Effects of Alignment Characteristics on Distance-based Tree Reconstruction

Roberto Tadeu Raittz, Jeroniza Nunes Marchaukoski, Dieval Guizelini, ALEXANDRE GORI DE CASTILHO, Guilherme Taborda Ribas

UNIVERSIDADE FEDERAL DO PARANÁ

Abstract

The analysis and comparison of nucleotide and amino acid sequences are the cornerstones of computational biology and bioinformatics studies. Trees are important tools for comparative studies between gene sequences, proteins and genomes. As well as in studies of phylogeny, protein homology and taxonomy based on molecular characteristics. The constructions of the trees are directly affected by the multiple alignment algorithms, by the measures of distances between the sequences and in the research ambiguity of the interpretation of the distance values, observed in the comparison of the sequences composition and the evolutionary events. In this work, we revisited classical hierarchical clustering methods, applied to nucleotide and amino acid sequences, to measure and identify the effects on the resulting dendrogram when one varies the size and number of aligned sequences. And, also, to verify whether any of those methods can reproduce the reference tree - or at least close for some alignment variation. The Average-linkage is widely studied and has known problems of consistency in phylogenetic reconstruction, especially when the distance matrix is not time corrected with evolutive models. However, with the nowadays softwares for simulating theoretical alignments and trees, it is possible drawing massive experiments to verify how the input data can affect the final tree obtained, and whether there are variations that can identify ranges where distance and linkage methods can be consistent. In the completed stages, 840 data sets were simulated and, from different clustering methods, 5, 880 trees were generated. The data sets were obtained by the Cartesian combination between size and sequence quantity. Where sizes ranged from 10 to 10, 000 bases, in nucleotide and amino acid sequences and; sequence amounts ranged from 25 to 100 sequences. We used the geodesic distance method to do trees comparisons. Among the methods evaluated for nucleotide sequences, the best methods are: single, complete, average and weighted. While for amino acid sequences, the highlighted methods are: median and centroid. We conclude that the larger the sequence size, the greater the consensus of the trees produced and, fairly often, the average-linkage method is not the most suitable for the reconstruction trees when compared to the other methods covered in this study. And, at this point in the research, none of these clustering methods reproduces the reference tree.

Funding: Capes

Link to Video: