34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Building Phylogenetic Trees by Using Gene Nucleotide Genomic Signals

Paul Dan Cristea, *Fellow, IEEE*

***Abstract*— Nucleotide genomic signal (*NuGS*) methodology allows a molecular level approach to determine distances between homologous genes or between conserved equivalent non-coding genome regions in various species or individuals of the same species. Therefore, distances between the genes of species or individuals can be computed and phylogenetic trees can be built. The paper illustrates the use of the nucleotide imbalance (*N*) and nucleotide pair imbalance (*P*) signals to determine the distances between the genes of several *Hominidae*. The results are in accordance with those of other genetic or phylogenetic approaches to establish distances between *Hominidae* species.**

I. Introduction

The Nucleotide Genomic Signals (*NuGS*s) methodology is based on the conversion of symbolic nucleotide sequences into genomic signals [1-4]. This approach allows the systematic use of signal processing methods in the analysis of genetic information. The genomic signal conversion used in this paper is a general purpose one-to-one mapping of symbolic genomic sequences to complex signals [4]. The *NuGS* methodology has been developed initially to reveal large scale features of DNA molecules [5,6] by analyzing genomes from GenBank [7], but it has been found that the approach is also adequate for the local analysis of DNA, in applications such as the study of pathogen variability on DNA extracted from patients [8-11], to detect the development of resistance to drugs [12]. This paper further develops on the application of *NuGS*s in the comparative analysis of homologous genes or conserved equivalent non-coding genome regions in various species, or in various individuals of the same species [13], specifically by establishing the distances between such entities, and by building the resulting phylogenetic trees [14].

In the following, we briefly present the basics of *NuGS*s methodology (Section II) and detail the aspects related to the use for computing distances between genomes based on the distances between selected coding or non-coding homologous regions (Section III). The procedure is illustrated for the cases of the HIV-1 Protease gene (Section IV) and for the mitochondrial DNA (mtDNA) genes of several Hominidae, including humans and the Neanderthal (Section V).

II. Genetic Information Digital Representation

There have been many attempts to represent nucleotide sequences as digital signals [1], most using some specific property of the nitrogenous bases (e.g., the electron-ion interaction potential [15]). The main purpose was to compute the Fourier transform of the signal, to detect the areas having a predominant third harmonics, which correspond to exons – the protein encoding segments [16,17]. Unfortunately, the resolution obtained this way was not satisfactory and better methods to find exons had to be developed, for instance searching for homologous or heterologous genes [18].

To analyze the nucleotide genomic sequences and to reveal their statistical symmetry, we have used a conversion from symbolic to numeric form, which maps the symbolic sequences to complex digital signals. The mapping we have used is an unbiased representation of nucleotide classes, which does not use the cardinality of numbers – their capability to express quantities, but the ordinality – the capability to order classes [2-4]. The method reveals surprising regularities, similar to Chargaff laws [19], not only in the distribution of nucleotides along DNA sequences, but also in the distribution of nucleotide pairs. The property has been found in all genomes – *archaea*, *bacteria* and *eukarya*. From the structural point of view, a genome appears to be more than a *plain text,* as it also satisfies regularities evoking the *rhythm* and *rhyme* of poems. These symmetries of nucleotide sequences would be difficult, if not impossible, to reveal by using only symbolic genomic sequences and standard statistical and pattern matching methods [20]. The mapping [21] attaches the complex numbers *a, c, g,* and *t* to the nucleotides *adenine*, *cytosine*, *guanine* and *thymine*, respectively, as it is shown in (1) and graphically represented in Fig. 1.

