[BIOINFO 2024]

Prediction of DNA marker candidates for species identification using k-mer based Machine learning model

Chi-hwan Kim and Gyoungju Nah*

Genome Analysis Center at National Instrumentation Center for Environmental Management, Seoul National University, Seoul, 08826, Korea
**Corresponding author: gnah.nicem@snu.ac.kr*

Searching for DNA marker sequences specific to one species is often challenging depending on the size of the available datasets. To alleviate such difficulties, we developed a method for the prediction of candidate DNA marker sequences for species identification using a k-mer-based Machine learning model. The method starts with counting all possible k-mers from the input genomes. Using k-mer frequency patterns, the machine learning model between k-mer and species is built. Finally, the k-mer sequences that significantly affect the model's efficacy are extracted as the candidates for species-specific DNA markers. We validated the method on the test dataset and found many predicted candidates were matched well with previously known species-specific markers.

Acknowledgments: This research was supported by a grant [22193MFDS471-3] from Ministry of Food and Drug Safety, Republic of Korea in 2024.