Codifying Primary Protein Structure as Peptides Frequencies Vector

An Efficient Alternative Method to Investigate Relationships among Genes and Organisms

Braulio R. G. M. Couto[1], Bruna A. Coimbra[1], Gabriel B. Tofani[2], Gustavo P. Irffi[2] and Cinthia T. V. Rocha[1]
[1]Instituto de Engenharia e Tecnologia- IET; [2]Instituto de Ciências Biológicas e da Saúde - ICBS
Centro Universitário de Belo Horizonte - UniBH
Belo Horizonte, MG, Brazil
braulio.couto@unibh.br, brunatecquimica@hotmail.com, bandeiragt@yahoo.com, palmer.gustavo@gmail.com, cinthia.rocha@prof.unibh.br

Marcos Augusto dos Santos
Departamento de Ciência da Computação - DCC
Universidade Federal de Minas Gerais - UFMG
Belo Horizonte, MG, Brazil
marcos@dcc.ufmg.br

***Abstract*—There are no uniquely correct methods for inferring phylogenies. In this scenario, Linear Algebra methods and visualization techniques are essential to better analyze complex systems and can be very helpful to categorize species. Thus, if proteins are represented as vectors, it will be possible to apply Linear Algebra methods in order to investigate relationships among genes and organisms. To investigate whether or not primary protein sequences and genomes can be represented as vectors and processed by Linear Algebra methods to generate accurate phylogenetic relationships, four sets of sequences were analyzed by classical phylogenetics techniques, based on pairwise multiple alignments, and by Linear Algebra and optimization methods. The results showed that primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space the relationships among species are consistent with classic phylogenetic trees.**

Keywords-protein sequences; phylogenetics; linear algebra methods.

I. INTRODUCTION

Phylogenetics or cladistics aims to reconstruct the evolutionary relationship among genes and organisms and to establish classifications that reflect those genealogies. Its fundamental axiom is that, as a product of evolution, nature is hierarchically ordered [1]. Unfortunately, there are no uniquely correct methods for inferring phylogenies, and many methods have been used [2]. Given an alignment and a tree scoring function, there are no efficient methodologies to calculate an optimal phylogeny. In this scenario, Linear Algebra methods and visualization techniques are essential to better analyze complex systems and can be very helpful to categorize species. So, if proteins are represented as vectors, it will be possible to apply Linear Algebra methods to investigate relationships among genes and organisms.

Here, we investigated whether or not primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space it generates relationships among species consistent with classic phylogenetic trees. The objective of our study is to answer three questions: Is it possible to represent proteins and genomes as tripeptide frequency vectors? Are phylogenetic trees constructed by using Euclidean distance between protein vectors consistent with phylogenetic trees constructed with alignments (classical phylogenetic trees)? Do images of genomes represented by multidimensional vectors and visualized in reduced tridimensional space generate relationships among species consistent with those described by classical phylogenetic trees?

In Section II we present the material and methods to codify proteins and genomes as tripeptide frequency vectors, and how to build phylogenetic trees by using Euclidean distance between protein vectors. In the same section, we show the visualization technique of genome vectors in reduced space. In Section III, we present and discuss the results. Section IV concludes the paper.

II. MATERIAL AND METHODS

Four sets of sequences were analyzed by classical phylogenetics techniques, based on pairwise alignments, and by Linear Algebra and optimization methods. Firstly, the origin of the Human Immunodeficiency Virus (HIV) was analyzed, retrieving from GenBank the three longest coding regions from seventeen different isolated strains of the Human and Simian immunodeficiency virus (SIV): the gag protein, the pol polyprotein and the envelope polyprotein precursor [3][4][5]. The second database was composed by the complete genome of five strains of *Chlamydophila pneumoniae* that were retrieved from the NCBI (National Center for Biotechnology Information) website [14].

 ISBN: 978-1-61208-335-3

The third dataset was similar to that used by reference [6] being composed by 59 whole mitochondrial genomes from the NCBI genome database, each one with 13 genes, totaling 767 proteins. Mitochondrial DNA (mtDNA) is one of the most fundamental evolutionary markers for phylogenetics due to the absence of recombination, high mutation rate and ease of sequencing. The following genes were analyzed: ATP Synthase F0 subunit 6 (ATP6), ATP Synthase F0 subunit 8 (ATP8), Cytochrome C Oxidase subunit 1 (COX1), Cytochrome C Oxidase subunit 2 (COX2), Cytochrome C Oxidase subunit 3 (COX3), Cytochrome B (CYTB), NADH Dehydrogenase subunit 1 (ND1), NADH Dehydrogenase subunit 2 (ND2), NADH Dehydrogenase subunit 3 (ND3), NADH Dehydrogenase subunit 4 (ND4), NADH Dehydrogenase subunit 4L (ND4L), NADH Dehydrogenase subunit 5 (ND5), and NADH Dehydrogenase subunit 6 (ND6). The 59 species were: Goldfish, Tasseled-mouth loach, Commom Carp, Zebrafish, Atlantic Cod, Olive flounder, Ornate Bichir, inbow trout, Atlantic salmon, Salvelinus, Brook trout, Redhead (Bird), Oriental Stork, White Stork, Peredrin Falcon, Red Junglefowl, Rook (Bird), Grey-headed Broadbill, Village Indigobird, Rhea, Ostrich, Cattle (Cows), Hippopotamus, Sheep, Wild boar, Dog, Cat, Grey Seal, Harbor seal, Blue Whale, Fin Whale, Jamaican Fruit Bat, Nine-Banded Armadillo, Virginia Opossum, Wallaroo, European Headhog, European rabbit, Platypus, White Rhinoceros, Asinus, Horse, Indian rhinoceros, Western Gorilla, Human, Orangutan, Chimp Vellerosus, African Bush Elephant, Guinea Pig, House Mouse, Edible Dormouse, Brown Rat, European mole, Aardvark, American Alligator, Ryukyu odd-tooth snake, Mole Skink, Green Sea Turtle, Painted Turtle, and African helmeted turtle.

