Analisis kekerabatan pada barisan DNA SARS-CoV-2 berdasarkan pembentukan pohon filogenetik dengan metode Hierarchical dan K-Means Clustering menggunakan Multiple Encoding Vector dan K-Mer = Implementation of Hierarchical and K-Means Clustering Methods using Multiple Encoding Vector in analyzing kinship in SARS-CoV-2 DNA Sequences

Banjarnahor, Evander, author

Abstrak

Berdasarkan data WHO pada pertengahan Juli 2021 lebih dari 185,2 juta orang di seluruh dunia terinfeksi virus corona atau Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Virus ini menyerang penapasan manusia yang dapat mengakibatkan infeksi paru-paru pada manusia dan bahkan dapat menyebabkan kematian. Tercatat bahwa lebih dari 4 juta orang di seluruh dunia meninggal akibat terinfeksi virus corona. Di Indonesia sendiri pada pertengahan Juli 2021 tercatat lebih dari 2,4 juta orang ternfeksi virus corona dan lebih dari 65,4 ribu orang meninggal akibat terinfeksi virus corona. Berdasarkan data tersebut, perlu dilakukan analisis kekerabatan virus SARS-CoV-2 untuk mengurangi penyebaran dan memberikan batasan sosial dari negara satu dengan negara lainnya. Identifikasi kekerabatan dari virus covid-19 dan penyebarannya dapat dilakukan dengan cara pembentukan pohon filogenetik dan clustering. Pada penelitian ini pohon filogenetik akan dibangun berdasarkan metode Hierarchical Clustering dengan menggunakan metode Multiple Encoding Vector dan K-Mer berdasarkan translasi DNA kodon menjadi asam amino. Jarak Euclidean akan digunakan untuk menentukan matriks jarak. Penelitian ini selanjutnya menggunakan metode K- Means Clustering untuk melihat penyebarannya, dimana nilai k ditentukan dari jumlah centroid yang dihasilkan dari metode Hierarchical Clustering. Penelitian ini mengambil sampel barisan DNA SARS-CoV-2 dari beberapa negara yang tertular. Dari hasil simulasi, nenek moyang SARS-CoV-2 berasal dari China. Hasil analisis juga menunjukkan bahwa leluhur covid-19 yang paling dekat dengan Indonesia berasal dari India, Australia dan Spanyol. Selain itu dari hasil simulasi dihasilkan bahwa barisan DNA SARS-CoV-2 terdiri dari 9 cluster dan cluster keenam adalah kelompok yang memiliki anggota paling banyak. Hasil analisis juga menunjukkan bahwa metode ini sangat opitimal dalam pengelompokan data dengan nilai 97.4%.

.....Based on WHO data in middle of July 2021, Coronavirus or Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is infecting more than 185.2 million people worldwide. The virus attacks human breathing, which can cause lung infections and can even cause death. More than 4 million people worldwide have died due to being infected with the coronavirus. In Indonesia alone, in mid-July 2021, there were more than 2.4 million people infected with the corona virus and more than 65.4 thousand people died from being infected with the corona virus. Based on those covid-19 survivor data, it is necessary to carry out a kinship analysis of the coronavirus to reduce its spreading. Identification of the kinship of the covid- 19 virus and its spread can be done by forming a phylogenetic tree and clustering. This study uses the Multiple Encoding Vector method and K-mer based on translation DNA codon to amino acid in analyzing sequences and Euclidean Distance to determine the distance matrix. This research will then use the Hierarchical Clustering method to determine the number of initial centroids and cluster, which will be used later by the

K-Means Clustering method kinship in SARS-CoV-2 DNA sequence. This study took samples of DNA sequences of SARS-CoV-2 from several infected countries. From the simulation results, the ancestors of SARS-CoV-2 came from China. The results of the analysis also show that the closest ancestors of covid-19 to Indonesia came from India, Australia and Spain. In addition, the ancestors of SARS-CoV-2 came from China. The SARS- CoV-2 DNA sequence is also consisted of 9 clusters, and the sixth cluster is the group that has the most members. The results also show that this method is very optimal in a grouping of data with a value of 97.4%.