Taxonomic identification of metagenomics reads based on sequence features by Support Vector Machines (SVM)

Tahila Andrighetti[1], Ney Lemke[1]

[1]*Universidade Estadual Paulista Júlio de Mesquita Filho, Departamento de Física e Biofísica, Botucatu, São Paulo, Brasil*

The acknowledgement of the importance of microbiota composition is increasing steadily after the advent of metagenomics. This approach allows sequencing and analyzing genetic material from a microbial community without the need of microbial culture. Since 99% of microorganisms are not culturable, metagenomics is the standard methodology to investigate microbiomes composition and dynamics. However, the actual output data of metagenome sequencing consists of a bunch of DNA fragments originated from various microorganisms. Moreover, the lack of reference genomes in databases challenges taxonomic identification of unknown organisms in these samples and unbiased estimates for the performance of the proposed methodologies. In this work, we evaluated the predictive power of Support Vector Machine (SVM) learning tool on taxonomic classification in phyla of unknown metagenomics DNA reads. To simulate the identification of unknown microorganisms, we used Gammaproteobacteria sequences excluding *Escherichia coli* as the training set in SVM. From the trained model, we classified the sequences of *E. coli* and analyzed if they were correctly assigned on Gammaproteobacteria group. The tests were performed for 100, 400 and 1000 bp test sequences to evaluate the influence of size on the prediction. The simulations were performed using the following DNA measurements as SVM input: GC content, di, tri and tetraplet entropy, di, tri and tetranucleotides frequencies (2, 3 and 4-mers), dinucleotide abundance and tetranucleotide derived z-score correlations (TETRA). We tested sets of measurements composed by all parameters but excluding one to compare the relative impact of each measure. We found that the groups which excluded TETRA are less suitable for the most of sizes tested, specially for 100 bp. The other groups showed AUC values higher than 0.7 for prediction of unknown sequences. The use of sequence features is an interesting approach to characterize sequences of not fully sequenced organisms.