Net7: a new tool for bacterial comparative genomics: massive tracing of vertical and horizontal gene flux between genome elements

Marina Manrique[1], Pablo Pareja-Tobes[1], Eduardo Pareja-Tobes[1], Marta Brozynska[1], Eduardo Pareja[1], Raquel Tobes[1]

[1]Oh no Sequences! , Era7 Bioinformatics , Granada, Spain.

rtobes@era7.com

Abstract. Identification and epidemiologic studies of pathogenic bacteria are mainly based on genotyping of a set of selected genes. The availability of Next Generation Sequencing technologies allows doing an exhaustive genotyping of the strains involved in outbreaks sequencing their whole genome. To have the sequences of the complete set of genes of a bacteria opens new strategies of analysis that can provide insight into their provenance and also into their possible evolution in the next future. The knowledge in these two directions can delineate the strategies of intervention including from prevention to treatment and epidemiologic surveillance.We have developed a tool to detect the complete set of similarity relationships of each protein from a genome element with all bacterial and archaeal proteins present in Uniprot, and hence with all the genome elements and taxonomic units to which the connected proteins pertain. This new tool for bacterial comparative genomics allows a massive tracing of vertical and horizontal gene flux between genome elements, based on the analysis of the similarity between their proteins. The tool analyzes similarity relationships that can be fixed to 90% or 100% of similarity threshold. The tool provides the data needed to obtain network representations with Hiveplot or gephi.The building of the network is based on Bio4j (http://bio4j.com/). Bio4j is a bioinformatics graph based DB including most data available in UniProt KB (SwissProt + Trembl), Gene Ontology (GO), UniRef (50,90,100), RefSeq, NCBI taxonomy, and Expasy Enzyme DB developed by Era7 Bioinformatics research group Oh no sequences!. The current version of Bio4j (0.7) includes 530.642.683 relationships and 76.071.411 nodes. Bio4j uses Neo4j technology, another Open Source project. Performance is one of the main advantages of the platform. In Bio4j data is organized in a way semantically equivalent to what it represents thanks to the graph structure. That means that queries which would even be impossible to perform with a standard Relational DB, just take a couple of seconds with Bio4j. Bio4j is an open source platform released under AGPLv3.Bio4j is freely available. Net7 will also be released under AGPLv3 Open Source license.