diagnostics

MDPI

Article

Ensemble Machine Learning Model to Predict SARS-CoV-2 T-Cell Epitopes as Potential Vaccine Targets

Syed Nisar Hussain Bukhari [1,*], Amit Jain [1], Ehtishamul Haq [2], Abolfazl Mehbodniya [3] and Julian Webber [4]

[1] University Institute of Computing, Chandigarh University, NH-95, Chandigarh-Ludhiana Highway, Mohali 140413, India; amit_jainci@yahoo.com

[2] Department of Biotechnology, University of Kashmir, Srinagar 190006, India; haq@uok.edu.in

[3] Department of Electronics and Communication Engineering, Kuwait College of Science and Technology, Kuwait City 13133, Kuwait; a.niya@kcst.edu.kw

[4] Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan; jwebber@ieee.org

* Correspondence: nisar.bukhari@gmail.com

Abstract: An ongoing outbreak of coronavirus disease 2019 (COVID-19), caused by a single-stranded RNA virus called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused a worldwide pandemic that continues to date. Vaccination has proven to be the most effective technique, by far, for the treatment of COVID-19 and to combat the outbreak. Among all vaccine types, epitope-based peptide vaccines have received less attention and hold a large untapped potential for boosting vaccine safety and immunogenicity. Peptides used in such vaccine technology are chemically synthesized based on the amino acid sequences of antigenic proteins (T-cell epitopes) of the target pathogen. Using wet-lab experiments to identify antigenic proteins is very difficult, expensive, and time-consuming. We hereby propose an ensemble machine learning (ML) model for the prediction of T-cell epitopes (also known as immune relevant determinants or antigenic determinants) against SARS-CoV-2, utilizing physicochemical properties of amino acids. To train the model, we retrieved the experimentally determined SARS-CoV-2 T-cell epitopes from Immune Epitope Database and Analysis Resource (IEDB) repository. The model so developed achieved accuracy, AUC (Area under the ROC curve), Gini, specificity, sensitivity, F-score, and precision of 98.20%, 0.991, 0.994, 0.971, 0.982, 0.990, and 0.981, respectively, using a test set consisting of SARS-CoV-2 peptides (T-cell epitopes and non-epitopes) obtained from IEDB. The average accuracy of 97.98% was recorded in repeated 5-fold cross validation. Its comparison with 05 robust machine learning classifiers and existing T-cell epitope prediction techniques, such as NetMHC and CTLpred, suggest the proposed work as a better model. The predicted epitopes from the current model could possess a high probability to act as potential peptide vaccine candidates subjected to in vitro and in vivo scientific assessments. The model developed would help scientific community working in vaccine development save time to screen the active T-cell epitope candidates of SARS-CoV-2 against the inactive ones.

Keywords: COVID-19; SARS-CoV-2; T-cell epitope; peptide-based vaccines; machine learning; random forest; ensemble learning; voting ensemble

Citation: Bukhari, S.N.H.; Jain, A.; Haq, E.; Mehbodniya, A.; Webber, J. Ensemble Machine Learning Model to Predict SARS-CoV-2 T-Cell Epitopes as Potential Vaccine Targets. *Diagnostics* **2021**, *11*, 1990. https://doi.org/10.3390/diagnostics11111990

Academic Editor: Tim Duong

Received: 13 September 2021
Accepted: 21 October 2021
Published: 26 October 2021

1. Introduction

An infection outbreak caused by a novel coronavirus has proliferated rapidly around the world. The World Health Organization (WHO) designated the disease as COVID-19 [1,2]. The pathogen was named SARS-CoV-2 by the Coronaviridae Study Group (CSG) [3]. The pathogen has resulted in 225,488,491 COVID-19 cases and 4,644,376 deaths worldwide as of September 13, 2021, posing a significant challenge to public health worldwide [4]. Furthermore, because SARS-CoV-2 keeps on circulating, the chances of mutations in the virus also increases. The recent delta variant with *Pango lineage* as AY.1, AY.2, AY.3,