$$a = 1+j, \quad t = 1-j, \quad g = -1+j, \quad c = -1-j. \tag{1}$$

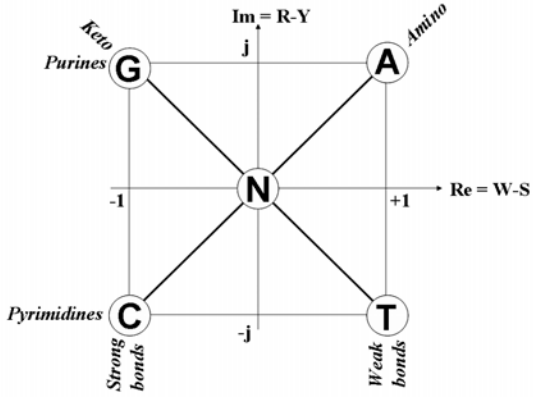

Figure 1. Complex representation of nucleotides

P. D. Cristea is with the University "Politehnica" of Bucharest, BioMedical Engineering Center, Spl. Independentei 313, 060042 Bucharest, sect.6, Romania (phone: +40-21-316 9569; fax: +40-21-316 9568; e-mail: pcristea@dsp.pub.ro).

Using the phases of the complex representations in (1), two *NuGS*s can be associated to a sequence:

- The *cumulated phase* – the sum of the phases of the complex representations of nucleotides in a sequence, from the first to the current h^{th} sample in the sequence:

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

where $Nu(k)$ is the k^{th} 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 occurrences of adenine, cytosine, guanine and thymine nucleotides, respectively, in the first h samples of the sequence of n_b nucleotides, $N(h)$ – the *nucleotide imbalance* – a *NuGS* that describes the statistics of the distribution of nucleotides in the sequence, and n_b – the number of nucleotides in the sequence;

- The *unwrapped phase* – the phase of the elements in a sequence corrected by adding $2m\pi$, $m \in Z$, Z – the set of integers, so that the absolute value of the difference of phase between two successive entries of the sequence be smaller than π.

$$\begin{aligned}&\theta_u(1) = \arg(C\{Nu(1)\}), \\ &\theta_u(h) = \arg(C\{Nu(k)\}) + 2m\pi,\ m \in Z, \\ &\text{so that } \left|\theta_u(h)-\theta_u(h-1)\right| < \pi, h \in \{2,\ldots,n_b\}.\end{aligned} \tag{3}$$

For the mapping (1) it can be shown (Cristea, 2005a):

$$\begin{aligned}&\theta_u(h) = \theta_u(1) + \frac{\pi}{2}[n_+(h) - n_-(h)] = \theta_u(1) + \frac{\pi}{2}P(h), \\ &h \in \{2,\ldots,n_b\},\end{aligned} \tag{4}$$

where n_+ is the number of positive pairs (A→G, G→C, C→T, T→A) and n_- is the number of negative pairs (A→T, T→C, C→G, G→A) formed by the first h samples of the sequence, $h \in (2,\ldots, n_b\}$, $P(h)$ is the *nucleotide pair imbalance* – a *NuGS* that describes the statistics of the distribution of pairs of nucleotides in the sequence. For long sequences, $\theta_u(1)$ is negligible. As they have a direct statistical significance, it is convenient to use the *nucleotide imbalance* (N) and the *nucleotide pair imbalance* (P) in genomic signal analysis, instead of the *cumulated phase* (θ_c) and *unwrapped phase* (θ_u), respectively.

For the comparative analysis of genomes, as used when studying gene variability, one considers a set of *NuGS*s derived from various individuals in a group, among which occur mutations. In such cases, every signal S_k, corresponding to a certain individual k (k = 1, m), can be characterized by the pair of signals: R – the *reference* against which we want to compare the set of signals, *e.g.*, a signal that expresses their common trend, O_k – the *offset* of S_k with respect to R,

