New approach for genomic comparison of invasive and non-invasive strains of Streptococcus pyogenes

Suzane de Andrade Barboza[1], Caio Rafael do Nascimento Santiago[1], Luciano Antonio Digiampietri[1],

1 Universidade de São Paulo

Abstract

Streptococcus pyogenes or Group A streptococcal (GAS) is a uniquely human Gram-positive pathogen related to a wide range of invasive and non-invasive diseases, having the fourth highest mortality rate among bacterial pathogens. The diversity of clinical outcomes of these infections can be explained by the acquisition of exogenous genetic material, mostly composed of virulence factors such as adhesins or phage toxins. One of the main virulence factors is M protein, which hypervariable region is used for GAS classification. Molecular epidemiology studies showed a genotype M/pathogenicity relation, which is being intensively investigated by genomic comparisons. However, little is known about different invasive levels observed within strains sharing the same genotype. This lack of information occurs due to two main reasons: (1) recent studies limit their comparisons by genotype or pathology analysis, disregarding non-invasive strains, and (2) software limitations concerning closely related genomes comparisons, which include difficulties in performing global alignments considering large genome rearrangements and the identification of strains' exclusive genes. In order to overcome these difficulties, two main strategies have been used in the comparison of 55 GAS genomes (28 genomes from invasive strains, 25 from non-invasive strains and 2 from isolates with unknown invasive profile): phylogenetic and gene network analysis, based on the identification of homologous genes. After performing local alignments with all genes of all genomes, homology relations were defined considering seven defined parameters: minimum identity percentage, minimum alignment percentage, minimum alignment length, maximum number of mismatched positions, maximum number of gap positions, maximum e-value, and minimum bit-score. The resulting graph is a gene network representation, where each homologous gene group is a cluster composed of nodes representing the genes of the genomes. Each genome is represented by a color, which allow us to identify gene sets exclusively found on invasive strains or strains related to a certain disease. A distance matrix of the genomes has then been calculated based on presence or absence of genes for each group of genes created previously, and a cladogram (representing the phylogenetic relationships) of all genomes was constructed, grouping strains with similar gene composition. Relating this information with the disease/virulence profile, we aim to better understand the relation between GAS genotypes and gene acquisition. These graphical representations will accelerate the identification of the virulence factors that could explain certain isolates' invasiveness and alternative genes for the production of an anti-streptococcal vaccine.

Funding: Capes