Shimantika Sharma, School of Biology, Georgia Institute of Technology, Atlanta, GA, USA
Alexander Bucksch, School of Biology & School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA
Joshua S. Weitz, School of Biology & School of Physics, Georgia Institute of Technology, Atlanta, GA, USA

G-DES: An Efficient Software For Microbial Gene-Level Diversity Estimation

Microbial genomes are dynamic; they frequently undergo different events, such as deletions, insertions or mutations, which change the sequence of genes and intergenic regions. These physical processes can induce genomic diversity in a microbial population. Advances in sequencing methods have increased our understanding of the structure and content of genomes. Previously, genomic differences were categorized in terms of the frequency of occurrence of gene variants (known as alleles). The analysis of changes in allele frequency is the foundation of population genetics. However, now, genomic differences can be measured in terms of difference in gene composition across a genome (rather than nucleotide differences between genes).Thus, there is an increasing need to quantify this "genomic variability" within a species.

The core and pan genome concepts have been proposed as one way to quantify genomic diversity within a group of organisms, e.g., within a species or genus[1]. The core genome is the set of genes found in every organism within a group. The pan genome is the set of all genes found within organisms of a group, including core genes and genes which appear in a fraction of genomes. Multiple attempts have been made to estimate the size of pan and core genomes in hopes of quantifying openness and closedness of a particular set of genomes to gene variation. Recent advances in the field pointed out that pan and core genome sizes are not robust to sampling and thus an alternative robust metric, **"genomic fluidity"** was proposed which can summarize the difference in gene content between genomes of closely related bacteria[2]. Genomic fluidity can be computed for a small number of sequenced genomes and can be used as a comparative metric between groups of closely related isolates. However, other forms of variation can also be measured, such as gene frequency distributions and the scaling of sample core and pan genome sizes. Hence, the scope of this project is to develop a microbial Gene-level Diversity Estimation Software (G-DES) to quantify microbial gene diversity in terms of two diversity metrics:

a. **Genomic fluidity**, which is defined as the ratio of unique gene clusters to the sum of gene clusters in pairs of genomes.

b. **Gene frequency distributions** which are defined as the frequency of genomes in which a particular gene occurs.

These two indices of gene diversity have the potential to enhance our understanding of gene compositional dynamics within individuals of the same species.

G-DES estimates the gene compositional differences between genomes of the same bacterial species and quantifies them by the two metrics given above. It is implemented as a collection of Perl modules. Usability is achieved by a GUI front-end giving access to all parameters. The only dependencies of the software will be Glimmer 3.0, NCBI Blast 2.2.25 and Bio-Perl 1.6.9. G-DES will then be tested against publicly available bacterial genome sequences. We utilize G-DES to quantify genomic composition within bacterial pathogens using hundreds of completely sequenced genomes.

References.

[1] Tettelin H, *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial pan-genome. *Proc Natl Acad Sci USA* 102:13950-13955

[2] Kislyuk AO, Haegeman B, Bergman N, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12:32