$$O_k = S_k - R \ \ (k = 1, m). \tag{5}$$

A natural choice for R, *e.g.*, when analyzing *Hominidae* mtDNA genes, is to use the signal for *Homo sapiens* (*Hs*) and compare each gene with its homologue in humans. Similarly, when monitoring pathogen variability to track the development of pathogen resistance to treatment, the natural choice for R is the wild type (*WT*) – usually downloaded from a genomic database, such as GenBank [7]. This is feasible when the variability is small enough, such as in the case of *Mycobacterium tuberculosis*, so that the pathogens in the isolates from various patients are not too different from the *WT*. But in the case of highly variable pathogens, such as *HIV*, the differences between individual signals and the *WT* signal might become too large, so that the reference must be constructed in terms of the set of signals. Most times, the reference can be chosen as: (1) the *average* (mean) or any other *linear combination* of the signals in the set (weighted mean), including the choice of one of the signals in the set, (2) the *median*, and (3) the *mode step* [21]. As its name suggests, the *ModeStep* signal is built by selecting in each point the variation (*step*) that occurs the largest number of times (the *mode*) in that point in the set of signals. The starting point is the median of the starting points.

The resolution can be further improved by using the *digital derivatives* of the offsets. This is particularly useful to identify punctual mutations (one nucleotide genetic variations), or to determine the distance between the individual signals and the reference, corresponding to step variations in the offsets.

III. Distances between Genes and Species

The *distance between two homologous genes,* G_1 and G_2, from different species or from different individuals, is defined as the sum of the absolute values of the differences between the *NuGS*s, s_{G1} and s_{G2}, describing the two genes:

$$d(G_1, G_2) = \sum_{k=1}^{L} \left| s_{G_1}(k) - s_{G_2}(k) \right|. \tag{6}$$

A one nucleotide genetic variation adds 1-3 units to the distance between two genes, depending on the signal used (N or P) and the inter-changed nucleotides. When the sequences do not have the same length, and the mutations are not one-nucleotide, but multi-nucleotide substitutions, inserts or deletions, a pre-processing of the signals involving the alignment of the sequences is necessary. Currently, we have applied the method only for short nucleotide sequences, where a visual alignment using the shape of the signals was feasible and yielded reproducible results (tested by different human operators). A software, which calls BLAST [20] for alignment, is being developed for handling large sequences.

The *distance between two species* A and B, from the point of view of the genes in some specified set of genes (set$G$), is defined as the Euclidian distance in a space in which each gene in set$G$ is an independent coordinate. Thus, the distance between two species A and B is calculated as the square root of the sum of the squares of the distances between all the genes included in set$G$:

$$d(A, B) = \sqrt{\sum_{G \in \text{set}G} d_G^{\,2}}. \tag{7}$$

IV. Distances Between *HIV*-1 Protease Genes in Isolates from Three Classes of Patients

We illustrate the method of measuring the distance between homologous genes by considering the *Human immunodeficiency virus* type 1 (*HIV*-1) protease enzyme (*PR* - the proteinase) gene. The purpose is the analysis of *HIV*-1 variability under the effect of antiretroviral treatment with *PR* inhibitors, and the detection of the development of pathogen drug resistance. *HIV*-1 makes many of its proteins in one long chain, and *PR* has the essential role of cutting this polyprotein into the proper pieces, at the proper timing. Consequently, *PR* is an important target for the current drug anti-*HIV* therapy. *PR* inhibitors bind to the enzyme catalytic area and block its action. Sequences genotyped from patient isolates in the laboratory of the National Institute of Infectious Diseases "Prof. Dr. Matei Bals", Bucharest, Romania, have been used in our study. The strain typical to the *HIV* infection in Romania has been classified as *subtype F* and its drug susceptibility has been studied as early as 1997 by Apetrei *et al.* [22]. Taking into account the mutations identified in the 121 analyzed sequences, the samples were classified from the point of view of the resistance to the current antiretroviral compounds active against *PR* and *RT* (reverse transcriptase) into three groups: sensitive (S), resistant (R) and multi-resistant (M).

