Recent Developments in the Analysis of Nucleotide Sequences by using Nucleotide Genomic Signals

Paul Dan Cristea
Biomedical Engineering Center
University "Politehnica" of Bucharest
Bucharest, Romania
pcristea@dsp.pub.ro

Abstract—The nucleotide genomic signal (NuGS) methodology allows to use powerful signal processing methods in the analysis of genomic data. The method reveals surprising global regularities in the distribution of nucleotides and pairs of nucleotides along genomic sequences in archaea, bacteria and eukarya. Both manifest and hidden regularities can be detected, the method being also effective for the local study of nucleotide sequences, such as in the analysis of pathogen variability. This is important for detecting the development of pathogen resistance to treatment.

Keywords-Nucleotide genomic signals, Nucleotide imbalance, Nucleotide pair imbalance, Genomic symmetries.

I. INTRODUCTION

The paper gives an overview of our current work on Nucleotide Genomic Signal (NuGS) representation and analysis and reports several new results obtained by using this approach. The NuGS methodology is based on the conversion of symbolic nucleotide sequences into digital genomic signals [1]-[3]. There have been several attempts to represent nucleotide sequences as digital signals, most using some specific property of the nitrogenous bases, e.g., the electron-ion interaction potential [4], as the key property to identify the nucleotides. Unfortunately, the resulting representation is *biased*, *i.e.*, it is adequate specifically for the study of the phenomena in which the chosen property is essential. The representation we defined and used is *unbiased*, meaning that it is adequate for a large range of problems related to DNA analysis. To achieve this versatility, the representation is not based on the *cardinality* of numbers – their capacity to express and handle quantities, but their *ordinality* – the capacity to classify objects, to specify and handle the order of objects [1]-[3].

Consequently, the NuGS methodology can be used successfully for a large range of problems, from global to local analysis of nucleotide features. The large scale analysis of genomic sequences reveals symmetries of extant DNA sequences, maintained for entire genomes [5], but also ancestral features, which existed in DNA sequences, but disappeared during evolution [6]. The local analysis of nucleotide sequences is important for the study of gene dynamics [7], especially in the context of pathogen variability [8]-[12], leading to the development of pathogen resistance to treatment [10]-[17]. Such results can help in the fast diagnosis and early assessment of drug efficiency, allowing a systematic use of the recent advances in molecular medicine to support clinical decision.

The striking regularities of the NuGSs make a genome appear to be more than a *plain text,* as it satisfies restrictions evoking the *rhythm* and *rhyme* of poems. The regularities also allow to *predict* the nucleotides in a DNA sequence, using a methodology similar to time series prediction, and to estimate the cell self repair potential in processes such as replication, transcription or crossover [3].

The following sections of the paper present a brief reminder of the basics of NuGS methodology, essentially consisting in the conversion of symbolic nucleotide sequences into digital genomic signals and give typical results for archaea, bacteria and eukarya genomes. Most features revealed this way would be difficult to detect with the standard approach using only symbolic sequences and statistical analysis [18, 19].

II. DNA STRUCTURE AND NUCLEOTIDE CLASSES

A nucleotide molecule comprises two strands which wind one around the other to form a double helix. Each strand is a heteropolymer containing four different, but closely related, monomers – the *nucleotides*. Each nucleotide comprises three elements: a deoxyribose (a sugar), a phosphate, and a nitrogenous bases: A – adenine, C – cytosine, G – guanine, and T – thymine (or U – uracil, for RNA sequences). The nitrogenous base gives the identity of a nucleotide. According to Crick-Watson paradigm, the two strands are complementary, as only the pairs A–T and C-–G normally exist [20].

The four nucleotides can be arranged in classes according to the three main dichotomies in their biochemical properties:

(1) molecular structure – A and G are purines (R), containing two cycles, while C and T are pyrimidines (Y), comprising only one cycle;

(2) strength of the link between the paired nucleotides on the two strands – the bases A and T are linked by two hydrogen bonds (W - weak bond), while the bases C and G are linked by three hydrogen bonds (S - strong bond);

(3) radical content – A and C contain an amino (NH3) group (M class) in the major grove, while T and G contain a keto (C=O) group in that position (K class).

To fully conserve the symmetry of the nucleotides and to classify them according to the three mentioned dichotomies, we have defined a vector nucleotide tetrahedral representation [1]. After a recursive process aimed at simplifying the nucleotide representation, it has been found that it was possible to give up the less important amino-keto separation, without loosing information, and use a simpler two-dimensional nucleotide representation.

