World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:14, No:03, 2020

Phenotype Prediction of DNA Sequence Data: A Machine and Statistical Learning Approach

Authors : Mpho Mokoatle, Darlington Mapiye, James Mashiyane, Stephanie Muller, Gciniwe Dlamini

Abstract : Great advances in high-throughput sequencing technologies have resulted in availability of huge amounts of sequencing data in public and private repositories, enabling a holistic understanding of complex biological phenomena. Sequence data are used for a wide range of applications such as gene annotations, expression studies, personalized treatment and precision medicine. However, this rapid growth in sequence data poses a great challenge which calls for novel data processing and analytic methods, as well as huge computing resources. In this work, a machine and statistical learning approach for DNA sequence classification based on k-mer representation of sequence data is proposed. The approach is tested using whole genome sequences of Mycobacterium tuberculosis (MTB) isolates to (i) reduce the size of genomic sequence data, (ii) identify an optimum size of k-mers and utilize it to build classification models, (iii) predict the phenotype from whole genome sequence data of a given bacterial isolate, and (iv) demonstrate computing challenges associated with the analysis of whole genome sequence data in producing interpretable and explainable insights. The classification models were trained on 104 whole genome sequences of MTB isoloates. Cluster analysis showed that k-mers maybe used to discriminate phenotypes and the discrimination becomes more concise as the size of k-mers increase. The best performing classification model had a k-mer size of 10 (longest k-mer) an accuracy, recall, precision, specificity, and Matthews Correlation coeffient of 72.0%, 80.5%, 80.5%, 63.6%, and 0.4 respectively. This study provides a comprehensive approach for resampling whole genome sequencing data, objectively selecting a k-mer size, and performing classification for phenotype prediction. The analysis also highlights the importance of increasing the k-mer size to produce more biological explainable results, which brings to the fore the interplay that exists amongst accuracy, computing resources and explainability of classification results. However, the analysis provides a new way to elucidate genetic information from genomic data, and identify phenotype relationships which are important especially in explaining complex biological mechanisms.

Keywords : AWD-LSTM, bootstrapping, k-mers, next generation sequencing
Conference Title : ICBCBBE 2020 : International Conference on Bioinformatics, Computational Biology and Biomedical Engineering
Conference Location : London, United Kingdom
Conference Dates : March 12-13, 2020