

Studi QSAR *in-house molecules* sebagai Inhibitor *Papain-like Protease (PLpro)* SARS-CoV-2 menggunakan Metode Ensemble

Ananda Fitri Karimah, Annisa Aditsania, Isman Kurniawan

Fakultas Informatika, Universitas Telkom, Bandung

anandafitri@students.telkomuniversity.ac.id, aaditsania@telkomuniversity.ac.id,
ismankrn@telkomuniversity.ac.id

Abstrak

Sudah lebih dari satu tahun sejak *World Health Organization* (WHO) menyatakan bahwa penyakit *Coronavirus disease 19* (COVID-19) adalah pandemi. Penyakit ini telah terkonfirmasi oleh WHO sebanyak 205.338.159 kasus positif per tanggal 13 Agustus 2021. Salah satu protein yang diproduksi oleh SARS-CoV-2 yang menjadi kunci penting dalam mereplikasi virus SARS-CoV-2 adalah *Papain-like Protease (PLpro)* SARS-CoV-2, protein ini menjadi salah satu target penting karena terlibat dalam replikasi virus. *Quantitative structure-activity relationship* (QSAR) adalah pendekatan komputasi kimia yang digunakan dalam salah satu langkah *drug discovery*, kelebihan menggunakan QSAR salah satunya adalah dapat mengurangi kegagalan mahal akibat banyaknya penelitian yang dilakukan. Penelitian ini membangun model klasifikasi QSAR *in house molecules* menggunakan metode *ensemble learning*, yaitu algoritma *Random Forest* dan *AdaBoost* untuk memprediksi nilai aktif pada dataset. Dataset yang digunakan pada penelitian ini berjumlah 90 buah dan dibagi menjadi 2 kategori yaitu nilai 1 yang berartikan molekul aktif dan 0 untuk molekul tidak aktif. Penelitian ini juga menggunakan dua jenis deskriptor *fingerprint* yaitu *PubChem* dan *Extended*. Tujuan dari penelitian ini adalah membangun model prediksi QSAR, menilai dan menganalisa kinerja model-model yang dibangun dan menilai kinerja algoritma yang dipilih, juga melihat pengaruh yang diberikan oleh jenis *fingerprint*. Hasil terbaik yang didapatkan pada penelitian ini yaitu kombinasi *PubChem* dan algoritma *Random Forest* dengan nilai akurasi dan F1-Score berturut-turut bernilai sebesar 81,48% dan 88,89%.

Kata kunci : *quantitative structure-activity relationship* (QSAR), *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), *Random Forest*, *AdaBoost*, *coronavirus disease 19* (COVID-19)

Abstract

It has been more than a year since the World Health Organization (WHO) declared that Coronavirus disease 19 (COVID-19) is a pandemic. This disease has been confirmed by WHO as many as 205,338,159 positive cases as of August 13, 2021. One of the proteins produced by SARS-CoV-2 which is an important key in replicating the SARS-CoV-2 virus is Papain-like Protease (PLpro) SARS-CoV-2, this protein is an important target because it is involved in viral replication. Quantitative structure-activity relationship (QSAR) is a computational chemical approach used in one of the drug discovery steps, one of the advantages of using QSAR is that it can reduce costly failures due to the large number of studies conducted. This study builds a QSAR in house molecules classification model using the ensemble learning method, namely the *Random Forest* and *AdaBoost* algorithms to predict the active values in the dataset. The dataset used in this study amounted to 90 pieces and was divided into 2 categories, namely a value of 1 which means active molecules and 0 for inactive molecules. This study also uses two types of *fingerprint* descriptors, namely *PubChem* and *Extended*. The purpose of this study is to build a QSAR prediction model, assess and analyze the performance of the models built and assess the performance of the selected algorithm, also see the effect given by the type of *fingerprint*. The best results obtained in this study were the combination of *PubChem* and the *Random Forest* algorithm with accuracy and F1-Score values of 81.48% and 88.89%, respectively.

Keywords: *quantitative structure-activity relationship* (QSAR), *severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2), *Random Forest*, *AdaBoost*, *coronavirus disease 19* (COVID-19)
