Utilization of Machine-Learning Methodologies in Order to Understand Complex Evolutionary and Functional Links among Bacterial Genomes

IFCS 2013-2013 conference of the International Federation of Classification Societies, Tilburg University, The Netherlands, July 14-17, 2013.

O. Poirion[1] and B. Lafay[2]

1 Laboratoire AMPERE Ecole Centrale de Lyon, FRANCE
olivier.poirion@ec-lyon.fr
2 Laboratoire AMPERE Ecole Centrale de Lyon, FRANCE
benedicte.lafay@ec-lyon.fr

Abstract

We are searching for evolutionary trends among genome maintenance-related genes present on the replicon sets (i.e., chromosomes and plasmids) of bacterial genomes. Traditional bioinformatic and phylogenetic methods are not adapted to large scale and high-dimensional study. We thus developed a semi-supervised analytical pipeline relying on data-mining methodologies. Generic unsupervised (SOM, K-means, Bayesian networks) and supervised (SVM, decision trees, boosting) classification methods were combined with specific bioinformatic algorithms based on sequence homology search (BLAST). Through this approach, important evolutionary processes could be characterized among genome-integrated plasmids and chromosomes. We here report on the inherent difficulties (input data bias, high-dimensional analysis, noise) and the applied methodology, and conclude on the significance of the data-mining methodology in knowledge discovery.

Keywords
comparative genomics, homology search, gene ontologies, classification, analytical pipeline