Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

IS DNA CODE PERIODICITY ONLY DUE TO CUF – CODONS USAGE FREQUENCY?

Mariusz Zoltowski, zoltm@ieee.org,
Biomed. Signals Analysis Lab., L.Rydygier Collegium Medicum, Nicolaus Copernicus Univ. of Torun

Abstract - **The triplet code for proteins and functional RNA has been either from the universal pattern of ancient RNA (-H1) [1], with a key role of an uneven codon usage frequency (CUF) in the periodic patterns origination, or a reading frame monitoring device (RFMD -H2) [2- 4]. H1 has lately been upheld [1] but in a single sequence sensitive way [1].**

Since H1 and H2 are not mutually exclusive [2, 3, 4], a single sequence-wise sensitive approach by a resonant recognition model (RRM) has become the attempt described in this paper to challenge H1 and H2 in eukaryotes case as a novelty.

In the RRM model [5, 6, 7] two bio-molecules interact favorably provided they both obey a common frequency and opposite phases consensus in their delocalized electron energy (DEE-) distributions [5].

Hence it has been possible to learn how well the DEE-s of the mRNA and of the ribosome match each other at 1/3 Hz - that applied to both the original and the CUF preserving randomly shuffled genomic data across the well known Bursét and Guigo collection of 570 coding vertebrates' genes.

The matching of RRM patterns reduces to harmonics phase comparison of the relevant DEE-s, a task by a digital phase locked loop (DPLL) [8, 9, and 10]. The DPLL phase control to meet the RRM phase matching case is quantified into a small number of classes to describe the mRNA-ribosome interaction in a categorical way.

Concerning the resulting lack of RRM preferences between the original and shuffled genes codons, it is concluded that either H1 is consistent with H2 or that an uneven CUF is sufficient for the gene rhythm of period 3 a single sequence wise.

An application to exons recognition and perspectives are also addressed. [1]

I. INTRODUCTION

Periodicity with period by 3 nucleotides (bases) in exons, as genes parts potentially able to encode amino acids of protein, has been well known for some time [1-4]. Firstly, reflecting correlations along coding sequence between nucleotides caused by the asymmetry in base composition at the three coding places of codons, it has been suggested that this occurs due to the universal $(RNY)_n$ pattern inherited from ancient RNA-H1 [2, 3]. Secondly, it has been proposed that the $(G-non\,G-N)_n (\sim (RNY)_n)$ pattern in the mRNA is responsible for the correct reading frame monitoring i.e. serving a RFMD [2] on the synchronization purpose of proteins translation-H2.

The H1 has been upheld by a key role of CUF in the period 3 rhythms which builds on an evolutionary algorithms outcome. Unfortunately this is not a sensitive enough way for a single gene sequence [1], whereas the RFMD –H2 holds in every translated in cell's ribosome gene case.

By the RFMD [5, 6] cell's ribosome is to act as an mRNA transcript receiver which introduces phase locking into proteins translation. Therefore the case has been found suitable to be considered in a novel, digital signal processing (DSP) way, which parallels a communication device case.

II. MATERIAL AND METHODS

Some experimental prokaryotes data on mRNA and tRNA – the transferring of amino acids according to the genetic code for their assembly into new protein - crosslink with 16S rRNA (ribosomal RNA). It has been indicated that the same two or three parts of 16S rRNA are possible binding sites of mRNA [2]. Those sites are exposed on the surface of the 30S ribosome, which is single stranded and highly conserved [2, p.650] and exhibits the *NNCNNCNNC - (NNC)* structure which is complementary to the major periodical motif $(G\,non-G\,N)_n$ of mRNA [2][2, 3]. Thus the unique translation frame alignment by G-C contacts has been singled out between mRNA and 16S rRNA and named as the reading frame monitoring device (RFMD). The RFMD is to keep a correct open reading frame (ORF) among the three possible in a single DNA strand case. This is so because the genetic code - by nucleotides triplets (codons) coding for 20 amino acids is comma-less.

The idea behind RFMD is quite close to that of the communication application, where different data patterns inserted into the main data stream serve for the purpose of synchronization to maintain limits of important information. Herein RFMD parallels synchronous receivers [8-10]. Turning to DNA, the three ORF-s of the genetic code can be observed as 0-phase, 1-phase or 2-

[1] I am very grateful to Keith Miller for improving my English

[2] C-cytosine, G-guanine, A- adenine, T-thymine (U- uracil), N-any base; R – G or A; Y – C or T

phase ORF, depending on which one of the subsequent nucleotides triple starts a codon.

To consider the RMFD by DSP approach, a mapping from the {A, T (U), C, G} set into numeric data should be done. Though there are different ways of such numeric assignment [4, 5, 7], the one which is intrinsic to the resonant recognition model (RRM, [5]) has been adopted.

