An Approach to Study Taxonomic Distribution of Genes: Biofilm Production as a Model

Antonio Gilson Gomes Mesquita[1], Sabrina Sondre de Oliveira Reis Margarido[1], José Miguel Ortega[2], Tetsu Sakamoto[3],

1 Universidade Federal do Acre, LabGenMol
2 Universidade Federal de Minas Gerais, Laboratório de Biodados
3 Universidade Federal de Minas Gerais. Laboratório de Biodados

Abstract

Taxonomic distribution of the orthologous of a gene of interest is not a trivial task to accomplish. Manual inspection of phylogenetic trees consist in the most analytical procedure to investigate the lowest common ancestor (LCA) and the descendant clades that have or not the gene, since deletions and lateral gene transfers may occur. Here we present the analysis of the taxonomic distribution of genes involved in production of bacterial biofilm. The query genes have been selected from the protein database UniProt and the taxonomic distribution was analyzed with software TaxOnTree. Originally, the software run a BLAST search limited by e-value (1e-5), similarity threshold (e.g. 30%) and builds phylogenetic trees using MUSCLE to obtain a multiple alignment, TrimAL to edit the alignment, FastTree to build the phylogenetic tree and colorizes the branches according to any taxonomic rank of choice, e.g. phylum, class, order, etc. turning the user into a taxonomy expert. However, since our interest was on bacterial genes, the limit of 200 proteins to generate the trees often did not surpass the analysis of a genus, due to the presence of great number of taxonomically closely related orthologues. Thus we further developed TaxOnTree to be able to present a restrict number of hits from a chosen clade, e.g. phylum. Development of the tool to provide biologically meaning outputs was developed in parallel to the study of biofilm genes. We could map the clade of origin of a specific gene (LCA) and the relationship of it to more distant ones. The genes of interest were: (i) phaZ, almost all clusters contained bacteria from the order Burkholderiales, two clusters containing Chromobacterium violaceum, one with sequences from phylum Ascomycota, one with Actinobacteria, along with a cluster of animal and protozoa sequences; (ii) phaE, restricted do Cyanobacteria of distinct classes; (iii) phaB restricted to Proteobacteria from the class Betaproteobateria; (iv) phaZ1, showing a cluster with Chromobacterium violaceum and Cyanobacteria associated with Proteobacteria from order Neisseriales and other clusters from Actinobacteria and Ascomycota, and one with animal and protozoa sequences; (v) phaZ2, with the same distribution of phaZ1; and (vi) phaC from Chromobacterium violaceum and present in Actinobacteria, Cyanobacteria and also Porifera and Mollusca.

Funding: Universidade Federal de Minas Gerais. Laboratório de Biodados.