III. Nucleotide Genomic Signals

The mapping we have used [1] attaches complex numbers (*a, c, g, t*) to nucleotides (adenine, cytosine, guanine and thymine, respectively), as follows:

$$a = 1 + j, \quad t = 1 - j, \quad g = -1 + j, \quad c = -1 - j. \qquad (1)$$

These complex representations of the nucleotides have the same absolute value $\sqrt{2}$, and the phases:

$$\arg(a) = \frac{\pi}{4}, \ \arg(t) = -\frac{\pi}{4}, \arg(g) = \frac{3\pi}{4}, \arg(c) = -\frac{3\pi}{4} \qquad (2)$$

Two NuGSs are directly associated to the phases (2) of the nucleotides in a sequence:

- The *cumulated phase*, which is the cumulated sum of the phases of the complex representations of nucleotides in the sequence, from the first to the current one (the h^{th} sample in the sequence of n_b bases)

$$\begin{aligned} \theta_c(h) &= \sum_{k=1}^{h} \arg(C\{Nu(k)\}) \\ &= \frac{\pi}{4}\left[3\left(n_G(h) - n_C(h)\right) + \left(n_A(h) - n_T(h)\right)\right] \\ &= \frac{\pi}{4} N(h), \ h \in \{1,2,\ldots,n_b\}, \end{aligned} \qquad (3)$$

where: $Nu(k)$ is the kth nucleotide in the sequence, $C\{Nu(k)\}$ – its complex representation (1), $n_A(h)$, $n_C(h)$, $n_G(h)$ and $n_T(h)$ – the number of adenine, cytosine, guanine and thymine nucleotides, respectively, among the first h samples of the sequence. $N(h)$ is the nucleotide imbalance, a signature of the distribution of nucleotides in the sequence;

- The unwrapped phase, which is the phase of the elements in a sequence, corrected by adding 2m□, m∈Z, Z – the set of integers, so that the absolute value of the differences between the phases of any two successive elements in the sequence becomes smaller than π

$$\begin{aligned} &\theta_u(1) = \arg(C\{Nu(1)\}), \\ &\theta_u(h) = \arg(C\{Nu(k)\}) + 2m\pi, \ m \in Z, \\ &\text{so that } \left|\theta_u(h) - \theta_u(h-1)\right| < \pi, h \in \{2,\ldots,n_b\}. \end{aligned} \qquad (4)$$

For the mapping (1), the unwrapped phase can be expressed in terms of n_+ - the number of positive pairs (A→G, G→C, C→T, T→A) and n_- - the number of negative pairs (A→T, T→C, C→G, G→A) formed by the first *h* samples of the sequence, $h \in (2,\ldots, n_b\}$ [2]:

$$\begin{aligned} \theta_u(h) &= \theta_u(1) + \frac{\pi}{2}[n_+(h) - n_-(h)] \\ &= \theta_u(1) + \frac{\pi}{2} P(h), \quad h \in \{2,\ldots,n_b\}. \end{aligned} \qquad (5)$$

$P(h)$ is the *nucleotide pair imbalance* – a signature of the distribution of the pairs of nucleotides in the sequence. For long sequences $\theta_u(1)$ is negligible. Because of their direct statistical significance, it is convenient to use the nucleotide imbalance (*N*) and the nucleotide pair imbalance (*P*), instead of the cumulated phase (θ_c) and the unwrapped phase (θ_u), respectively.

IV. Regularities of Genomic Signals

We give below several typical examples of phase analysis results for the NuGSs of eukaryote, archaea and bacteria genomes.

Fig. 1 presents the nucleotide imbalance (*N*) and the nucleotide pair imbalance (*P*) along the DNA sequence of *Homo sapiens* chromosome 11, downloaded from GenBank [19].

N is approximately zero at the scale of the figure. This is in accordance to Chargaff's second law [21], which states that the frequencies of occurrence of purines and pyrimidines along an eukaryote DNA molecule tend to be the same in (3), so that n_G - n_C and n_A - n_T , hence *N*, are close to zero.

