[BIOINFO 2024]

Comparative genomics applied to understand taxonomic relationship among *Neisseria* species

HyunJun Lee[1], Yoseop Kim[2], Ho-Jin Gwak[2], and Mina Rho[1,2,*]

[1] *Department of Biomedical Informatics, Hanyang University, Seoul, Korea*
[2] *Department of Computer Science and Engineering, Hanyang University, Seoul, Korea*
** Corresponding author: minarho@hanyang.ac.kr*

Calculating interspecies distances based on genomic information is crucial for accurate species classification. Conventionally, average nucleotide identity (ANI) based on whole genome sequences has been used for taxonomic classification. However, two major challenges exist. First, recovering complete genome sequences from clinical samples is often difficult. Therefore, evolutionary distances are frequently estimated based on specific marker genes, such as 16S ribosomal RNA (rRNA) or core genes within a taxon. However, discordance in evolutional distance among different genes can impact the results. Second, ANI estimates may vary depending on the genomic regions analyzed, potentially causing misclassification of closely related species that should not be taxonomically distinct.

In this study, we conducted comparative genomics approaches on 40 *Neisseria* species. Core genes and dispensable genes were identified, and evolutionary distances were analyzed using full-length 16S rRNA genes, core genes, dispensable genes, and whole-genome sequences. Our result revealed consistent evolutionary relationships across gene sets, with some discordance in 16S rRNA genes. Moreover, the identity distribution varied depending on the method used. Specifically, ANI analysis produced evolutionary relationships similar to those derived from core genes but showed an identity distribution approximately 5–10% lower.