Penerapan algoritma K- Means clustering pada pengelompokan barisan DNA virus hepatitis B (HBV) = Application of K-Means algorithm in clustering the DNA sequences of hepatitis B virus (HBV) / Nova Yuniarti

Nova Yuniarti, author

Abstrak

[ABSTRAK
Berdasarkan data WHO tahun 2014, diperkirakan sekitar 15 juta orang di dunia yang terinfeksi hepatitis B (HBsAg+) juga terinfeksi hepatitis D. Infeksi hepatitis D dapat terjadi bersamaan (koinfeksi) atau setelah seseorang terkena hepatitis B kronis (superinfeksi). Penyakit hepatitis B disebabkan oleh virus HBV dan penyakit hepatitis D disebabkan oleh virus HDV. HDV tidak dapat hidup tanpa HBV. Hepatitis D erat hubungannya dengan infeksi virus HBV, sehingga sangat realistis bila setiap usaha pencegahan terhadap hepatitis B, maka secara tidak langsung mencegah hepatitis D. Pada tesis ini akan dibahas bagaimana hasil pengelompokan barisan DNA HBV menggunakan algoritma k-means clustering dengan menggunakan perangkat lunak R. Dimulai dengan mengumpulkan barisan DNA HBV yang diambil dari GenBank, kemudian dilakukan ekstraksi ciri menggunakan n-mers frequency, dan hasil ekstraksi ciri barisan DNA tersebut dikumpulkan dalam sebuah matriks dan dilakukan normalisasi menggunakan normalisasi min-max dengan interval [0, 1] yang akan digunakan sebagai data masukan. Jumlah cluster yang dipilih dalam penelitian ini adalah dua dan penentuan centroid awal dilakukan secara acak. Pada setiap iterasi dihitung jarak masing-masing objek ke masing-masing centroid dengan menggunakan Euclidean distance dan dipilih jarak terpendek untuk menentukan keanggotaan objek di suatu cluster sampai akhirnya terbentuk dua cluster yang konvergen. Hasil yang diperoleh adalah virus HBV yang berada pada cluster pertama lebih ganas dibanding virus HBV yang berada pada cluster kedua, sehingga virus HBV pada cluster pertama berpotensi berevolusi dengan virus HDV menjadi penyebab penyakit hepatitis D.
<hr>
ABSTRACT
Based on WHO data, an estimated of 15 millions people worldwide who are infected by hepatitis B (HBsAg+) are also infected by hepatitis D. Hepatitis D infection can occur simultaneously with hepatitis B (co infection) or after a person is exposed to chronic hepatitis B (super infection). Hepatitis B is caused by the HBV virus and hepatitis D is caused by HDV virus. HDV can not live without HBV. Hepatitis D virus is closely related to HBV infection, hence it is really realistic that every effort of prevention against hepatitis B can indirectly prevent

hepatitis D. This thesis discussed the clustering of HBV DNA sequences by using k-means clustering algorithm and R programming. Clustering processes is started with collecting HBV DNA sequences that are taken from GenBank, then performing extraction HBV DNA sequences using n-mers frequency and furthermore the extraction results are collected as a matrix and normalized using the min-max normalization with interval [0, 1] which will later be used as an input data. The number of clusters is two and the initial centroid selected of cluster is choosed randomly. In each iteration, the distance of every object to each centroid are calculated using the Euclidean distance and the minimum distance are selected to determine the membership in a cluster until two convergent clusters are created. As the result, the HBV viruses in the first cluster is more virulent than the HBV viruses in the second cluster, so the HBV viruses in the first cluster can potentially evolve with HDV viruses that cause hepatitis D., Based on WHO data, an estimated of 15 millions people worldwide who are
infected by hepatitis B (HBsAg+) are also infected by hepatitis D. Hepatitis D infection can occur simultaneously with hepatitis B (co infection) or after a person is exposed to chronic hepatitis B (super infection). Hepatitis B is caused by the HBV virus and hepatitis D is caused by HDV virus. HDV can not live without HBV. Hepatitis D virus is closely related to HBV infection, hence it is really realistic that every effort of prevention against hepatitis B can indirectly prevent hepatitis D. This thesis discussed the clustering of HBV DNA sequences by using k-means clustering algorithm and R programming. Clustering processes is started with collecting HBV DNA sequences that are taken from GenBank, then performing extraction HBV DNA sequences using n-mers frequency and furthermore the extraction results are collected as a matrix and normalized using the min-max normalization with interval [0, 1] which will later be used as an input data. The number of clusters is two and the initial centroid selected of cluster is choosed randomly. In each iteration, the distance of every object to each centroid are calculated using the Euclidean distance and the minimum distance are selected to determine the membership in a cluster until two convergent clusters are created. As the result, the HBV viruses in the first cluster is more virulent than the HBV viruses in the second cluster, so the HBV viruses in the first cluster can potentially evolve with HDV viruses that cause hepatitis D.]