Construction of alternative algorithm for development of phylogenetic trees using InterPro entries and frequency vectors of tri-peptide

Lucas Felipe Silva[1], Dalila Dominique Duarte Rocha[2], Lara Maria Silva Miranda[1], Matheus Allef Cruz[1], Thiago do Carmo Librelon Rocha[1], Bráulio Roberto Gonçalves Marinho Couto[1], Marcos Augusto dos Santos[3]

[1]Centro Universitário de Belo Horizonte (Unibh); [2]Universidade Federal de São João del-Rei; [3]Universidade Federal de Minas Gerais (UFMG)

In the study of the evolution of species, the use of phylogenetic trees to verify the relationship among them is essential. However, the evolutionary reconstruction organisms using traditional phylogenetic methods can be affected by errors, e.g., misalignments or by using a limited number of genes. In addition, alignment methods of complete sequences of genomes are impossible because they demand a huge computational effort. In this context, representation of proteins as vectors in multidimensional space opens up possibilities for the application of linear algebra methods to investigate such relationships. The InterPro database integrates predictive models or "signatures "of proteins, describing it as a tool in the study of evolutionary processes. This research looks for answers to the following questions: a) genomes analyzed by linear algebra methods, using proteins as vectors of the frequency of tripeptide and InterPro entries can generate valid phylogenetic relationships from a biological viewpoint? b) What is the computational performance of linear algebra techniques when used to generate phylogenetic trees, compared to classical methods? Two sets of data were used: complete genomes were analyzed from 14 species of plant models, and 317 complete Eukaryotic genomes, retrieved from the UniProt database (http://www.uniprot.org/proteomes/). Classical methodology (pair-to-pair alignments) and linear algebra technique with tri-peptide vectors and frequency vectors of InterPro entries were applied to both datasets. In classical analysis, the sequences of the genomes were tested in the programs: MEGA, ClustalW, Clustal Omega, MUSCLE, BioEdit, and CLC Sequence. Needleman-Wunsch global alignment algorithm was used to generate data necessary to build classical trees. Unfortunately, there is no valid results due to computer problems and none of the programs submitted supported the large amount of data (for example, in the plant models data, evaluated genomes had, on average, approximately 13 million amino acids). To use linear algebra methods, we developed a matrix with the presence (1) or absence (0) of InterPro entries for each complete genome analyzed. In another matrix representation, all protein sequences of each complete genome were transformed into vectors of tri-peptides frequencies. By using Linear Algebra algorithm, it was possible to construct phylogenetic trees that showed similar results for both representation vectors, tri-peptides and InterPro entries. It has been found that the distribution of species in the dendrograms was generated according to the taxonomy presented in the literature. The results showed that genomes can be evaluated using linear algebra techniques.