NEIGHBORHOOD CLUSTERING TO ANALYSE ANTIMICROBIAL RESISTANCE IN BACTERIAL GENOMES

Valuable insight into gene function and evolution can be obtained by analysing the conservation of gene order among prokaryotic genomes. Although there are many existing tools to visualize and compare the order of genes between genomes, only a few support the analysis of large numbers of genomes in an efficient and organized way. Obtaining precise functional predictions is particularly important in the case of antimicrobial resistance (AMR) genes, as subtle differences in similarity patterns can mean the difference between an organism being treatable or resistant to one or more antibiotics. Databases such as the Comprehensive Antibiotic Resistance Database (CARD) provide high-quality predictions, but there is a significant gray area (“Loose hits") where genes may or may not confer AMR.

We introduce an approach to compare the neighborhoods of AMR genes in genomes with different degrees of relatedness, to provide additional insight into their potential function. Our approach uses a predictive tool to first identify AMR genes with differing degrees of confidence, then applies novel similarity measures and application of the UPGMA, MCL and DBSCAN graph-clustering techniques to identify patterns of similarity among gene neighborhoods. This analysis is complemented by phylogenetic analysis to assess the similarity of identified genes as well as their neighborhoods. We also provide a graphical tool to visualize the gene content in sets of neighborhoods.

AMR gene neighborhoods were observed to be very similar within closely related members of species including *Salmonella* Heidelberg. Examination of the neighborhoods of Loose hits provided more information on their probable functions. The proximity of some Loose hits to other AMR genes in many neighborhoods provided additional evidence for their function, whereas in other cases the Loose hits were isolated and likely not associated with AMR. We also considered a more-diverse set of genomes that encompassed *Salmonella* as well as the frequent pathogens *Klebsiella pneumoniae, Citrobacter, Enterobacter and Escherichia coli.* In this set we found cases where seemingly poor Loose predictions were associated with clusters of AMR genes, and instances where gene order was surprisingly similar across distantly related genomes which is potentially suggestive of recent transmission of AMR genes between pathogenic organisms.

Our method provides significant new insights in the function of candidate AMR genes, and these refined predictions can be used for prediction of resistance at the organism level as well as inference of the processes by which AMR can be transmitted among organisms.