The last database analyzed was composed by mitochondrial D-loop sequences for the Hominidae taxa (pongidae) [7]-[10]. We used this dataset because mitochondrial D-loop is very useful for comparing closely related organisms provided that it is one of the fastest mutating sequence regions in animal DNA. The origin of modern man is a highly debated issue that has recently been tackled by using mtDNA sequences. The limited genetic variability of human mtDNA has been explained in terms of a recent common genetic ancestry: all modern-population mtDNAs originated from a single woman who lived in Africa less than 200,000 years ago [7][8][9]. This family embraces the gorillas, chimpanzees, orangutans and the humans. Species description (and GenBank access code): German Neanderthal (AF011222), Russian Neanderthal (AF254446), European Human (X90314), Puti Orangutan (AF451972), Jari Orangutan (AF451964), Mountain Gorilla Rwanda (AF089820), Western Lowland Gorilla (AY079510), Eastern Lowland Gorilla (AF050738), Chimp Schweinfurthii (AF176722), Chimp Vellerosus (AF315498), Chimp Verus (AF1767310), and Chimp Troglodytes (AF176766). Both HIV and primates analysis (the first and fourth datasets) were based on demos of the bioinformatics toolbox from MATLAB (The MathWorks, Inc).

In order to visualize the genomes, we must represent each one as a point in space. The distance between the points should represent the differences in the genomes as a whole. Therefore, we might expect similar species to be close together in space. The genome proteins were represented as vectors of frequencies of groups of amino-acids. In this study, a sliding window of size 3 was used to measure the frequency (tripeptide frequency vectors). To represent the genome we used the vector sum of all its proteins. Therefore, we can obtain a database of genomes, S, as a rectangular matrix, X, where each line corresponds to one of the n genomes:

$$X = (X_1, X_2, ..., X_n)^T = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{pmatrix}$$

With 20 amino-acids, each primary protein structure is codified as a vector with $m = 20^3 = 8{,}000$ dimensions. In the above matrix, X_{ij} represents frequency o tripeptide i in the genome j (i = 1, 2, …n; j = 1, 2, …m; where n = number of genomes and m = 8,000 possible tripeptides). To codify a complete genome, the vector representations of genes from individual species are summed [6][11][12]. Therefore, both genes and genomes are represented as tripeptide frequency vectors with 8,000 dimensions. To generate a suitable visualization of these vectors, it is necessary to reduce the dimensionality of the space, with the minimum loss of information. Reference [13] presented a visualization technique to analyze chemical databases. We used that technique as a basis to develop a method for using genomes to visualize relationships among species in reduced space (3D). Actually, the high-dimensional visualization problem in $\Re^m$ can be formulated as a distance-geometry problem: to find n points in low space (2D or 3D) so that their interpoint distances match the corresponding values from $\Re^m$ as closely as possible. When a representation in reduced space, Y, is generated for the database matrix X, we can calculate an error function **E** as the following:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_{ij} - \gamma_{ij})^2$$

where δ_{ij} is the Euclidean distance between protein vector (or genome) i and j in the original space, represented in the matrix X, and γ_{ij} is the Euclidean distance between protein vector (or genome) i and j in the reduced space, represented in the matrix Y. The best representation of X in the reduced space will be the Y with the minimal associated error function. Therefore, we must solve an unconstrained optimization problem. Many methods can be used to solve this problem [13]. In this study, we used a technique based on the interior-reflective Newton method, actually the conjugated gradient Newton's method [11]. Classical phylogenetic tree reconstructions were made by using

 ISBN: 978-1-61208-335-3

pairwise alignments algorithms. Alternative phylogenetic trees were obtained by UPGMA or Neighbor-Joining methods using, as pairwise distances, the Euclidean distance between each protein or genome vector of the dataset.

III. RESULTS

Phylogenetic tree reconstruction of 17 strains of the Human (HIV) and Simian immunodeficiency virus (SIV) in the first dataset were made for each coding region. For the GAG coding region we used the Tajima-Nei method to measure the distance between the sequences and the unweighted pair group method using arithmetic averages (UPGMA) for the hierarchical clustering. Phylogenetic tree for the POL polyproteins was made by using the Jukes-Cantor method to measure distance between sequences and the weighted pair group method using arithmetic averages (WPGMA) for the hierarchical clustering. For the ENV polyproteins we used the normalized pairwise alignment scores as distances between sequences and the UPGMA method. Given that the three trees showed slightly different results, a consensus tree using all three regions was built using a weighted average of the three trees (Figure 1). All three sequences from the GAG, POL, and ENV regions of the 17 HIV and SIV strains were codified as tripeptide frequency vectors that can be visualized in tridimensional space (Figure 2). Phylogenetic tree of HIV and SIV viruses reconstructed by Neighbor-joining method using as pairwise distances the Euclidean distance between each tripeptide frequency vectors (Figure 3). Both phylogenetic trees (Figure 1 and 3) and the vectors in 3D space (Figure 2) illustrate the presence of two clusters and some other isolated strains. The most compact cluster includes all the HIV2 samples, and the second cluster contains the HIV1 strain (not as compact as the HIV2 cluster). From the trees it appears that the Chimpanzee is the source of HIV1.