Fig. 2 presents the nucleotide imbalance signals (*N*) of the *PR* genes for three cases of each type, as well as the signal for a standard sequence (*std*) downloaded from the NIH GenBank database (accession number NC 001802, length 297 bp [7]). The curves are difficult to follow in this figure, because of the close similarity of the signals (even when using different colors and different line styles). Actually, this is not a problem, as the offsets and, especially, their derivatives give the details necessary to measure the distance between the genes so that the representation of the signals is not necessary for measuring the distances.

Fig. 3 gives the offsets of nucleotide imbalance signals (*N*) in Fig. 2 with respect to the *std* signal, chosen as reference. The step variations in the offsets correspond to the points where there are genetic variations with respect to the *std* gene. The distance between two genes is computed from the local differences in the considered *NuGS*s.

In Fig. 4 we have represented the digital derivatives of the offsets shown in Fig. 3. The pulses on the lines in these figures correspond to the step differences in the offsets in Fig. 3. The distance between each gene and the std gene, computed with (6), is shown at the right of the lines. All the differences between the signals for each gene and the std signal, corresponding to the WT gene, have been considered for computing these distances. In parentheses are given the distances computed based only on the genetic variations known to be linked to the resistance to PR inhibitors, for which the corresponding codons are marked by vertical strips in the figure. The clustering after the distances between the patients' and the *std* genes is not in full accordance with the initial *S, R, M* classification of the patients, which is based on clinically observed resistance to the entire antiretroviral treatment, including both *PR* and *RT* inhibitors.

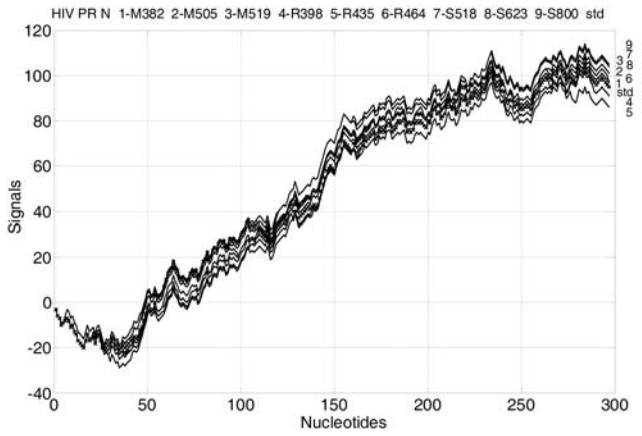

Figure 2. Nucleotide imbalance signals (*N*) of the *PR* genes for nine patients belonging to the sensitive (*S*), resistant (*R*) and multi-resistant (*M*) classes, as well as for a reference sequence downloaded from GenBank.

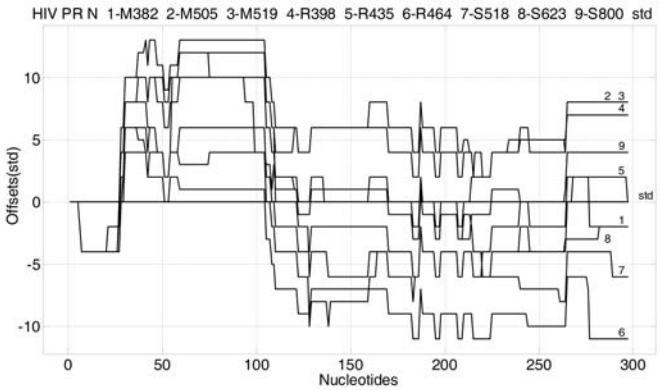

Figure 3. Offsets of the *N* signals in Fig.2 with respect to the *std* signal.

