Klasifikasi sekuens protein coronavirus menggunakan Metode K-Nearest Neighbor dan seleksi fitur algoritma genetika = Classification of coronavirus protein sequences using K-Nearest Neighbor method and feature selection genetic algorithm

Mufarrido Husnah, author

Abstrak

Coronavirus (CoV) adalah keluarga virus penyebab penyakit sistem pernapasan ringan hingga berat pada berbagai spesies hewan termasuk manusia. Salah satu spesies Coronavirus yang muncul pada akhir tahun 2019 yaitu SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) dan menimbulkan penyakit baru bernama Covid-19 (Coronavirus disease-2019) kemudian berstatus pandemi. Penyebaran Covid-19 yang cepat dan dengan tingkat kematian yang tinggi terus terjadi di berbagai negara. Oleh karena itu, deteksi dini patogen perlu dilakukan secara cepat dengan menggunakan data sekuens protein Coronavirus. Sekuens protein merupakan data struktur primer dari suatu protein yang memiliki 27 fitur berdasarkan discere. Dalam penerapannya, tidak semua fitur relevan dengan data yang digunakan sehingga perlu seleksi fitur untuk menghindari dimensi data yang tinggi dan tidak optimal. Seleksi fitur algoritma genetika memberikan fitur-fitur optimal pada data dan metode K-Nearest Neighbor (KNN) melakukan klasifikasi data sekuens protein Coronavirus dengan fitur hasil seleksi fitur algoritma genetika. Seleksi fitur algoritma genetika menghasilkan 11 fitur optimal yang meningkatkan performa running time metode klasifikasi KNN menjadi 0,0541 detik. Fitur optimal diperoleh dari karakteristik AA-count , secondary structure fraction , isoelectric point dan instability index. Hasil terbaik performa akurasi, spesifisitas beserta sensitifitas secara berurutan yaitu 96,68%, 98,7% dan 94,4% yang diperoleh pada nilai parameter K=3.
.....Coronaviruses (CoV) are a family of viruses that cause mild to severe respiratory system diseases in various animal species including humans. One of the Coronavirus species that emerged at the end of 2019 was SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) and caused a new disease called Covid-19 (Coronavirus disease-2019) then had a pandemic status. The rapid spread of Covid-19 and with a high death rate continues to occur in most of countries. Therefore, early detection of pathogens needs to be done quickly using Coronavirus protein sequence data. Protein sequences are primary structural data of a protein that has 27 features but not all of the existing features are relevant to the data used, so feature selection is necessary to avoid high and suboptimal data dimensions. The genetic algorithm feature selection provides optimal features to the data and the K-Nearest Neighbor (KNN) method performs the classification of Coronavirus protein sequences data with features resulting from the genetic algorithm feature selection. The genetic algorithm feature selection produces 11 optimal features that improve the running time performance of the KNN classification method. The average result of running time is 0.0541 second. The best results were accuracy performance, specificity and sensitivity are 96.68%, 98.7% and 94.4% respectively which were obtained at the parameter value K=3.