Nucleotide sequences of five isolated strains of *Chlamydophila pneumoniae* were converted into amino-acid sequences, codified as tripeptide frequency vectors and projected in reduced space (3D): strains J138 and TW-182 occupy the same region of space, near CWL029 strain and far from AR39 and LPCoLN. Before applying the method proposed herein, we tried to generate multiples alignments using the complete genomes of *Chlamydophila pneumonia* for reconstructing the classical dendogram (Figure 4) in MATLAB. An “out of memory error” occurred in a computer with an Intel(R) Core(TM) i5 processor and 6GB RAM memory. It was impossible to reconstruct the classical phylogenetic tree in that computer because the genome was more than 1,200,000 nucleotides long and there were more than 400,000 amino-acids. The computer time and memory consumption grew quickly due to the size of each genome, generating computational fatal error. This was not a problem for the Linear Algebra method proposed here: all five genomes were codified as tripeptide frequency vectors and projected in 3D space using the same computer. Phylogenetic tree reconstructed with the Euclidean distance between each tripeptide vectors (Figure 5) was equal to the classical dendogram in Figure 4. For the third dataset, amino-acid sequences of thirteen mitochondrial genes from 59 different species were codified as tripeptide frequency vectors and projected in reduced space (3D): we observed exactly 13 compact clusters related to each gene analyzed (Figure 6). When these genes vectors were summed and the resultant genome vectors projected in tridimensional space, the four classes analyzed formed compact and distinct clusters (Figure 7). Species from the same class (Actinopterygii, Aves, Mammalia, and Reptilia) were placed in a monophyletic branch in the phylogenetic tree of the 59 whole mitochondrial genomes reconstructed by Neighbor-joining method using as pairwise distances between species the Euclidean distance between each tripeptide frequency vectors. When the phylogenetic tree for Mammalia species was reconstructed by using Euclidean distance between vector species, we observed a pattern compatible with classical trees (Figure 8). The last database evaluated was composed by mitochondrial D-loop sequences for the Hominidae taxa. Pairwise distances using the Jukes-Cantor formula and the phylogenetic tree with the UPGMA distance method were reconstructed (Figure 9). In this classical phylogenetic tree, Human resembles Neanderthal and Chimp species. However, when amino-acid from these mitochondrial D-loop sequences for the Pongidae taxa were codified as tripeptide frequency vectors and projected in reduced space (3D), we observed that European Human was more closely related to Gorilla than to Chimp species (Figure 10). In the phylogenetic tree built using Euclidean distance between vector species (Figure 11) Neanderthal, Chimp and Orangutan form monophyletic groups. However, differently from the classical tree, Human and Gorilla species were in the same branch.

 ISBN: 978-1-61208-335-3

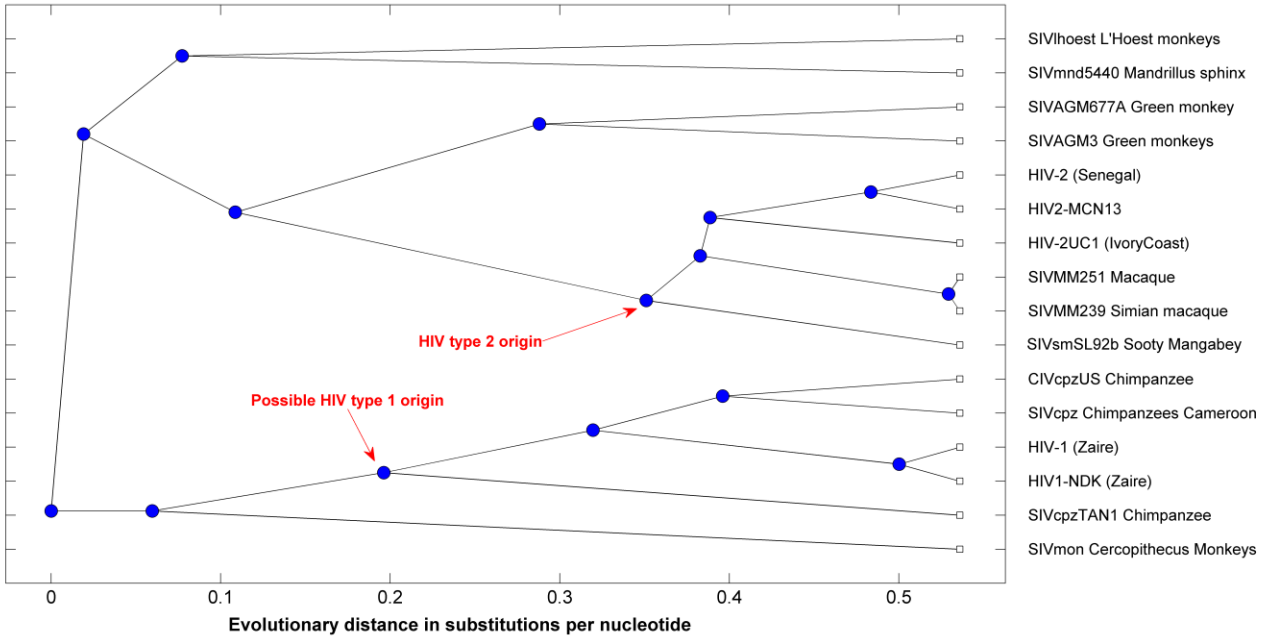

Figure 1. Phylogenetic consensus tree from multiple strains of the HIV and SIV viruses: the tree illustrates the presence of two clusters and some other isolated strains, showing possible origins for two characterized strains of human AIDS viruses, type 1 (HIV-1) and type 2 (HIV-2).