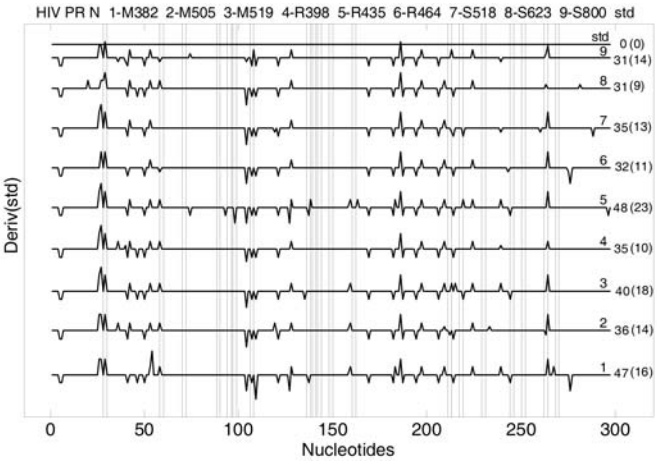

Figure 4. Derivatives of the offsets in Fig. 3. The numbers at the right of lines are the distances between each *PR* gene and the *std* gene. The numbers in parantheses are the distances produced by mutations that confer resistance to *PR* inhibitors.

Fig. 5 and 6 give the offsets and, respectively, the digital derivatives of the signals in Fig.2 with respect to the *ModeStep* reference – an expression of the common trend of the set of nine patient signals. The large distance between the *std* signal, which corresponds to the HIV-1 type B virus given in the GenBank database [7], and the patients'

ModeStep reference, which corresponds to signals for viruses that are all HIV-1 type F, is readily noticeable.

Table 1 gives the matrix of the distances between the *PR* genes of the virus isolates from the nine considered patients, whereas Fig. 8 contains the corresponding neighbor-joining phylogenetic tree built with an accessory application provided by BioEdit [14].

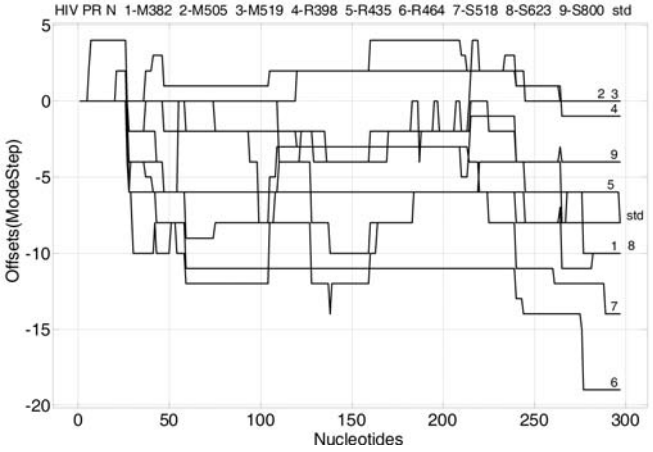

Figure 5. Offsets of the signals in Fig.2 with respect to the *ModeStep* reference.

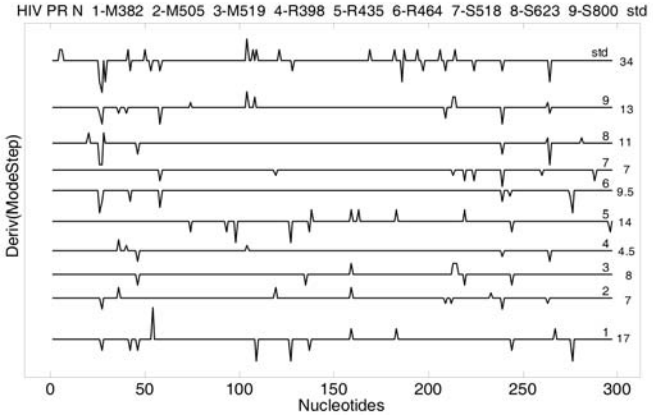

Figure 6. Derivatives of the offsets in Fig. 5. At the right of lines are the distances between each *PR* gene and the *ModeStep* reference.