P has an almost linear variation along the about 140 million nucleotides of the chromosome. According to (4), the linearity of *P* corresponds to a uniform statistical excess of the n_+ pairs with respect to the n_- pairs. This is a general characteristics of *all* investigated genomes. We have shown ([2, 6]) that recombination conserves this regularity, while local random mutations, such as uncorrelated SNPs (single nucleotide polymorphisms), would destroy it.

The readily noticeable non-regularity in the interval 15,172 – 15,376 Kbp in Fig.1 is probably caused by an error. Better curated (phase 3) releases have an even higher linearity, as can be seen in Fig. 2 that gives θ_c and θ_u for chromosome 1, of *H. sapiens* that contains more than 228 million nucleotides.

The NuGSs of archaea and bacteria show even more pregnant regularities, with typical features for each taxon and with parameters defining a "physiognomy' of each genome. Fig. 3 gives the *N* and *P* signals for the complete genome of *Escherichia Coli* K12, accession U00096 [19], the first completely sequenced genome.

The nucleotide pair imbalance *P* is also a linear function, like in the case of eukaryotes, but the slope is negative.

A significant change appears for the nucleotide imbalance *N*, which is no longer close to zero along the entire sequence, but piece-wise linear. The circular DNA of *E.coli* is divided by the features of *N* in two

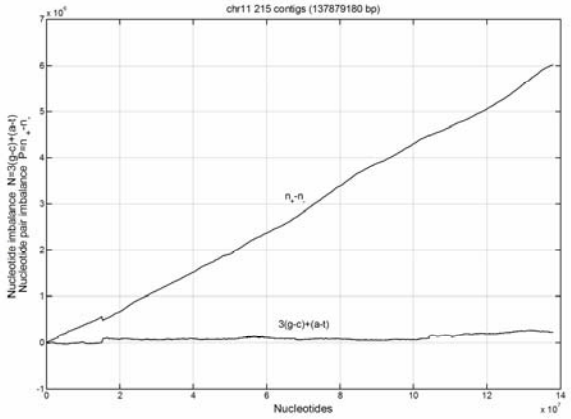

Figure 1. Nucleotide imbalance (N) and nucleotide pair imbalance (P) along concatenated contigs of Homo sapiens ch.11 ([19]).

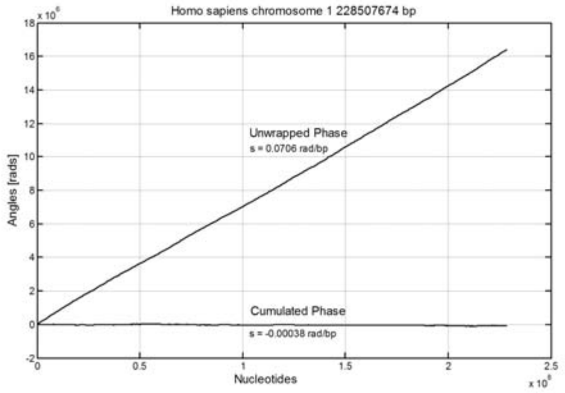

Figure 2. Cumulated phase and unwrapped phase for “phase-3” data of *Homo sapiens* chromosome 1 [19].

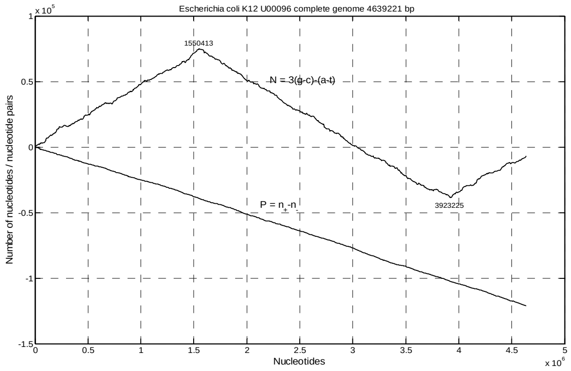

Figure 3. *N* and *P* for *Escherichia Coli* strain K12 (U00096 [19], 2009, length 4,639,221 bp).