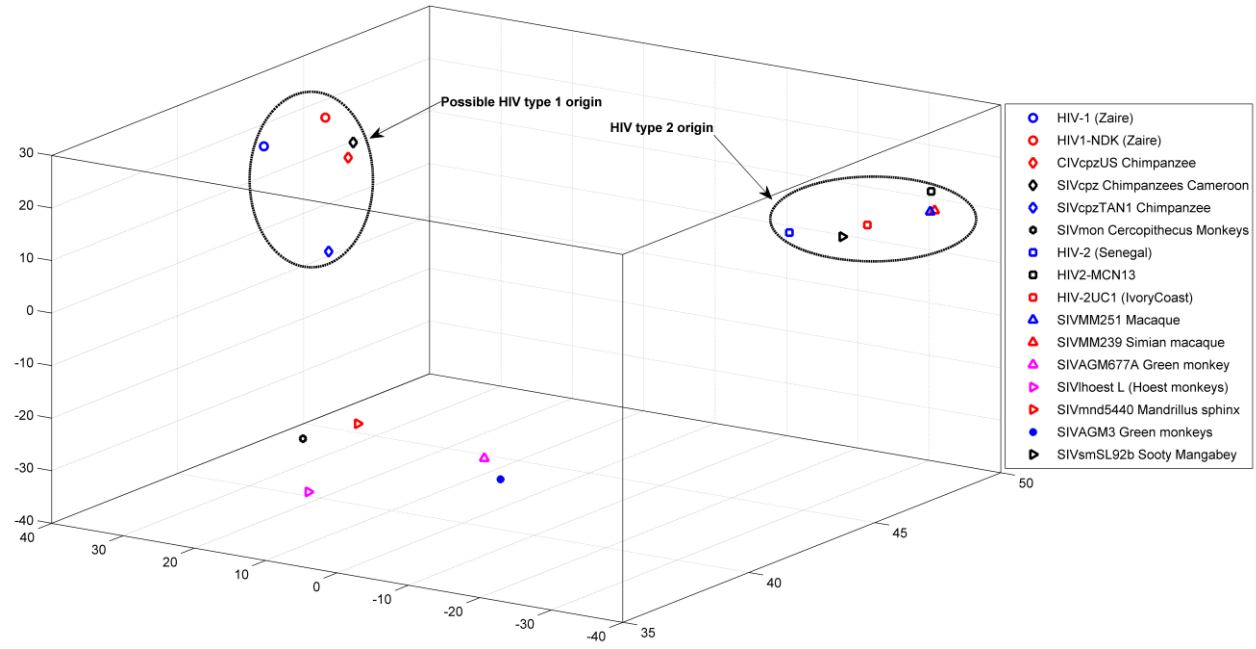

Figure 2. Nucleotide sequences from the GAG, POL, and ENV coding regions of seventeen different isolated strains of the Human and Simian immunodeficiency viruses were converted into amino-acid sequences, concatenated, codified as tripeptide frequency vectors and projected in reduced space (3D): the two clusters containing HIV1 and HIV2 are identified, as in the phylogenetic tree (Figure 1).

 ISBN: 978-1-61208-335-3

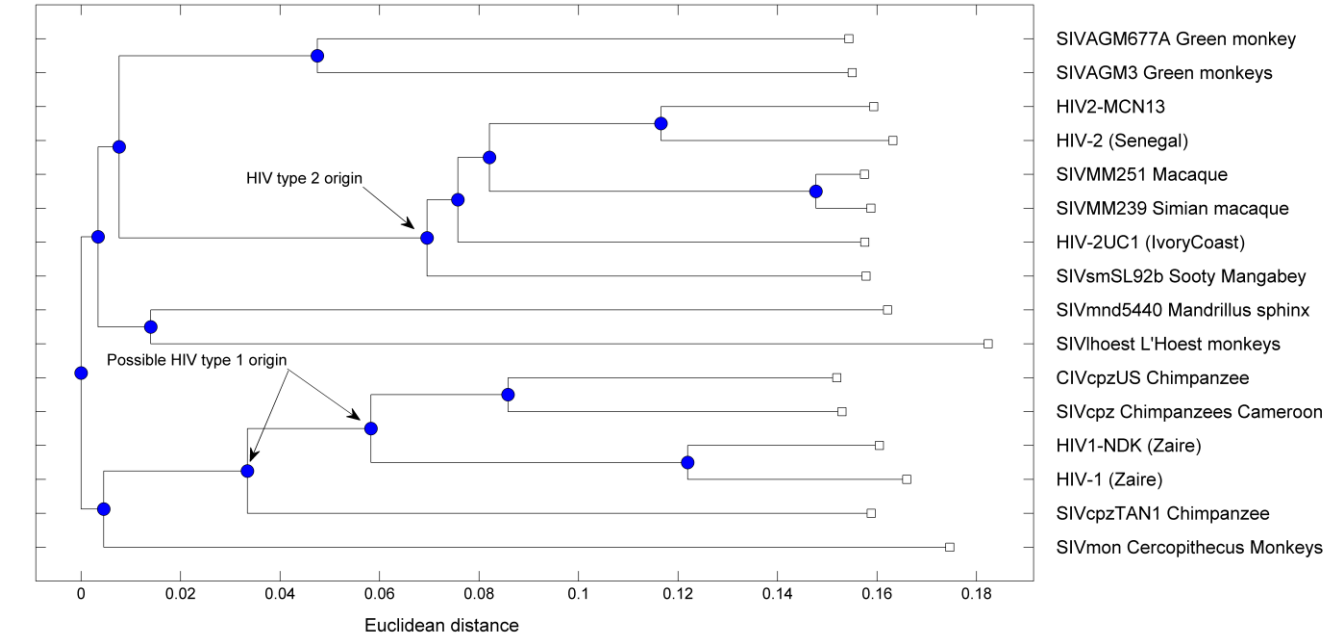

Figure 3. Phylogenetic tree of HIV and SIV viruses reconstructed by Neighbor-joining method using the Euclidean distance between each tripeptide frequency vectors: the pattern obtained is identical as that showed in classical phylogenetic tree from Figure

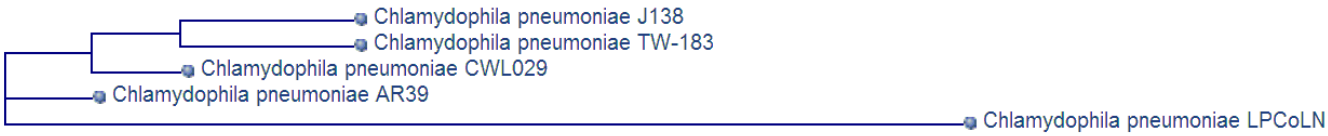

Figure 4. Classical dendogram, constructed based on genomic BLAST, for five strains of *Chlamydophila pneumoniae* (available at reference [14]).