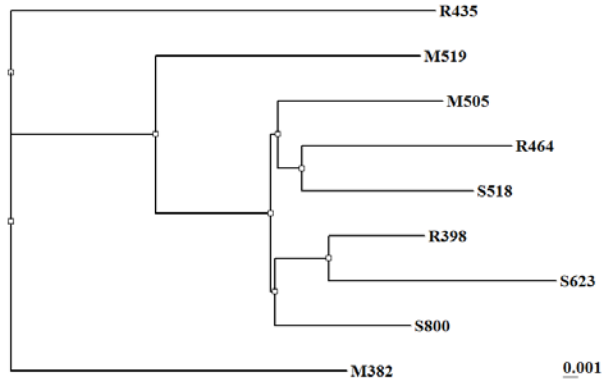

Figure 7. The neighbor-joining phylogenetic tree built with the distance matrix of the PR M382-505-519, R398-435-464, S518-623-800 genes in Table 1.

TABLE 1 MATRIX OF THE DISTANCES BETWEEN THE *PR* GENES FOR NINE PATIENS

	M382	M505	M519	R398	R435	R464	S518	S623	S800
M382	0.0000	0.0458	0.0491	0.0459	0.0488	0.0422	0.0569	0.0569	0.0573
M505	0.0458	0.0000	0.0383	0.0173	0.0525	0.0244	0.0244	0.0314	0.0210
M519	0.0491	0.0383	0.0000	0.0312	0.0522	0.0492	0.0312	0.0456	0.0279
R398	0.0459	0.0173	0.0312	0.0000	0.0526	0.0280	0.0243	0.0208	0.0208
R435	0.0488	0.0525	0.0522	0.0526	0.0000	0.0636	0.0527	0.0673	0.0531
R464	0.0422	0.0244	0.0492	0.0280	0.0636	0.0000	0.0245	0.0279	0.0246
S518	0.0569	0.0244	0.0312	0.0243	0.0527	0.0245	0.0000	0.0352	0.0210
S623	0.0569	0.0314	0.0456	0.0208	0.0673	0.0279	0.0352	0.0000	0.0244
S800	0.0573	0.0210	0.0279	0.0208	0.0531	0.0246	0.0210	0.0244	0.000

V. DISTANCES BETWEEN MTDNA GENES AND SPECIES IN THE *HOMINIDAE* FAMILY

The *NuGS* methodology has also been applied to analyze the mitochondrial DNA (mtDNA) [13], specifically in the comparative analysis of the mt genes for the *Hominidae* family [23], and for mammals in general. We briefly show here the results obtained in computing inter-genes and inter-species distances by using the *N* signals for the 37 mitochondrial DNA (mtDNA) genes. Seven members of the *Hominidae* family (*Hs* - *Homo sapiens*, *Hsn* - *Homo sapiens neanderthalensis*, *Papa* - *Pan paniscus*, *Pat* - *Pan troglodytes*, *Popya* - *Pongo pygmaeus abelii*, *Popy* - *Pongo pygmaeus*, *Gg* - *Gorilla gorilla*), and four other mammals (*Ms* - *Macaca sylvanus, Mm* - *Mus musculus*, *Bm* - *Balaenoptera musculus*, *Oa* - *Ornithorhynchus anatinus*) are considered. As previously discussed, only the digital derivatives of the offsets of the signals are necessary to establish the distances between a set of genes.

The procedure is illustrated in Fig. 8 which shows the digital derivatives of the offsets of the *N* signals in with respect to *Homo sapiens* signal for the case of the tRNA-Met gene. The indices of the lines are marked on the left, and the distances between each gene and the *Hs* gene are shown on the right. As expected, all the *Hominidae* (including the extant *Hsn*), as well as *Ms* – a forest monkey, are significantly closer to *Hs* than the other examples of mammals, which include not only the house mouse (9 – Mm), but also the blue whale (10 – *Bm*) and the platypus, the only mammal that lays eggs (11 – *Oa*).

