Establishment of a pipeline for 16S-based metagenomic studies of Mycobacterium leprae

Felipe Borim Corrêa[1,2], Gabriel da Rocha Fernandes[2]

Universidade Federal de Minas Gerais, Programa Interunidades de Pós-Graduação em Bioinformática, Belo Horizonte, MG, Brazil[1], Centro de Pesquisas René Rachou, Fundação Oswaldo Cruz, Biosystems Informatics and Genomics Group, Belo Horizonte, MG, Brazil[2]

Identifying the pathogen Mycobacterium leprae is not a simple task and leprosy still remains a health problem all over the world. Sequence analysis of 16S ribosomal RNA has been used to perform metagenomic studies. However, 16S-based techniques are known to have limitations because of the biases mainly related to DNA extraction and PCR amplification. Whereas we can not simulate DNA extraction, the choice of primers is an important step since the hypervariable regions for distinguishment among taxa can vary, such as the amplification efficiency. The purpose of this study was to establish a pipeline for 16S-based metagenomic studies of Mycobacterium leprae. Methods were divided in two parts: candidates selection and candidates evaluation. In the first part we used Simulate_PCR to perform an in silico PCR with the 16S rRNA gene sequence of M. leprae TN strain (NC_002677.1). Amplicons were simulated using all viable pairings of 22 forward and 22 reverse primers and were filtered by maximum length of 550 nucleotides. Primer pairs were checked for possible cross dimerization and melting temperature range. Each selected amplicon was submitted individually to QIIME for taxonomy assignment with OTU similarity of 97%. Compatible OTU databases used were Greengenes versions 6oct2010, 29nov2010, 4feb2011 and Silva version 111. For candidates evaluation we performed an in silico PCR with the selected primers from the last step using Silva and Greengenes 16S fasta databases. In candidates selection we got a total of 18 amplicons assigned taxonomically to Mycobacterium leprae OTUs in at least one database. Maybe the better hypervariable regions for taxonomy assignment are V2 and V6 because only amplicons which covers these regions were assigned to M. leprae in all databases. In candidates evaluation we got 9 primer pairs which could amplify at least 50% of Bacteria domain and 1 primer pair of Archaea for both databases. We can conclude that there are more efficient hypervariable regions for Mycobacterium leprae identification in environmental samples, however in the amplification step we can have a huge loss of information. In further analysis we are going to evaluate taxonomic classification with a mock community dataset.