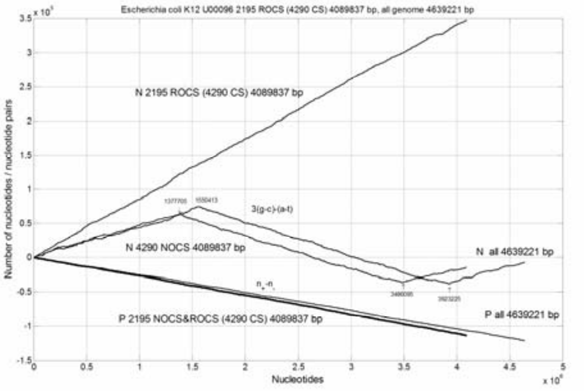

Figure 4. *N and* P for the complete genome of *E. Coli* K12 (U00096 [19]), before and after re-orienting its coding segments.

distinct segments: one having $3n_G + n_A$ in excess to $3n_C + n_T$, for which (3) gives a positive slope of *N*, the other having the opposite property. The separation points have a biological significance: the minimum of *N*, at 3,923,225 bp, corresponds to the *origin of genome replication*, whereas the maximum, at 1,550,413 bp, corresponds to the *terminus of replication*.

Two main features, shared by most bacteria, can readily be identified in Fig.3: (1) A remarkable good linearity of *P*, characterized by a very small root mean square error per nucleotide of the linear fitting to *P* and a large ratio of the resulting linear per non-linear variation, corresponding to a smooth strait line; (2) An approximately piece-wise linear *N*, locally affected by recombination. Recombination, as well as specifically correlated single nucleotide polymorphisms (SNPs), change *N*, but not *P*, which remains linear for all genomes, including archaea, bacteria and eukarya [2].

Re-orienting all exons of a sequence in the same direction reveals a surprising “hidden” symmetry of DNA molecules [5]. Figure 4 presents the NuGSs before and after such a re-orientation for the complete genome of *E.coli*. The following signals are shown:

- *N* and *P* for the whole genome (length 4,6392,21 bp), also represented in Fig. 3;
- *N* and *P* for the concatenated 4,290 exons, with a total length of only 4,089,837 bp, after extracting the non-coding segments, but maintaining the initial orientation of the exons in the genome. These sequences are called NOCS – non-re-oriented coding segments;
- *N* and *P* for the concatenated 4,290 exons, after the re-orientation of the 2195 exons that where in the negative direction in the sequence downloaded from GenBank [19]. These sequences are called ROCS – re-oriented coding segments.

The most remarkable result is the approximately linear shape of *N* after the re-orientation of the exons (ROCS), which suggests a highly regular ancestral genomic structure, from which the current nucleotide structure, described by a piece-wise linear *N* shape specific to each species, has evolved. The nucleotide imbalance *N* is linked to a potential governing the transversal interactions between DNA-DNA and DNA-protein molecules. To allow only specific transversal intermolecular interactions in processes such as replication, transcription and crossover, a well defined potential along the DNA molecule and, correspondingly, a well defined *N* signal is requested.

The NOCS and ROCS plots of *P* are superposed, because of the invariance of *P* resulting from the conservation of the direct (n_+) and inverse (n_-) numbers of nucleotide pairs when reversing a segment of a DNA double helix, while simultaneously switching its strands to maintain the 5’→ 3’ positive sense ([1], [2]).

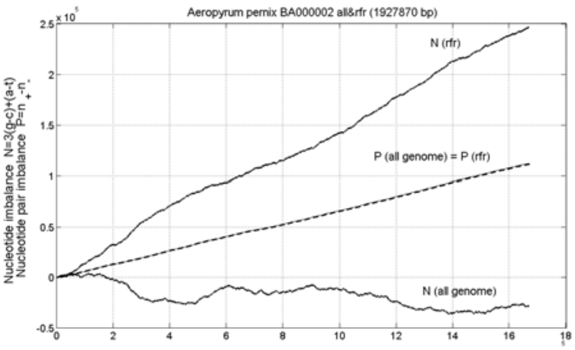

Figure 5. *N* and *P* of *Aeropyrum pernix* (BA00002 [19]) complete genome, before and after re-orienting (re-framing) its exons.

As already mentioned, *P* is also linear for archaea, but *N* is irregular, as shown in Fig. 5 for *Aeropyrum pernix* (BA00002 [19]), an archaea that lives in hot springs. The figure comprises the *N* and *P* signals for the complete genome (1,669,695 bp), initially (all) and after re-framing (rfr) – the re-orientation of its 1428 negatively oriented exons, while leaving unchanged the 1227 positively oriented exons and the non-coding segments. When all exons are arranged in the same direction, the signal *P* remains unchanged, while the initially (irregularly) decreasing *N* becomes an (almost) linearly increasing signal. These results support the hypothesis of a regular ancestral genomic structure, existing in both archaea and bacteria.

