12th International Meeting on Microbial Epidemiological Markers (IMMEM XII)
Dubrovnik, Croatia
18–21 September 2019

276 MultidimensionalSequence Typing (MDST): a new method to localize particular sequences in the universeof bacterial genomes

Miguel D. Fernandez-de-Bobadilla[1], Alba María Talavera Rodríguez[2;3], Fernando Baquero[1], Teresa Coque[1], Val Fernandez Lanza[2]

[1]IRYCIS, Microbiology, Madrid, Spain, *[2]IRYCIS, Bioinformatics Unit, Madrid, Spain*, *[3]IRYCIS, Infectious Diseases, Madrid, Spain*

Background: One of the major tasks in microbial epidemiology is to type the isolates with the high precision required to track outbreaks in different locations. Up to now the most extensively sequence-based tool is Multi Locus Sequence Typing (MLST) but the obtained accuracy is insufficient in the age of NGS. Based in the principles of MLST, two different ways of sequence typing have been developed: cgMLST and wgMLST. These approaches have two major weakness: i) the nomenclature (ST designation) in unrelated with the strain closeness; and ii) the method is too sensitive to small genetic changes (a single SNP in one gene produces a new sequence type). To overcome these limitations, we designed a new typing method providing data in Euclidean coordinates. Given a strain (as part of a certain species), a few coordinates are required to locate the strain in the Euclidean species space, so that we can immediately visualize the relationship of that strain with the rest of the strains represented in the universe of sequences.

Materials/methods: We present a new pipeline able to provide a natural classification of specie. Given a species database with *N* genomes we define a sub-set of *M* genomes representing the diversity of the whole dataset. This *M* genomes dataset is the *pre-coordinate system*. Principal Components Analysis (PCA) is applied to the *pre-coordinate system* to create the Euclidean space. MASH is used to calculate the distance of a query genome with the *pre-coordinate system*. Given a query strain, MASH creates *M* coordinates (one distance per genome) stored in a vector. The vector is then transformed by using the projection matrix of the PCA. Finally, we store the *K* Principal Components that contains the 95% of variability.

Results: We have tried our methods with several species (*Escherichia coli, Klebsiella pneumoniae, Enterococcus faecium* or *Staphylococcus aureus*) demonstrating that out method is universal and can be applied to any species.

Conclusions: We present an innovative universal typing system able to classify thousands of genomes of any species in a fast and scalable way, adding pseudo-phylogenetic information in the nomenclature.