SNPs Detection in *Klebsiella pneumoniae*

Nicholas Costa Barroso Lima*[1], Maurício Egídio Cantão[2] and Marisa Fabiana Nicolás[1]

[1]Laboratório de Bioinformática (LabInfo), Laboratório Nacional de Computação Científica (LNCC). Petrópolis-RJ, Brasil;
[2]EMBRAPA Suínos e Aves. Concórdia-SC, Brasil.

Single nucleotide polymorphisms, SNPs, are common genome variations and may be associated with different phenotypes. This type of polymorphism is responsible for most of the genetic variability (90%) in a human genome, as discovered by the Human Genome Sequencing Project. Recently, several studies have focused on the detection of this type of polymorphism in bacterial genomes for use in bacterial strain typing and phylogeny reconstruction. In this work we developed a methodology for in silico detection and filtering of SNPs for bacterial genomes in order to analyze its prevalence.

In order to detect the SNPs, a methodology based on sequence alignment algorithms and filters developed in PERL were used in order to obtain a reliable final set of these polymorphisms. The occurrence of SNPs fits the Poisson probability distribution, as they are events that occur in an interval, in this case, coding sequences (CDSs). Within this context, we calculated the expected frequency of SNPs in CDSs using a Poisson probability distribution. Therefore, CDSs that exceeded the expected frequency of SNPs could be subjected to different selective pressure. The methodology was tested and evaluated for the draft genome of *Klebsiella pneumoniae* isolate Kp13, a bacteria causing nosocomial infection. This genome was set as a reference and was compared with three other *K. pneumoniae* genomes: *K. pneumoniae* 342, *K. pneumoniae* subsp. *pneumoniae* MGH 78578; and *K. pneumoniae* NTUH-K2044. *K. pneumoniae* strain 342 has a different niche and life style regarding the other two strains of the same species, MGH 78578 e NTUH-K2044. *K. pneumoniae* 342 lives associated with plants, and this bacterium is capable of nitrogen fixation and the other two are pathogens.

K. pneumoniae 342 showed the largest number of SNPs when compared with *K. pneumoniae* isolate Kp13. Whereas, MGH 78578 and NTUH-K2044 showed similar number of SNPs. The common subset of SNPs amongst the three comparisons showed an expected frequency of 1-3 per CDSs. Subsequently, the analysis of the product of CDSs with more than four SNPs of that common subset revealed no direct association with pathogenicity. On the other hand, the unique subset of SNPs for the comparisons of Kp13 with MGH 78578 or NTUH-K2044 showed an expected frequency of 1-8 and 1-7 SNPs per CDSs, respectively. Remarkably for this comparison, we found that the product of CDSs harboring more SNPs than the expected frequency were associated with major virulence factors, resistance factors and bacterial fitness. These results indicate a tendency of fixation of SNPs associated with a pathogenic lifestyle.

*nicholas@lncc.br