V. Prediction of Nucleotide Sequences

Another way to explore the correlations and regularities in the genomic signals is to predict the nucleotides in a DNA sequence based on the knowledge about the preceding nucleotides, in a way similar to *time series prediction* [3]. Such an approach can be seen as an evaluation of the feasibility of nucleotide error correction at the level of DNA replication and transcription, or RNA translation. Despite the involvement of vastly distinct systems in these processes, there are many functional similarities.

The nucleotide prediction methodology can be improved, reducing both its complexity and the required computing time, by using a two step procedure, as shown in the block diagram of Fig. 6. The first step uses a principal component analysis (PCA) block which retains only the high variance components of the input signal, those necessary to reconstruct it with a given accuracy. The second step uses an artificial neural network (ANN) to perform the actual prediction of the next nucleotide. After training the system on signals satisfying a rather mild regularity condition (a statistical periodicity modeled by a circulate matrix input, condition satisfied by most "natural" signals), the PCA stage actually performs a DFT of the input, passing from the time (space) domain, to frequency domain. Correspondingly, the ANN generates the estimate of the next sample in the sequence by implementing a (restricted) inverse DFT transform.

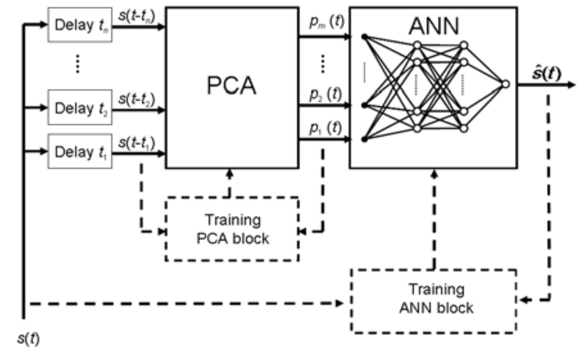

Figure 6. Two stage structure of a nucleotide prediction system [3].

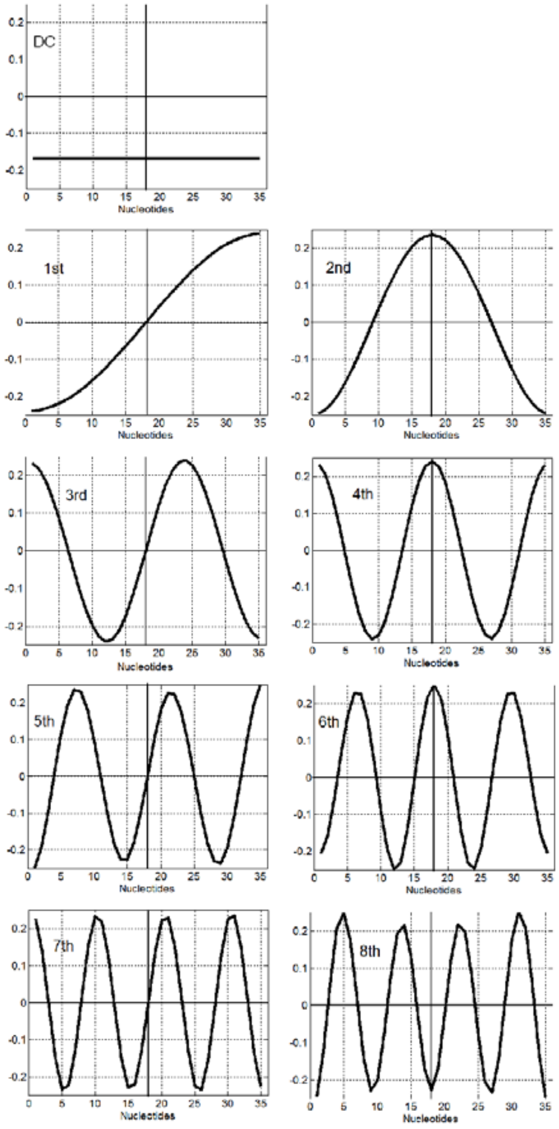

Figure 7. Content of the rows in the matrix describing the functioning of the PCA block in Fig. 6.

