *Science*, USA, 28 October 2009, Available: http://webinar.sciencecareers.org/genestoproteins/.

[14] L. Ponnala, "On finding poorly translated codons based on their usage frequency*", Bioinformation,* 4, (2), USA, 2009, pp 63-65.