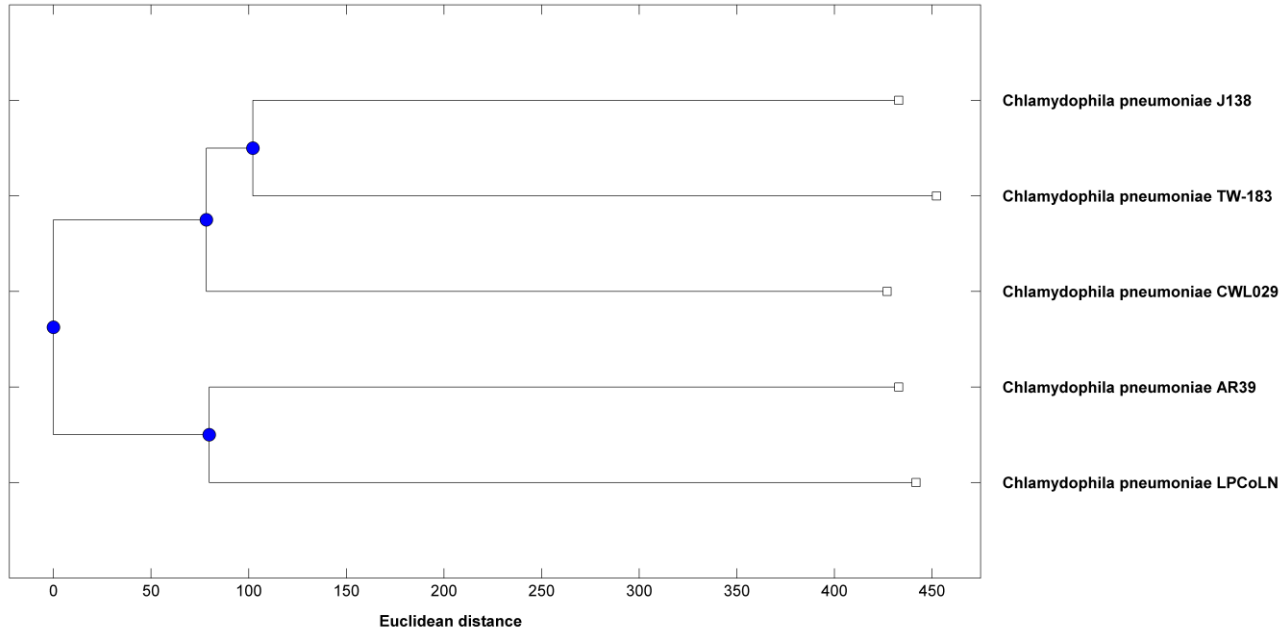

Figure 5. Phylogenetic tree of *Chlamydophila pneumoniae* strains reconstructed by Neighbor-joining method using as pairwise distances between strains the Euclidean distance between each tripeptide frequency vectors: there in no difference between this tree and the classical dendogram from Figure 4.

 ISBN: 978-1-61208-335-3

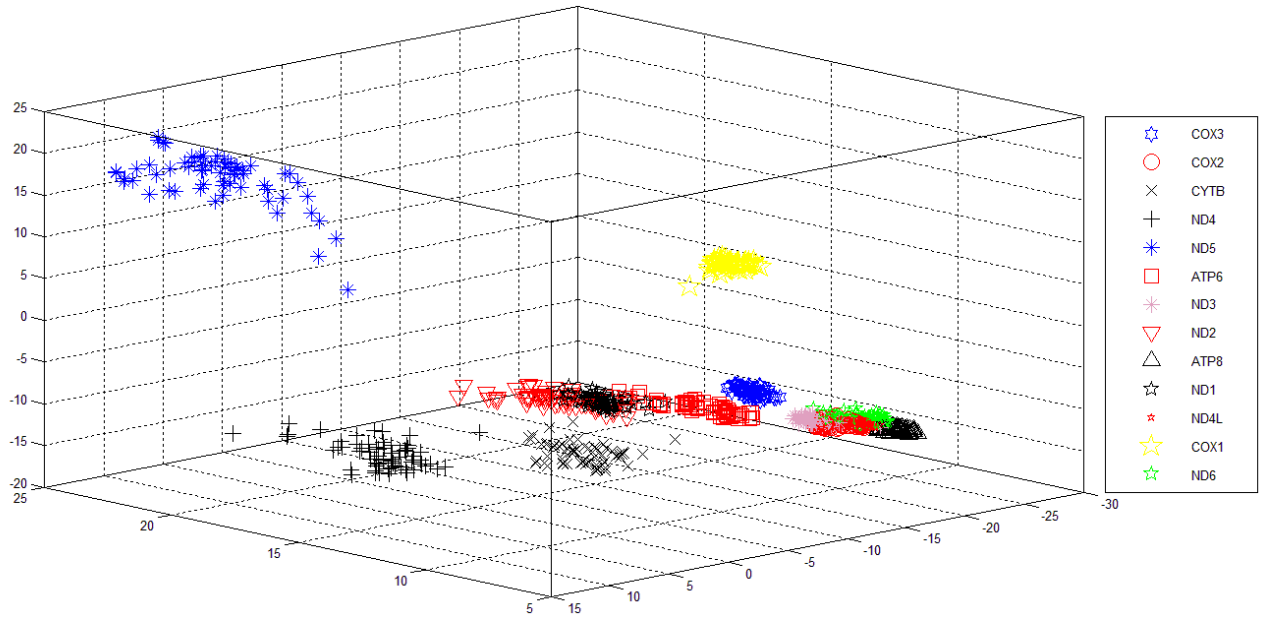

Figure 6. Visualization in reduced space (3D) of thirteen mitochondrial genes from 59 different species, codified as tripeptide frequency vectors: there are exactly thirteen compact and well defined clusters of genes.

