Linear Algebra Methods for Inferring Phylogenies Based on Peptides Frequencies Vectors: An Efficient Alternative Method to Investigate Relationships among Genes, Genomes and Organisms

Lara Maria Silva Miranda[1], Gabriel Bandeira Tofani[1], Gustavo Palmer Irffi[1], Lucas Felipe Silva[1], Matheus Allef Cruz[1], Thiago do Carmo Librelon Rocha[1], Bráulio Roberto Gonçalves Marinho Couto[1], Marcos Augusto dos Santos[2]

[1]*Centro Universitário de Belo Horizonte (UniBH),* [2]*Universidade Federal de Minas Gerais (UFMG)*

The objective of this paper is to answer four questions: Is it possible to represent proteins and genomes as tripeptide frequency vectors? Why Euclidean distance between protein vectors is better than the cosine as a metric to build phylogenetic trees? 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? Five 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 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. Wholegenome sequencing of MRSA isolates from 14 patients involved in a outbreak were the third database. The fourth dataset was composed by 59 whole mitochondrial genomes from the NCBI genome database, each one with 13 genes, totaling 767 proteins. The last database analyzed was composed by mitochondrial D-loop sequences for the Hominidae taxa (pongidae). 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. Computationally, and mathematically the proposed method simplifies the study of the evolutionary chain of genes and genomes. The computational load is substantially lowered and complete genomes can be easily analyzed in a very modest computer.