The RRM builds on the finding that distribution of delocalized electron energies (DEE-s) along protein plays an essential role in determining its biological activity [5, 6]. It was found that proteins having the same biological function (same target or receptor) share the same frequencies in DEE-s. Although receptors and ligands share the same characteristic frequency, the phase at a particular frequency is opposite between receptors and ligands-binding proteins and equals to $\pm\pi\, rad$.

Since the RFMD is also an interaction of the above case, it becomes clear that the characteristic frequency at 1/3 of mutual interaction can be assumed to be due to the codon-wise frame period by 3 bases. In the RRM, properties of delocalized electron of molecular interactions are by the electron – ion interaction potential (EIIP). Simple conversion to numeric form by EIIP-s assignment requires the bases EIIP-s [Ry], which are; A – 0.1260, T (U) – 0.1335, C – 0.1340 and G – 0.0806 [5].

Due to such a mapping, genes data, e.g. GenBank files, can be digitally processed. In relation to sampled time-waveforms, discrete time *n* of signal sample is substituted by a nucleotide place *n* in genomic sequence [4, 7]. Therefore both the "future" and "past" of discrete time signals find their genomic sequence equivalents by the "to the right" or "to the left" in respect of the current place *n* in the sequence.

To synthesize protein, an interaction between mRNA and cell's ribosome becomes "sticky" by their harmonic waveforms consensus in distributions of DEE-s rather than by their common pattern of interaction case. An obvious choice for the interaction frequency is at *~1/3* [Hz].

Turning to methods; to uncover rhythm at *1/3* [Hz] in a numeric series out of sampling at *1* [Hz], either a place-wise shift alleviating or the "left" – "right" symmetric digital band-pass filtering (around *1/3* [Hz]) is performed by:

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

which is Z-inverse of causal - $H_{BP}(z)$ and non-causal - $H_{BP}(z^{-1})$ FIR filters cascade case with $N_{FIR}+1$ taps of each. A filtered genomic signal $\mathbf{x}$ of length N is changed into the analytical one with the aid of a Hilbert transformer; $\mathbf{x_a} = \mathbf{x} + \mathbf{j}\,\mathbf{x}_H = \{x_a(n) =$

$$= |x_a(n)| \exp[j\varphi_a(n)] * \exp(j\frac{1}{3}2\pi n),\ n = 1,..., N\ \} \ (2).$$

Analytical $\mathbf{x_a}$ of (2), including Hilbert transformed $\mathbf{x}_H$, is at ~ *1/3* Hz. A counterpart rhythm of cell ribosome is accounted for by a conjugated rhythm (opposite stream running) in respect of (2) rhythm case which also includes a control term to allow a phase matching with the previous case. That is by

$$\{r(n) = \exp(\mp j\pi) * \exp[-(j\frac{1}{3}2\pi n + \varphi(n-1))],$$

$$n = 1,..., N\ \} \qquad (3).$$

Either φ_a of (2) or φ of (3) stands for a reading frame shift conditioned on place *n,* whereas a place change by one base is equivalent to the phase shift by $2\pi/3$ [rad]. The RRM opposite phase consensus term has also been accounted for in (3) as a phase bias. With all the derivations in hand, one can consider a matching of the ribosome and the reading frame of m-RNA case by a phase-lock of two signals; the translation incoming one of (2) and the reference one of (3). That phase-lock of the contacting RRM frames occurs with a misalignment phase difference lack, so is by; $\varphi^{err}(n) \rightarrow 0$ (4a).

The error φ^{err} is by the product of signals (2) and (3) i.e. by;

$$\varphi^{err}(n) = angle[-x_a(n) * r(n)] = \varphi_a(n) - \varphi(n-1) \ (4b).$$

Accordingly, an answer on how the reading frames of a ribosome and of mRNA are matched together is by a phase control correction term of (3) - φ, which is adjusted by: $\varphi(n) = \varphi(n-1) + K_1\varphi(n)^{err} \quad n = 1,...$ (5).

The adaptive algorithm of (5) is that of a digital-phase locked loop (DPLL) [8, 9 and 10] of I-order, whereas loop filter gain is by K_1.

In view of (4b) φ^{err} stands an instantaneous correction towards the RRM perfect "stickiness" case, while a net correction is by φ-contributing to the RRM image, that of mRNA and the cell's ribosome. This image is further quantized to categorize the molecules interaction in the eukaryotes case, whose ribosome is similar but larger

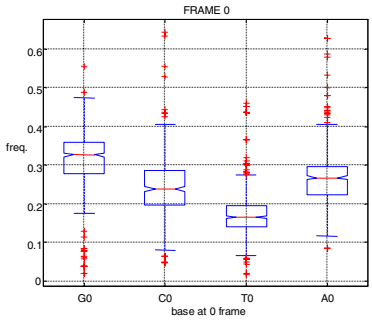
.