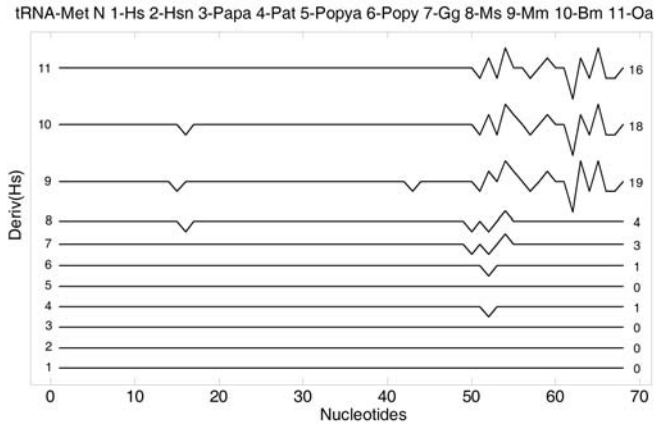

Figure 8. Derivatives of the offsets of the tRNA-Met gene *N* signals with respect to the *Hs* signal. The distances between each gene and the *Hs* gene are shown at the right of the lines.

Table 2 and Fig. 9 give the distances between the seven *Hominidae* species considered above, evaluated on the basis of the distances among their genes in the mtDNA respiratory chain. The succession of the genes along the horizontal axis corresponds to their position on the mtDNA nucleotide sequence. It is to notice the rhythmicity of the variation of the mt gene distances along the mtDNA molecule, which can not be fully attributed to the differences in the gene lengths. This result indicates the existence of "hotspots" in the mtDNA molecule from the variability point of view. Further research is needed to understand the mechanism and function of this surprising property of mtDNA.

TABLE 2
DISTANCES BETWEEN GENES IN THE *RESPIRATORY CHAIN* FOR THE *HOMINIDAE* FAMILY (*Hs* REFERENCE)

Species	Accession Genes	Complex I							CIII	Complex IV			
		mtND1	mtND2	mtND3	mtND4	mtND4L	mtND5	mtND6	mtCYB	mtCOX1	mtCOX2	mtCOX3	
		957 bp	1044 bp	346 bp	1378 bp	297 bp	1812 bp	525 bp	1141 bp	1542 bp	884 bp	681 bp	
Hs	NC001807	0	0	0	0	0	0	0	0	0	0	0	**0.00**
Hsn	NC011137	14	10	7	15	5	21	5	18	17	7	9	**42.47**
Pat	NC001643	91	104	43	139	24	215	54	134	141	67	79	**372.22**
Papa	NC001644	99	101	41	123	32	199	54	203	151	63	77	**392.76**
Gg	NC001645	103	154	44	169	28	287	52	152	177	91	96	**471.81**
Popya	NC002083	163	216	74	231	49	300	64	127	239	109	135	**576.12**
Popy	NC001646	170	220	81	241	54	373	71	197	257	104	144	**652.26**

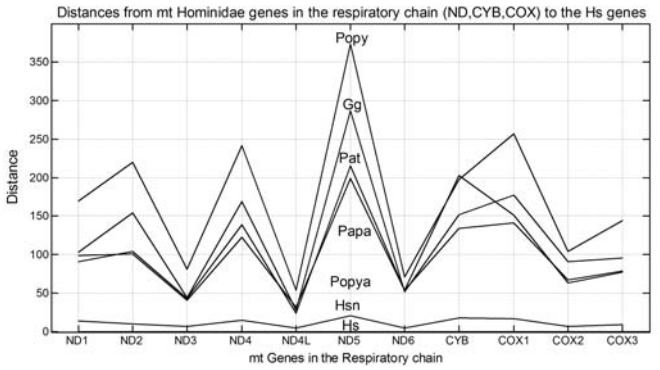

Figure 9. Distances between *Hominidae* family Respiratory chain mtDNA genes and the corresponding Hs mt genes.

VI. CONCLUSIONS

The paper applies the NuGS methodology for a molecular level approach to building phylogenetic trees based on the distances between selected sets of genes. The results are consistent with previous studies based on fossil findings, average genetic differences, or on identifying features uniquely shared between pairs of species.

