(1) Submission ID#1527335

A novel method for the classification of entire genomes of Neisseria meningitidis with a bag-of-words approach and machine learning

Author(s)

Marco Podda, n/a
Visiting Scientist
University of Pisa

Simone Bonechi, n/a
Assistant Professor
University of Siena

Andrea Palladino, n/a
Senior data scientist
GSK

Mattia Scaramuzzino, n/a
Research Fellow
GSK

Alessandro Brozzi, n/a
Senior Data Scientist
GSK

Guglielmo Roma, n/a
Head of Discovery Data Science Centre of Excellence
GSK

Alessandro Muzzi, n/a
Associate Director, Head of Bacterial Computational Genomics Team
GSK Vaccines

Corrado Priami, n/a
Professor
University of Pisa

Alina Sîrbu, n/a

Associate Professor
University of Pisa

Margherita Bodini, n/a
Senior Data Scientist
GSK

Background
Whole genome sequencing of bacteria has many applications in strain classification for surveillance purposes that would benefit from the use of machine learning. However, using entire bacterial genomes as input for machine learning algorithms poses difficulties due to the genome sizes, fragmented sequences, and substantial variability between strains which all impede direct comparison of the genomes without alignment or typing.

Aim/Methods
Our objective was to implement and test a machine learning method to classify bacterial genomes. Hereto, we developed a “bag-of-words” approach to first encode entire bacterial genomes using SentencePiece tokenization and then to analyze these with machine learning.

Results
After training and validation, we classified the capsule B group genotype in Neisseria meningitidis genomes with 99.6% accuracy in an independent test set. Using capsule B genomes, we showed that removing the capsule region from the genomic sequence significantly reduced the probability of capsule B classification of these genomes (p-value 4.66e-10, Wilcoxon paired t-test) compared to removing random regions of the same sequence length. Similarly, after training and validation, we classified the multifactor invasive/carriage phenotype in Neisseria meningitidis genomes with 90.2% accuracy in an independent test set. Using invasive genomes, we showed that removing 155 known virulence factors (VFs) from the genomic sequence significantly reduced the probability of invasive classification of these genomes (p-value 7.12e-72, Wilcoxon paired t-test) compared to removing random sequences of the same length. Subsequently, we studied genes predicted by the model to be important for invasiveness. In silico knockouts of all 2,808 Neisseria meningitidis genes confirmed that the model predictions align with our current understanding of the underlying biology. Four of the top 10 genes found to be most important for invasive phenotypes were known VFs (cssA, cssB, lbpA, cssC), while the other six (secA, NMB0060, ygfZ, tamB, NMB1803, fixN) are plausible novel VFs.

Conclusions
To our knowledge, this is the first machine learning method using entire bacterial genomes to classify strains and identify genes considered relevant by the classifier. This method may be useful for strain classification application towards population genomics, antibiotic resistance monitoring, outbreak investigation, and strain coverage prediction.