III. RESULTS

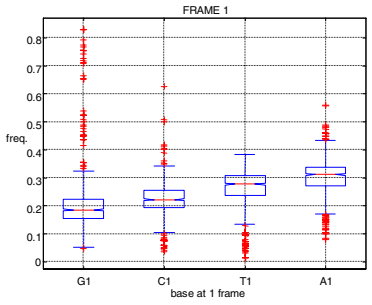
To mimic mRNA transcript, exons of every gene from the well known Bursét and Guigo collection of 570 vertebrates' genes were joined together and scanned for their RRM phase image by Sec.2 DPLL algorithm. The FIR band-pass filter quality factor was set ~100 at 1/3 Hz. The loop tracking as to fast or slow phase error correcting was kept by $K_1 \cong 0.5$. Required net phase corrections to maintain the RRM interaction "stickiness" have been

mapped into I-II-or III-category by phase image events of the net phase- φ belonging to $\pm a2\pi/3$ [rad] intervals whereas a takes 1/4, 1/2 and 1 respectively. The last category, IV, was to comprise net phase error events outside $\pm 2\pi/3$ radians i.e. of genes whose net frame shift was greater than by ±1 base. Across the genes collection results are by; **1**) codon nucleotides usage in all of the 3 codon places (in Figs 1-a,-b,-c); **2**) the categories histograms of the genes phase images (in Figs 2-a,-b) and; **3**) mRNA-ribosome interaction amplitude (Fig.3). In **2-3** the codons in normal order-i), randomly permuted–ii) and of complementary sequences–iii) are of concern. The median lengths of observed genes were; ~1050 b. in the IV, ~800 b. in the III, ~730 b. in the II and ~450 bases in the I-category.

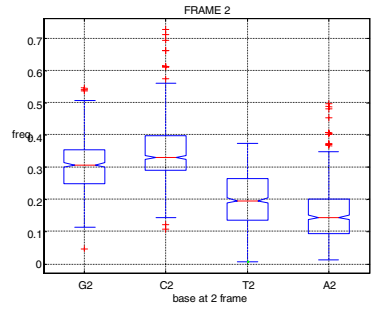
Figs.2a-b across-genes-collection-comparison is in a normal (a) and randomly permuted (b) codons case.



a)

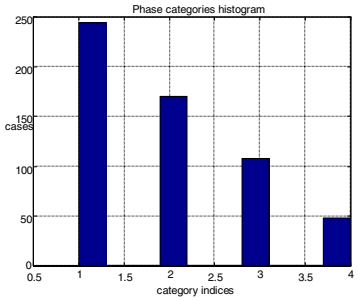


b)

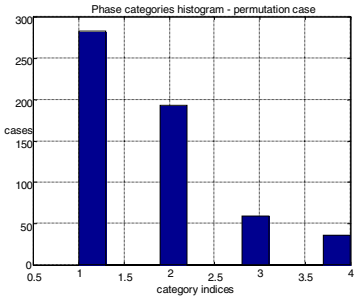


c)

Fig.1. G, C, T/U and A -bases frequency respectively in I -a), 1I -b) and III-c) place of codon across the vertebrates' genes set.



a)



b)

Fig.2.Genes categories of phase histograms; in normal (a) and permuted codons (b) case.

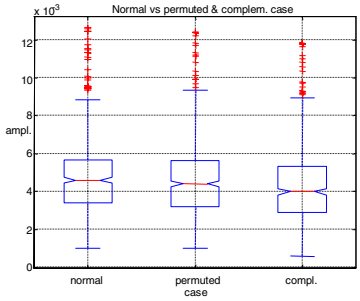


Fig.3.The interaction amplitude in the normal – (i), permuted –(ii) and complementary strand –(iii) case.

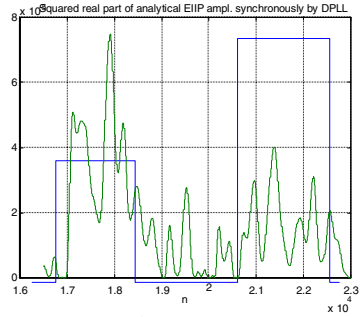


Fig.4a. *S.c.* genes; 1st-on Watson's (of lower rectangle) and 2nd- on complementary Crick's (of higher rect.) strand and their amplitudes detected synchronously on Watson's strand.

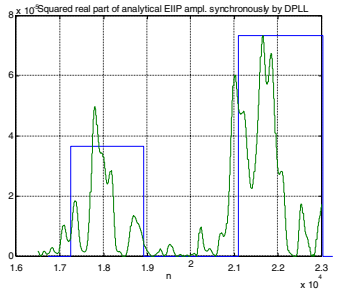


Fig.4b. *S.c.*; the same as in Fig.4a but with the amplitudes detected synchronously on Crick's strand.