An improved method using a software assisted sequence alignment is being developed.

REFERENCES

[1] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing Magazine*; vol.18, pp.8-20, 2001.

[2] 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.

[3] P.D. Cristea, "Genetic signal representation and analysis," *Proc. of Society of Photo-Optical Instrumentation Engineers* (*SPIE*), 4623, 77-84, 2002.

[4] P.D. Cristea, "Representation and Analysis of DNA sequences," in E.G. Dougherty, I. Shmulevici, J. Chen, Z. J. Wang, (ed.), *Genomic Signal Processing and Statistics*, 15-65, Book Series on Signal Proc. and Comm., Hindawi, New York, NY 10022, USA, 2005

[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.1, pp.132–137, 2004.

[7] Natl. Center for Biotechnology Information, Natl. Institutes of Health, Natl. Lib. Medicine, GenBank, *http://www.ncbi.nlm.nih.gov/genoms.*

[8] 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.

[9] P.D. Cristea, D. Otelea, R. Tuduce, "Genomic Signal analysis of HIV variability," *Proc. of SPIE BIOS 2005*, vol. 6, no. 14, pp. 362-372 2005.

[10] P.D. Cristea, D. Otelea, R. Tuduce, "Study of HIV Variability Based on Genomic Signal Analysis of Protease and Reverse Transcriptase Genes," *EMBC'05*, Shanghai, China, paper 1845, IEEE Cat: 05CH37611, 2005.

[11] P.D. Cristea, "Genomic Signal Analysis of Mycobacterium tuberculosis," *Progress in Biomedical Optics and Imaging, Proceedings of SPIE*, vol. 6447, pp. C1– C8, 2007.

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

[13] P. D. Cristea, R. Tuduce, "Hominidae mtDNA Analysis by Using Nucleotide Genomic Signals," *GSP 2011 – 2nd Intl Workshop on Genomic Signals Processing*, Bucharest, pp.61-65, 2011.

[14] T. Hall, BioEdit, Nov. 4, 2011, http://www.mbio.ncsu.edu/bioedit

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

[16] Akhtar, M., Epps, J., Ambikairajah, E., "On DNA numerical representations for period-3 based exon prediction," *Proc. of IEEE Workshop on Genomic Signal Processing and Statistics*, GENSIPS, pp. 1-4, 2007.

[17] J.A. Berger, S.K. Mitra, J. Astola, "Spectrum analysis for DNA sequences," *Proc. of the 7th Intl. Symp. on Signal Processing and its Applications*, vol. 2, pp.29-32, 2003.

[18] N. Chakravarthy, A. Spanias, L.D. Lasemidis, K. Tsakalis, "Autoregressive modeling and feature analysis of DNA sequences," *EURASIP J. of Genomic Signal Processing*, vol.1, pp.13-28, 2004.

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

[20] 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, 1997.

[21] P.D. Cristea, R. Tuduce, J. Cornelis, R. Deklerck, I. Nastac, M. Andrei, "Signal Representation and Processing of Nucleotide Sequences," *Proc. of the 7th IEEE Intl. Conf. on Bioinformatics and Bioengineering*, pp.1214-1219, Boston, USA, 2007.

[22] C. Apetrei *et al.*, "HIV Type 1 Diversity in Northeastern Romania in 2000-2001 Based on Phylogenetic Analysis of pol Sequences from Patients Failing Antiretroviral Therapy," *AIDS Research and Human Retroviruses*, vol. 19, no. 12, pp. 1155-1161, December 2003.

[23] P.D. Cristea, R. Tuduce, "Nucleotide Genomic Signal Comparative Analysis of Homo Sapiens and other Hominidae mtDNA," ISSCS 2009 – *International Symposium on Signals, Circuits and Systems*, Iaşi, Romania, pp. 5-8, 2009.