Penerapan metode pengelompokan hierarchical ordered partitioning and collapsing hybrid (Hopach) untuk menganalisis kekerabatan virus ebola = Application of hierarchical ordered partitioning and collapsing hybrid method to analyzing phylogenetically on ebola virus

Hengki Muradi, author

Abstrak

[Salah satu tujuan dalam studi ekpresi gen (DNA/Protein) adalah menemukan subbagian
yang penting secara biologis dan kelompok-kelompok dari gen-gen. Pengelompokan gen tersebut dapat dilakukan dengan metode hirarki maupun metode partisi. Kedua metode pengelompokan dapat dikombinasikan, dimana
dilakukan fase partisi dan hirarki secara bergantian, metode ini dikenal dengan metode Hopach. Tahap partisi dapat dilakukan dengan metode PAM, SOM, atau K-Means. Proses partisi dilanjutkan dengan proses Ordered, baru kemudian dikoreksi dengan proses agglomorative, sehingga hasil pengelompokan menjadi lebih akurat. Dalam menentukan kelompok utama digunakan ukuran MSS (Median Split Silhouette). MSS mengukur homogenitas hasil pengelompokan,
dimana hasil pengelompokan yang dipilih adalah yang meminimumkan MSS. Pada pengelompokan 136 barisan DNA Virus Ebola dari GeneBank. Proses
awalnya dilakukan pensejajaran global, dan dilanjutkan dengan perhitungan jarak genetik dengan menggunakan koreksi Jukes-Cantor. Pada penelitian ini didapat jarak genetik maksimum adalah 0.6153407 sedangkan jarak genetik minimum adalah 0. Selanjutnya matriks jarak genetik dapat dijadikan dasar untuk mengelompokkan barisan-barisan tersebut dengan menggunakan metode Hopach. Pada hasil pengelompokan Hopach-PAM, diperoleh kelompok utama sebanyak 10 kelompok dengan nilai MSS sebesar 0,8873843. Kelompok-kelompok virus ebola dapat diidentifikasikan berdasarkan subspesies dan tahun pertama kali mewabah.
Proses pensejajaran global dan pengelompokan Hopach-PAM menggunakan bantuan program open source R.
......One goal in the study of gene expression (DNA/Protein) is finding biologically important subsets and clusters of genes. Clustering these genes can be achieved by hierarchical and partitioning methods. Both clustering methods can be combined, where partition and hierarchy phases can be executed alternately, this method is known as a Hopach method. The partitioning step can be done by the PAM, SOM, or K-Means clustering method. The partition process continued with the process of Ordered, then corrected with agglomorative process, so that the clustminering results become more accurate. The main clusters determine by using MSS
(Median Split Silhouette). MSS is used to measure homogeneity of the clustering result, in which the clustering is selected to minimize its MSS. The clustering procceses of 136 DNA sequences of Ebola virus, are started by performing a global alignment, and continued with the genetic distance calculations using Jukes-Cantor correction. In this research we found the maximum genetic distance is 0.6153407, meanwhile the minimum genetic distance is 0. Furthermore, the genetic distance matrix can be used as a basis for clustering sequences in Hopach-PAM clustering method. Based on, the clustering results, we obtained 10

major clusters with MSS value of 0.8873843. Ebola virus clusters can be identified by subspecies and the first occoring year of their outbreak. We implemented the global alignment process and Hopach-PAM clustering algorithm using the open source program R.;One goal in the study of gene expression (DNA/Protein) is finding biologically important subsets and clusters of genes. Clustering these genes can be achieved by hierarchical and partitioning methods. Both clustering methods can be combined, where partition and hierarchy phases can be executed alternately, this method is known as a Hopach method. The partitioning step can be done by the PAM, SOM, K-Means clustering method. The partition process continued with the process of Ordered, then corrected with agglomorative process, so that the clustmineringresults become more accurate. The main clusters determine by using MSS (Median Split Silhouette). MSS is used to measure homogeneity of the clustering result, in which the clustering is selected to minimize its MSS. The clustering procceses of 136 DNA sequences of Ebola virus, are started by performing a global alignment, and continued with the genetic distance calculations using Jukes-Cantor correction. In this research we found the maximum genetic distance is 0.6153407, meanwhile the minimum genetic distance is 0. Furthermore, the genetic distance matrix can be used as a basis for clustering sequences in Hopach-PAM clustering method. Based on, the clustering results, we obtained 10 major clusters with MSS value of 0.8873843. Ebola virus clusters can be identified by subspecies and the first occoring year of their outbreak. We implemented the global alignment process and Hopach-PAM clustering algorithm using the open source program R., One goal in the study of gene expression (DNA/Protein) is finding biologically important subsets and clusters of genes. Clustering these genes can be achieved by hierarchical and partitioning methods. Both clustering methods can be combined, where partition and hierarchy phases can be executed alternately, this method is known as a Hopach method. The partitioning step can be done by the PAM, SOM, or K-Means clustering method. The partition process continued with the process of Ordered, then corrected with agglomorative process, so that the clustminering results become more accurate. The main clusters determine by using MSS (Median Split Silhouette). MSS is used to measure homogeneity of the clustering result, in which the clustering is selected to minimize its MSS. The clustering procceses of 136 DNA sequences of Ebola virus, are started by performing a global alignment, and continued with the genetic distance calculations using Jukes-Cantor correction. In this research we found the maximum genetic distance is 0.6153407, meanwhile the minimum genetic distance is 0. Furthermore, the genetic distance matrix can be used as a basis for clustering sequences in Hopach-PAM clustering method. Based on, the clustering results, we obtained 10 major clusters with MSS value of 0.8873843. Ebola virus clusters can be identified by subspecies and the first occoring year of their outbreak. We implemented the global alignment process and Hopach-PAM clustering algorithm using the open source program R.]