mtDNA Data Mining: A Global Analysis

Camilla Reginatto De Pierri[1], Bruno Thiago de Lima Nichio[1], Tetsu Sakamoto[2], José Miguel Ortega[2], Mauro Antônio Alves Castro[1], Roberto Tadeu Raittz[1]

Federal University of Paraná[1], Federal University of Minas Gerais[2]

Phylogenetic methods based on comparisons of whole genomes features are relevant to provide information of many species and genes, besides the understanding of the evolutionary relationships. The origin of mitochondria and the time of acquisition in the course of evolution are some of the targets in the current research have led to many points of view by evolutionary biologists. The lack of eukaryotic intermediaries complicates the analysis of the actual mitochondrial evolutionary history. Despite several studies conducted until now and the large amount of information already known about the last common mitochondrial ancestor, we still need more evidences about the time of the organelle acquisition. In order to clarify facts not explained yet, this research was conducted in two steps. We've used a data mining strategy to explore all aspects of mitochondrial DNA and then inferred a phylogenetic tree based on pairwise distance. For this, we used RefSeq data from 6,811 organisms, totalizing 98,933 mitochondrial proteins. We clustered all these mitochondrial proteins available on RefSeq database in MatLab environment, using RAFTS3GROUPS algorithm. This algorithm is one implementation of RAFT3 tool, which is used to find orthologs sequences on RAFTS3 database, according to the self-score value. Here, we used a self-score 50% similarity. To infer a version of the global tree, we use all mitochondrial sequences of RefSeq database, according to neighbor-joining method. As a result, we obtained 8431 clusters, of which 2159 has two or more sequences. As expected, in the analysis we noticed that the largest clusters are the 13 most common proteins among the species, that is responsible for oxidative phosphorylation. The biggest cluster had the product COX1 protein, followed by CytB. In addition to these 13 proteins, other less common have also been identified, especially in the fungi kingdom, protist and plantae organisms. Some hypothetical proteins were also detected. The clustering relation with the inferred phylogenetic tree was consistent in most of the tree branches analyzed. It shows that the phylogenetic tree proposed can be used as a reference for future research.