Fig. 7 shows the content of the matrix describing the functioning of the PCA block, after its training on „natural" signals. The practically important result is that the PCA matrix actually implements the DC and the first eight harmonics of a Discrete Fourier Transform (DFT) of the input signal [3]. This property allows to further reduce the complexity of the prediction system in Fig. 6, by replacing the PCA block with its pre-determined Fourier equivalent, thus avoiding a timeconsuming training step.

The prediction model shows a quite good efficiency, which is the effect of the multilevel regularities in the structure of genomic sequences.

The resulting architecture can be used to analyze and predict any "natural" signal, especially in the case of time series prediction.

VI. Retro-Viral Inserts in Bacteria

A clear example on how an insert of exogenous genetic material into the genome of a host organism can be revealed by using NuGS methodology is provided by the case of *Bacillus phage* SPBc2 complete genome (accession AF020713 [19, 22]) inserted into the complete genome of *Bacillus subtilis* (accession AL009126 [19, 23]). The plot marked '*wild*' in Fig. 1 gives the nucleotide imbalance *N* for the complete genome of *Bacillus subtilis* in the case of a bacterium attacked by the SPBc2 bacteriophage. The *Bacillus subtilis* (*Bs*) genome comprises 4,214,814 bp and 4,101 exons. The nucleotide imbalance *N* of the wild genome has two almost linear branches, delimited by the origin of replication (the superposed initial and final point of the circular *Bs* genome) and the terminus of replication (the maximum of *N* at 1941693 bp). As shown above, the 'hidden' ancestral structure of a genome can be revealed by re-orienting all exons in the same direction.

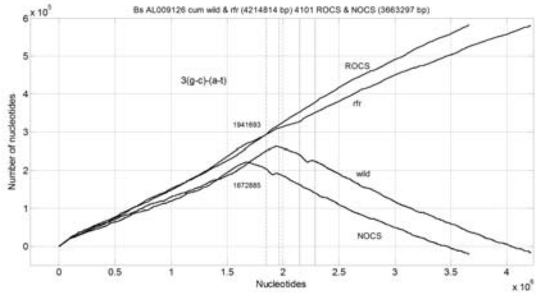

Figure 8. Nucleotide imbalance *N* of *Bacillus subtilis* (AL009126 [19, 23]): wild – complete genome (4,214,814 bp), rfr – reframed genome, NOCS – non-re-oriented coding segments (4,101 concatenated exons, 3,663,297 bp), ROCS – re-oriented coding segments.

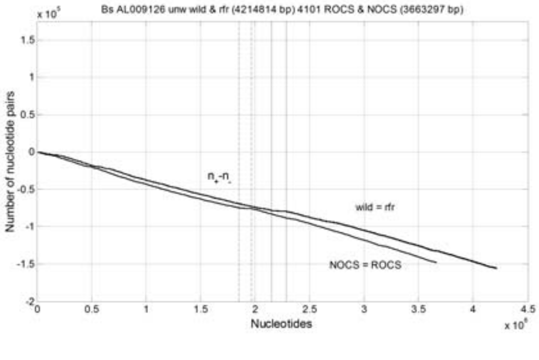

Figure 9. Nucleotide pair imbalance $P = n_+ - n_-$ of *Bacillus subtilis* (AL009126 [19, 23]): wild & rfr – complete and reframed genome, NOCS & ROCS – non-re-oriented and re-oriented coding segments. The four lines are superposed in two almost identical pairs. The insert initially in the interval 2151273-2285688 bp is marked.

It is noteworthy that, in the case of *Bs,* the introns contribute in increasing the linearity of the wild genome *N* signal. The NOCS (non-re-oriented coding segments) signal becomes the almost strait line, denoted by ROCS (re-oriented coding segments) in Fig. 8. The ROCS line has only one domain for the whole genome, a higher slope ($s_N \approx +0.1656$) and a better linearity. The distinct behavior in the region of the insert almost disappears, suggesting a putative common origin for the genetic material of the host and of the intruder.

As expected, the curve resulting after the re-orientation of the reading frames for the exons in the same positive direction, while keeping the introns unchanged, is less linear and displays two distinct domains. The nucleotide pair imbalance *P* for the complete genome ('wild') of *Bacillus subtilis*, the signal for the re-framed sequence ('rfr'), as well as for the non-re-oriented ('NOCS') and re-oriented ('ROCS') coding regions are given in Fig. 8. There are four lines, superposed in two pairs, as the re-orientation of the segments of a nucleic acid molecule conserves the distribution of the pairs of nucleotides and keeps constant the *P* signal. The region of the insert is clearly distinct from the rest of the host genome, revealing second order structural differences between the host and the insertion.

