Highly resolved phylogeny for Corynebacteriales

Nilson A. Da Rocha Coimbra[1,2], Vasco Azevedo[1], Aïda Ouangraoua[2]

[1] LGCM, Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [2] CoBIUS, Department of Computer Science, Universite de Sherbrooke, Sherbrooke, Quebec, Canada

The taxonomic classification of large numbers of organisms remains a hard task. In the past three decades, the 16S ribosomal RNA sequences were largely used as biological marker for phylogeny reconstruction of microbes. Nowadays, due the increase of genomic data and information produced and stored in databases, we are able to use all this whole information to reconstruct highly resolved phylogenies, by first detecting universal features in genomic data and then using them for taxonomic classification. In this work, our aim is to reconstruct the natural history and evolution of the Corynebacteriales family, in order to identify speciation in the CMNR group, the largest clade in the Actinobacteria domain. Genome data were retrieved from the RefSeq Database. The annotation in coding genes of each genome was extracted using in-house Python scripts. A customized version of Orthofinder, kindly provided by Dr. Emms, was used to cluster coding genes in families of orthologous genes. The content in coding genes of gene families was augmented using in-house Python scripts for protein prediction by homology. The phylogeny reconstruction was performed using GRIMM software and the PHYLIP software based on the augmented gene families. We collected 274 genome records of Corynebacteriales from NCBI RefSeq Database, with 1,013,515 proteins sequences in all genomes. Orthofinder predicted 27,165 clusters of homologous genes. Clusters containing paralogous genes were discarded. 22 universal, single-copy gene clusters were identified by Orthofinder. 55 additional universal clusters were obtained by augmenting existing clusters using protein prediction by homology. On total, 21,098 proteins partitioned into 77 universal clusters were predicted and used in order to reconstruct the phylogeny of Corynebacteriales.