Insights about the phylogenomic approaches to Staphylococcus aureus taxa clustering

Guilherme Coppini[1], Célio Dias Santos Júnior[1], Flávio Henrique Silva[1],

1 Federal University of São Carlos

Abstract

Staphylococcus aureus is a widespread bacteria involved in resistance-acquired infections. Rapid evolving rates of S. aureus make approaches as 16S rRNA phylogenetic trees less usual, being the most usual method to identify S. aureus strains Staphylococcal Protein A (SpA) phylogeny, which is based on a single protein-coding gene with a hyper-variable region X. But, could the phylogeny of a gene does reflect a complex network of strain phylogenetic relationships? Our main goal was to compare the clustering resolution and accuracy of whole-genome distance based approaches and protein-based phylogeny inside a S. aureus dataset, correlating taxa clustering to find a better strain phylogeny method. To do this, 168 S. aureus genomes available in NCBI were selected and had their proteins predicted by PRODIGAL software. SpA sequences were selected using BEAF pipeline with a manually curated database, and 5 partial or disrupted ORFs were discarded. SpA sequences were aligned with MAFFT, had their substitution model predicted by ModelGenerator and best model was selected through Aikaike Informative Criterion. LG+G+F+I substitution model was used in RAxML program to generate the consensus final tree, following the extended majority rule from a bootstrap with 100 pseudo replicates. Genomes were fragmented into 250 bp fragments by a home-made python script and were pairwised searched through Usearch local alignment, generating a total of 26,569 alignments. Total sum of alignments' bit-scores for each possible genome pairs were calculated and then used to calculate euclidean distances, generating a dissimilarity matrix, used to generate the final dendrogram by UPGMA method. Different topological inferences were summarized as Compare2trees's score of 0.4034, where the topologies were almost 60% divergent. When compared using ETE toolkit, the two constructions regarded a normalized Robinson-Foulds coefficient of 0.94 and a symmetric distance (RF) of 279.0. A frequency of edges in the SpA tree of 0.57 was also found in genomic distance dendrogram. The distance results provided by Phylo.io showed a poor topological conservation between both approaches. These results showed a different topological resolution between both approaches, reflecting a different strains sorting. The analysis of strains relation carried using phenotype data revealed an apparent better resolution for the whole-genome distance approach, where most strains were correctly clustered, although more tests are needed to confirm it. In this sense, we observed the importance of considering wide-genome approaches for taxa clustering, which despite not necessarily reflecting the phylogeny, could still be used to reflect the phenotype.

Funding: CNPq, CAPES