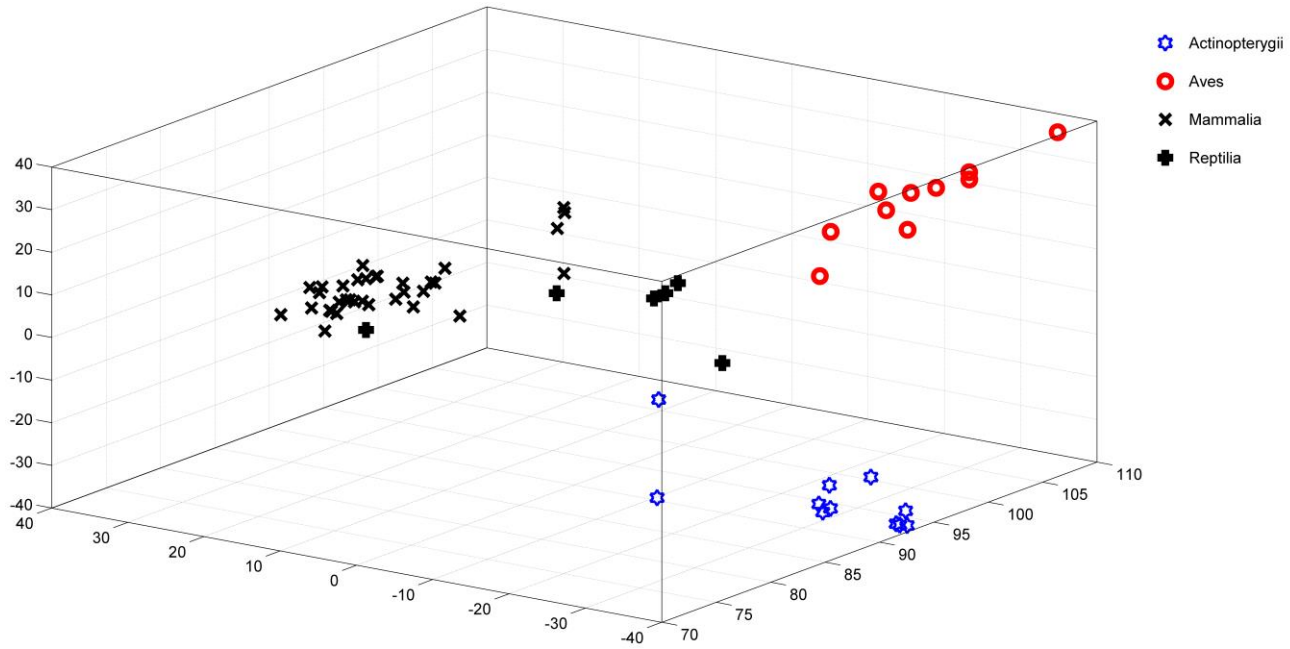

Figure 7. Amino-acid sequences of thirteen mitochondrial genes from 59 different species were codified as tripeptide frequency vectors, summed to represent each specie vector and projected in reduced space (3D): there are exactly four compact clusters related to the four class analyzed (Actinopterygii, Aves, Mammalia, and Reptilia).

 ISBN: 978-1-61208-335-3

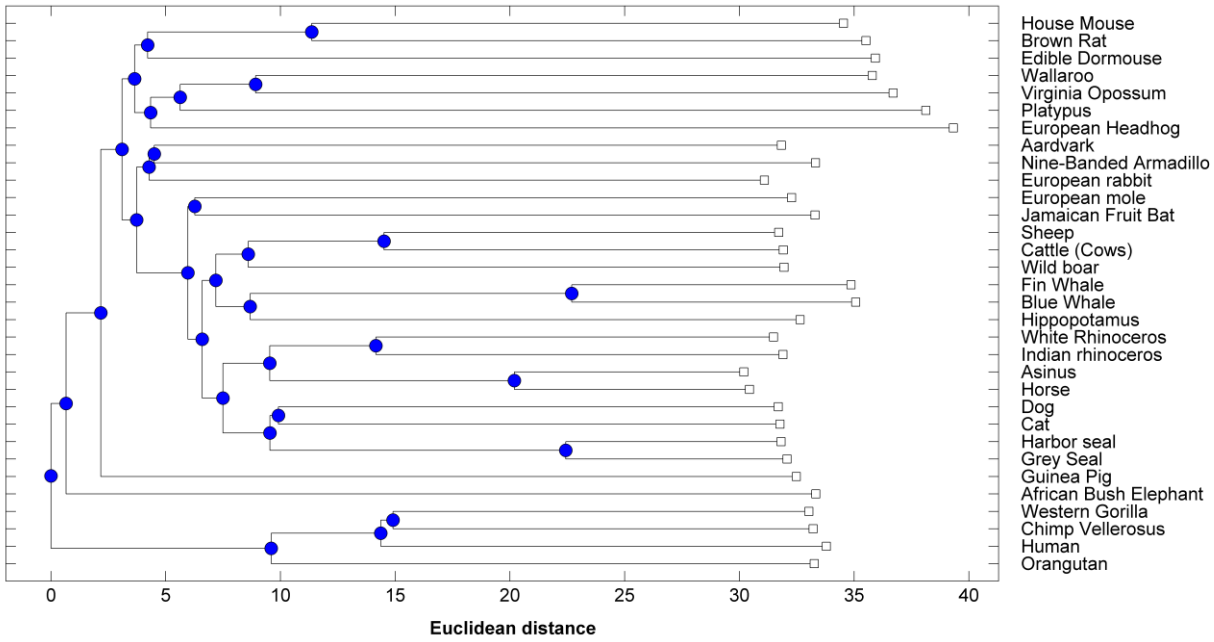

Figure 8. Phylogenetic tree of whole mitochondrial genomes from 32 Mammalia species reconstructed by Neighbor-joining method using Euclidean distance between each sum vector.

