[BIOINFO 2025]

Development of machine learning model for prediction in virulence level of pathogenic *Escherichia coli*

Yoojung Hwang[1,2], Woo Young Cho[3], Mi-Ran Seo[3], Nam Dahyun[4], Yeun-Jun Chung[1,5,6], and Seung-Hyun Jung[1,2,5*]

[1] *Department of Medical Sciences, Graduate School of The Catholic University of Korea, Seoul, Republic of Korea*

[2] *Catholic Research Institute for Human Genome Polymorphism, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

[3] *ConnectaGen, Hanam, Republic of Korea*

[4] *Basic Medical Science Facilitation Program, Catholic Medical Center, The Catholic University of Korea, Seoul, Republic of Korea*

[5] *Department of Microbiology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

[6] *Department of Biochemistry, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea*

**Corresponding author: hyun@catholic.ac.kr*

Pathogenic *Escherichia coli* (*E. coli*) is a major cause of foodborne illness globally, capable of inducing severe symptoms such as hemorrhagic diarrhea. To predict *E. coli* virulence, we analyzed 1,030 isolates collected from food, human, animal, and environmental sources using Whole-Genome Sequencing (WGS). A custom bioinformatics pipeline was applied to extract genomic features, and strains were categorized into high, intermediate, or low pathogenicity based on investigated pathotypes and human infection history. Four machine learning algorithms— Gradient Boosting Machine (GBM), Random Forest (RF), and Support Vector Machines (SVM) with linear and radial basis function (RBF) kernels —were trained and evaluated for predictive performance. The best-performing model, selected based on validation set performance, achieved an accuracy of 0.9224 and Area Under the Curve (AUC) of 0.9748. Its predictive capability was further supported by evaluation on external datasets. Feature importance analysis further enabled the identification of key virulence genes, supporting the development of a simplified, cost-effective predictive model. These results demonstrate the potential of genome-based machine learning approaches in enhancing the early identification and risk assessment of pathogenic *E. coli*, although further refinement is needed for intermediate-level strains.