IV. DISCUSSION

Also in a eukaryotes case a *G/R- non G–N* pattern of Sec.1 is preserved in view of G, C, T/U and A bases preferences in the first, second and third codon place case as shown in Figs.1a-c respectively.

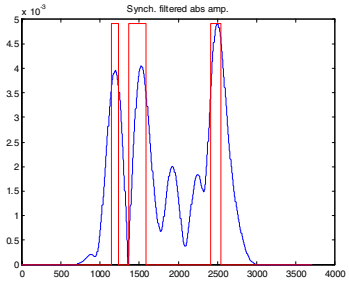


Fig.5**.** Detected amplitude of **PIKEGLOBIN** gene whose Gene Bank exons are by rectangles

Since in view of Figs.2a-b and Fig.3 a conserved within codons permutation CUF case is followed by the relevant RRM "stickiness" classes similarity, one can conclude that non-uniform CUF is sufficient either for a plausible mRNA – ribosome RFMD translation case or the period 3 rhythm origination. This is so since the method considered so far can also detect the period 3 rhythms and those slightly deviated in frequency in a way of PLL [10] based spectrum analyzer without any RRM interpretatation imposed. Hence the above appear as the H1 and H2 compatibility evidence case from a single-sequence-wise RRM study. An application to genes exons recognition is addressed in a Fig.5 sample gene case.

Also a Fig.3 difference between a normal (i) and a complementary (iii) case of the RRM amplitudes has shown discriminating preferences for a correct strand of gene location; e.g. *S.cerevisiae* genes of Figs.4a-b are by their greater amplitudes on their just strands of their location respectively. It was discovered, too, that the RRM phase category correlates with a gene's length, so that, by more frequent phase error events, interaction breaks can be attributed to longer proteins as the "technical" ones in their assembly near the ribosomes[3].

In view of this paper, optimal communication signals detection seems to parallel the cell's translation by the RRM and complete the coding theory approach to ribosomal translation [11].

However, 1) there are genes without the period 3 properties [7] and 2) one can show a CUF sufficiency for the period 3 rhythms by a single seq. sensitive method not related to the ribosome by polyphase but deviated in frequency signals [12]. For this reason further biological evidence on the RFMD could reliably exclude it from an Occam's razor case i.e. expression comparison of genes which code for similar proteins but differ by the period 3 rhythm pattern can provide some new insight into the problem.

.

V. REFERENCES

[1] S.T. Eskesen, F.N. Eskesen, B.Kinghorn and A. Ruvinsky, "Periodicity of DNA in exons", *BMC Molecular Biology*, pp.5-12, 18 August 2004.

[2] E.N. Trifonov, "Translation Frame Code and Frame-monitoring Mechanism as Suggested by the Analysis of mRNA and 16S rRNA Nucleotide Sequences", *J.Mol.Biol.* pp.643-652, 1987.

[3] E.N. Trifonov,"3-, 10.5-, 200- and 400-base periodicities in genome sequences", *Physica A*, (249), N-H Elsevier, pp.511-516, 1998.

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

[5] I. Cosic, "Macromolecular Bioactivity: Is it Resonant Interaction between Macromolecules? - Theory and Applications", *IEEE Trans. Biomed. Eng.,* IEEE, NJ, pp. 1101-1114, Dec. 1994.

[6] H.Pirogova, Q.Fang, M.Akay and I.Cosic, "Investigation of the Structural and Functional Relationships of Oncogene Proteins", *Proc. IEEE,* IEEE, NJ, pp.1859-1867, Dec. 2002.

[7] P.P. Vaidyanathan, "Genomics and Proteomics: A signal Processor's Tour", *IEEE Circuits and Systems Mag*., IEEE, NJ, pp.6-29, 4th Quart. 2004.

[8] W.C. Lindsey, C.M. Chie, "A survey of digital phase-locked loops", *Proc. IEEE*, IEEE, NJ, pp.410-431, 1981.

[9] M. Zoltowski, "Some advances and refinements in digital phase-locked loops (DPLL-s)", *Signal Processing*, Elsevier, The Nederlands, pp.735-789, 2001.

[10] F.M.Gardner, *Phaselock techniques, 3rd ed*., Wiley & The IEEE Press, 2005.

[11] "DNA as Digital Data, Communication Theory and Molecular Biology", *IEEE Eng. in Medicine & Biol. Mag*., IEEE, NJ, vol.25, no 1, Jan./Feb., 2006.

[12] A.Rushdi and J.Tuqan," Gene identification using the Z-Curve representation", ICASSP 2006, pp.II-1024-1027