Poster C-12

Do Genome-Escale Phylogenies Actually Detect True Trees?

Authors:
Karla Yotoko (*PUC-RS*)
Sandro Bonatto (*PUC-RS*)

Short Abstract: We tested, using different phylogenetic methods, and three different sets, if the use of complete genomes recovers a convergent "true" tree. Our data point to differences between DNA and aminoacid trees, as well to long branch attractions, suggesting that complete genomes do not assure the recover of the "true" tree.

Long Abstract:
In this work, we have tested the hypothesis that one phylogeny constructed with complete genomes represents the “true” tree, what would be evidenced by the fact that several different phylogenetic methods recover the same tree with high supported values. To do this, we attempted to reconstruct the phylogeny of three different orders of bacteria based on the complete set of orthologous genes found among each set of sequences. Our data sets consist in 13 OTUs of Lactobacillales (480 gene partitions), 11 OTUs of Bacillales (664 gene partitions) and 17 OTUs of Enterobacteriales (236 gene partitions). In our tests, we used the three most popular phylogenetic methods: Maximum Likelihood (ML), Maximum Parsimony (MP) and Neighbor-Joining (NJ), performed with both nucleotide and amino-acid sequences data. Except for ML analyses applied to the amino-acid sequences, which were performed using the program ‘tree-puzzle 5.0’, all trees were constructed using the program ‘PAUP* 4.0’. Since ML approaches require a choice of appropriated substitution models, we used the program ‘ModelTest 3.06’ to infer the best model for each DNA sequences set and the program ‘HYPHY’ to infer the best model for the protein data sets. Our results showed that all phylogenetic methods recovered the same tree for the order Lactobacillaes, whereas for Bacilalles, two alternative topologies were found, with high support values: one based on the DNA sequences and the other based on the amino-acid sequences. For the order Enterobacteriales, the scenario seems to be more complex: ML and MP methods applied to DNA sequences and MP method applied to protein sequences recovered one topology, whereas ML and NJ methods applied to protein sequences and NJ method used in DNA sequences recovered another one. We performed the Shimodaira-Hasegawa test and found significant differences between the two trees recovered for each bacterial order. One could imagine that including more OTUs and less gene partitions (as well as in Enterobacteriales) can collaborate to the inaccuracy of the phylogenetic inference. However, in Bacterialles, we included only 11 OTUs and 664 genes, and the incongruence also had appeared. Our next hypothesis is that the incongruence arose from different phylogenetic signals presented by different gene partitions in each data set. To test this hypothesis, we have built the ML tree of each gene partition (using its respective model of nucleotide substitution inferred by the 'ModelTest' program) and we have searched, within these trees, for the branches of the concatenated phylogenies. For the Lactobacillales data set, all branches presented in the concatenated tree were found in at least 30% of the partitions (160-478), with bootstrap average value of 83% (± 13.7). For the Bacilalles data set, the branches of the DNA topology

were found from 261 to 662 topologies, with bootstrap average value of 82,7% (± 17.3). The difference between the DNA and the protein topologies of Bacillales consists in only one branch. The branch presented exclusively in the concatenated protein topology was found in 153 partition topologies, with bootstrap average value of 60.4% (± 19.6), whereas the branch presented only in the concatenated DNA topology was found in 261 partitions with bootstrap average value of 67.6% (± 20.6). From these results, one can conclude that the differences could hardly be assigned to different phylogenetic signals among gene partitions. It is simpler to suppose that there are some putative differences among the protein and DNA data sets, which regards further investigations. For the Enterobacteriales data set, the branches of the topology 1 [ML, MP (DNA) and NJ (Protein)] were found in at least 69 topologies (69-236), with bootstrap average value of 84,6% (± 10.7). The topology 2 [ML, MP (protein) and NJ (DNA)] has one branch different from the topology 1, which was found in 95 topologies with a bootstrap average value of 64.4 (± 21.1), against a branch of the topology 1 found in 116 partitions, with bootstrap average value of 75% (± 18.0). In this case, we cannot discard the differences among the partitions in terms of phylogenetic signal. However, for both cases, we cannot discard the putative influence of the long branch attraction, which can affect the topologies due to several factors. Fortunately, new complete genomes were recently sequenced and published in the web, including those ones that can be useful to break the long branches included in the discrepancies found for Enterobacillales and Bacillales data sets. The next steps of this work is to include these new OTUs in order to test the influence of the long branches, and to construct the partition trees using the protein sequences and other methods in order to search for the differences between DNA and protein sequences as well as among the methods. As a preliminary conclusion, our data show that the use of complete genomes can contribute to find robust topologies, which can be used to reinforce several hypothesis concerning the phylogenetic history of a given set of organisms. Nevertheless, different topologies still arise from complete genomes, what clearly shows that the inclusion of several gene partitions, such as complete genomes, cannot assures the recover of the true tree, which also depends of the set of organisms included in each analysis.