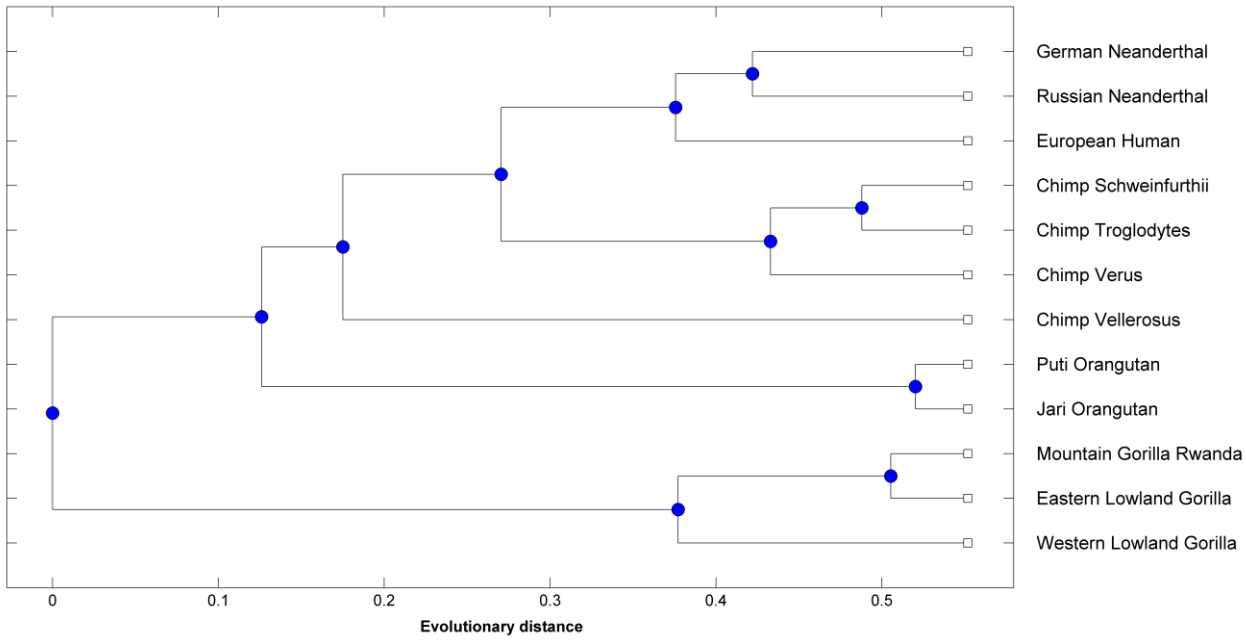

Figure 9. Classical phylogenetic tree from mitochondrial D-loop sequences for the Hominidae taxa reconstructed by pairwise distances using the Jukes-Cantor formula and the UPGMA distance method.

 ISBN: 978-1-61208-335-3

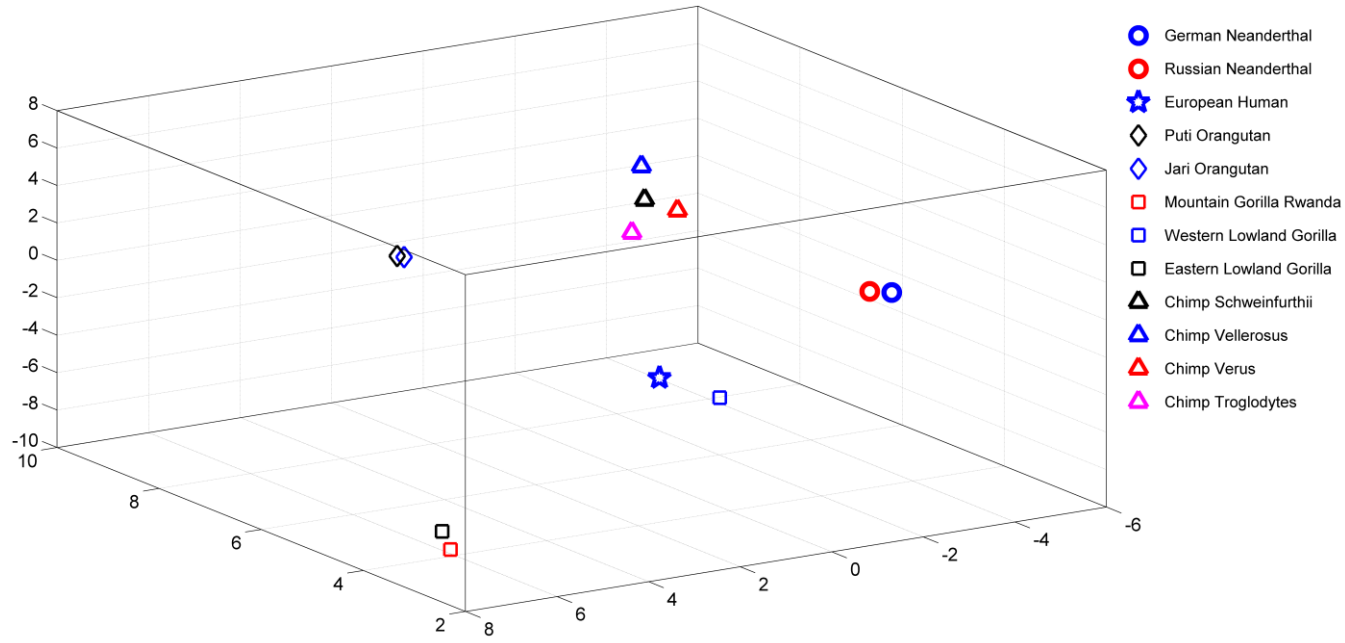

Figure 10. Mitochondrial D-loop sequences for the Hominidae taxa were codified as tripeptide frequency vectors and projected in tridimentional reduced space: groups of Neanderthal, Orangutan, Gorilla, and Chimp species occupy specific regions in the space. European Human is closer to Gorilla than Chimp.

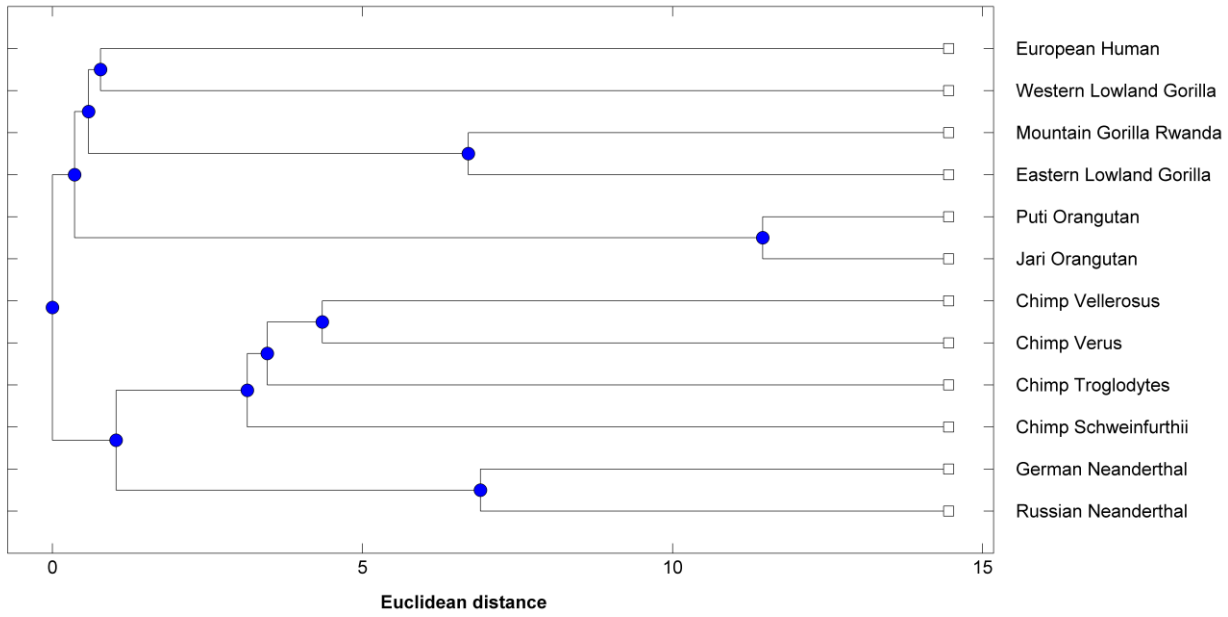

Figure 11. Phylogenetic tree from mitochondrial D-loop sequences for the Hominidae taxa reconstructed by using the Euclidean distance between specie vectors: Human and Gorilla are at the same branch.

 ISBN: 978-1-61208-335-3

IV. CONCLUSION

Primary protein structure from four different datasets were codified as peptides frequencies vector, visualized in reduced space (3D) and analyzed by using Euclidean distance as pairwise distances between genes and species. According to our results, the Linear Algebra method proposed seems to work very well, i.e., primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space they generate relationships among species consistent with classic phylogenetic trees. We illustrated by the previous results biological and computational benefits of such a methodology.

Therefore, it is possible to represent proteins and genomes as tripeptide frequency vectors and the analysis done with these vectors are consistent with classical analysis using multiple alignments. Besides, phylogenetic trees constructed by using Euclidean distance between vectors of proteins are compatible with trees constructed with alignments (classical phylogenetic trees). Images of genomes represented by multidimensional vectors and visualized in reduced tridimensional space generate relationships among species consistent with those described by classical phylogenetic trees.

Computationally, and mathematically the proposed method simplifies the study of the evolutionary chain of genes and genomes. It is a richer model to investigate relationships among sets of organisms than classical phylogenetics trees, because it takes into account not only pairwise distances but also the geometric position of each genome in a multidimensional space. As a result, computational load is substantially lowered and complete genome, as shown in *Chlamydophila pneumonia* analysis, can be easily done in a modest computer.

REFERENCES

[1] J. V. Crisci, L. Katinas, and P. Posadas, Historical Biogeography: an Introduction, Harvard University Press, Cambridge, 2003.

[2] M. Salemi and A. M. Vandamme, The Phylogenetic Handbook: a Practical Approach to DNA and Protein Phylogeny, 1st ed., Cambridge University Press, Cambridge, 2003.

[3] F. Gao, *et al*., "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes," Nature 397(6718), 1999, pp. 436-41.

[4] H. W. Kestler 3rd, *et al*. "Comparison of simian immunodeficiency virus isolates," Nature 331(6157), 1998, pp. 619-622.

[5] M. Alizon, S. Wain-Hobson, L. Montagnier, and P. Sonigo, "Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients," Cell 46(1), 1986, pp. 63-74.

[6] G. W. Stuart, K. Moffett, and J. J. Leader, "A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes," Molecular Biology and Evolution 19(4), 2002, pp. 554–562.

[7] I. V. Ovchinnikov, A. Götherström, G. P. Romanova, V. M. Kharitonov, K. Lidén, and W. Goodwin, "Molecular analysis of Neanderthal DNA from the northern Caucasus," Nature 404(6777), 2000, pp. 490-493.

[8] A. Sajantila, *et al*., "Genes and languages in Europe: an analysis of mitochondrial lineages," Genome Research 5(1), 1995, pp. 42-52.

[9] M. Krings, A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo, "Neandertal DNA sequences and the origin of modern humans," Cell 90(1), 1997, pp. 19-30.

[10] M. I. Jensen-Seaman, and K. K. Kidd, "Mitochondrial DNA variation and biogeography of eastern gorilas," Molecular Ecology 10(9), 2001, pp. 2241-2247

[11] L. S. Marcolino, B. R. G. M. Couto, and M. A. Santos, "Genome Visualization in Space," Advances in Soft Computing, 2010, pp. 225-232.

[12] A. Seetharam and G. W. Stuart, "Whole genome phylogenies for multiple Drosophila species," BMC Research Notes 2012, vol. 5, p. 670.

[13] D. Xie and T. Schlick, "Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization," In: C. A. Floudas, P.M. Pardalos (eds), "Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches," vol. 40, Kluwer Academic Publishers, Dororecht/Boston/London, 2000, pp. 267-286.

[14] National Center for Biotechnology Information [Online] Available from: <http://www.ncbi.nlm.nih.gov/genome/171> 2014.02.28

 ISBN: 978-1-61208-335-3