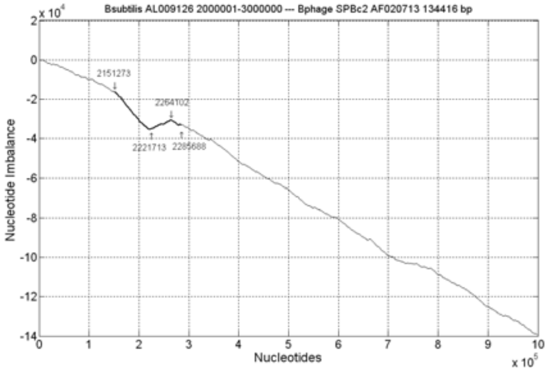

Figure 10. Aligned *N* signals of the 2000001-3000000 bp interval for the *B. subtilis* genome (AL009126 [19, 23]) and the inversed *Bacillus phage* SPBc2 complete genome (AF020713 [19, 22], 134,416 bp).

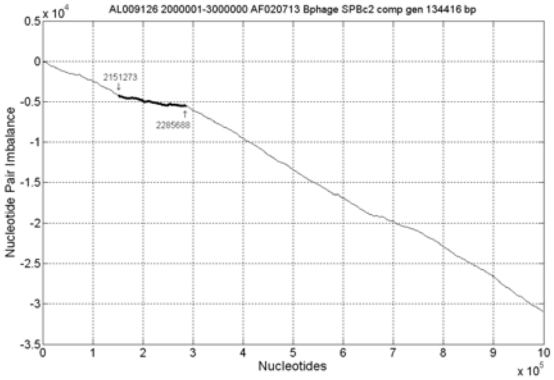

Figure 11. Aligned *P* signals of the 2000001-3000000 bp segment of *B. subtilis* genome (AL009126) and of the inversed *Bacillus phage* SPBc2 complete genome (AF020713).

A BLAST [28] search of the genomic databases [] for the segment 2151000 - 2264000 bp, approximately corresponding to the insert in Fig. 8 and Fig. 9, apart of the expected hit to the *Bs* complete genome (AL009126 [19, 23]) and with several other strains of *Bs*, returns a perfect match with a sequence covering most of the *Bacteriophage* SPBc2 genome (AF020713 [19, 22]), with a total score of 1.354e+05. Good matches also result for other bacteriophage genes, primarily for the SP-beta methyltransferase gene.

The *N* signal is approximately piece-wise linear. The local regularity of the SPBc2 genome is higher than of the neighboring areas of the *Bs* genome. On the other hand, the *P* signal is less regular. As mentioned above, it appears that, after proper re-orientation and alignment, the viral complete genome *N* and *P* signals superpose exactly over the domain 2151273-285688 bp of the *Bs* genome, as shown in Fig. 10 and Fig. 11.

VII. Conclusions

The study of nucleotide sequences by using NuGSs reveals regularities in the structure of DNA and RNA molecules. This approach has been applied to describe the structure of nucleotide sequences, both in the current state of extant species and in a putative ancestral state, from which they could have evolved.

The NuGSs method also provides a natural and efficient framework for the prediction of nucleotide sequences. Such a methodology can help locate areas in DNA and RNA molecules that have distinct statistical features and specific functionalities such as inserts.

References

[1] P. D. Cristea, "Conversion of Nitrogenous Base Sequences into Genomic Signals," Journal of Cellular and Molecular Medicine, vol. 6, no. 2, pp. 279–303, 2002.

[2] P. D. Cristea, "Chapter 1: Representation and analysis of DNA sequences," in Genomic Signal Processing and Statistics, E. Daugherty, I. Shmulevich, J. Chen, and Z.J. Wang, eds, Eurasip Book Series on Signal Processing and Communications, Hindawi Publ. Corp., pp. 15–65, 2005.

[3] P. D. Cristea, Rodica Tuduce, J. Cornelis, R. Deklerck, I. Nastac, M. Andrei, "Signal Representation and Processing of Nucleotide Sequences," Proceeding of the 7th IEEE Intl. Conf. on Bioinformatics and Bioengineering (IEEE BIBE 2007), pp. 1214-1219, Harvard Medical School, Boston, USA, October 14-17, 2007.

[4] I. Cosic, "Macromolecular bioactivity: Is it Resonant Interaction between Molecules? - Theory and Applications," IEEE Trans. On BME, vol. 41, pp. 1101-1114, 1994.

[5] P. D. Cristea, "Large Scale Features in DNA Genomic Signals," Signal Processing [Special Issue on Genomic Signal Processing], Elsevier, vol. 83, pp. 871–888, 2003.

[6] P. D. Cristea, "Genomic Signals of Re-Oriented ORFs," Eurasip – Journal on Applied Signal Processing, [Special Issue on Genomic Signal Processing], vol. 2004, no.1, pp. 132–137, January 1, 2004.

[7] P. D. Cristea, "Invariants of DNA Genomic Signals," in Biomedical Applications of Micro- and Nanoengineering II, edited by D.V. Nicolau, Proceedings of SPIE, vol. 5651, SPIE, Belingham, WA, pp.115-125, 2005.

[8] P. D. Cristea, "Genomic Signal Analysis of Mycobacterium tuberculosis," Progress in Biomedical Optics and Imaging, Proceedings of SPIE, vol.6447, pp. C-1 – C8, 2007.

[9] S. T. Cole, R. Brosch, J. Parkhill, et al., "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence," Nature, vol. 393, no. 6685, pp. 537-544, 1998.

[10] P. D. Cristea, D. Otelea, Rodica Tuduce, "Genomic Signal analysis of HIV variability," SPIE - BIOS 2005, Proceedings of SPIE, vol. 6, no. 14, pp. 362-372, 2005.

[11] P. D. Cristea, D. Otelea, Rodica Tuduce, "Study of HIV Variability Based on Genomic Signal Analysis of Protease and Reverse Transcriptase Genes," EMBC'05 - 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Shanghai, China, IEEE Catalog no: 05CH37611; ISBN: 0-7803-8740-6, ISSN: 1094-687X, paper 1845, September 1-4, 2005.

[12] P. D. Cristea, "Genomic Signal Analysis of Pathogen Variability," Progress in Biomedical Optics and Imaging, Proceedings of SPIE, vol. 6088, pp. P1-P12, 2006.

[13] P. F. Barnes, D. L. Lakey, and W. J. Burman, "Tuberculosis in patients with HIV infection," Infectious Disease Clinics of North America, vol. 16, pp. 107–126, 2002.

[14] J. M. Musser, "Antimicrobial Agent Resistance in Mycobacteria: Molecular Genetic Insights," Clinical Microbiology Reviews, pp. 496–514, Oct., 1995.

[15] A. Telenti et al, "Detection of rifampin-resistance mutations in Micobacterium tuberculosis," Lancet, vol. 341, pp. 647-650, 1993.

[16] M. J. Torres et al, "Rapid Detection of Resistance Associated Mutations in Micobacterium tuberculosis by LightCycler PCR," J. Clin. Microbiol., vol. 3194-3199, 2000.

[17] I.C. Shamputa, L. Rigouts, F. Portaels, "Molecular genetic methods for diagnosis and antibiotic resistance detection of mycobacteria from clinical specimens," APMIS, vol. 112, pp. 728–752, 2004.

[18] S. F. Altschul, T. L. Madden, et. al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, vol. 25, pp. 3389-3402, http://www.ncbi.nlm.nih.gov/blast/, 1997.

[19] GeneBank, NIH - National Institutes of Health, National Centre for Biotechnology Information, National Library of Medicine, (NCBI/GenBank), http://www.ncbi.nlm.nih.gov/, 2009.

[20] J. D. Watson, F. H. C. Crick, "A structure for deoxyribose nucleic acid," Nature, vol. 171, no. 4356, pp. 737–738, 1953.

[21] E. Chargaff, "Structure and function of nucleic acids as cell constituents," Federal Proceedings, vol.10, pp. 654–659, 1951.

[22] V. Lazarevic, B. Soldo, A. Dusterhoft, H. Hilbert, C. Mauel and D. Karamata, "Introns and intein coding sequence in the ribonucleotide reductase genes of *Bacillus subtilis* temperate bacteriophage *Spbeta*," Proc. Natl. Acad. Sci. U.S.A., vol. 95, no. 4, pp. 1692-1697, 1998.

[23] F. Kunst et al, "The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*," Nature, vol. 390, no. 6657, pp. 249-256, 1997.