Metode Triclustering delta-Trimax Melalui Pendekatan Two-Way K-Means Menggunakan Geneontology Data Ekspresi Gen = delta-Trimax Method through Two-Way K-means Approach Using Gene Ontology on Gene Expression Data

Teguh Saputra, author

Abstrak

Analisis triclustering merupakan teknik yang mampu mengelompokkan data 3 dimensi secara bersamaan, sehingga dapat diperoleh sub-ruang dari data 3D yang terdiri dari subset observasi (gen), subset kondisi (kondisi) dan subset konteks (waktu). Analisis triclustering yang dilakukan pada penelitian ini yaitu metode delta-Trimax melalui pendekatan   two-way   K-means.   Tujuan   dari   metode delta-Trimax yaitu menemukan tricluster yang memiliki nilai minimum dari three-dimensial mean square residual (ð3) dan volume maksimum. Pendekatan two-way K-means digunakan untuk membentuk suatu populasi awal agar dapat mengurangi beban komputasi dan membantu membentuk tricluster yang lebih baik. Metode ini akan diimplementasikan pada data ekspresi gen kultur HAE (Human Airway Epithelial) yang terinfeksi virus SARS-CoV, SARS-dORF6, SARS-BatSRBD, dan H1N1. Implementasi dilakukan dengan 9 simulasi dan diperoleh simulasi terbaik dengan nilai threshold dari perhitungan MSR sebesar 0.0435, threshold = 1.7 dan sebanyak 24 tricluster terbentuk berdasarkan penilain triclustering quality index (TQI). Dari himpunan tricluster tersebut diperoleh informasi mengenai perbandingan pola ekspresi gen pada virus SARS-CoV, SARS-dORF6, SARS-BatSRBD dengan virus influenza H1N1. Terdapat 7 tricluster yang memiliki kesamaan pola ekspresi gen di setiap kondisi dan 8 tricluster yang diduga memiliki perbedaan kondisi antara setiap variasi virus SARS- CoV dengan virus influenza H1N1. Pada tricluster lainnya juga diperoleh informasi hanya beberapa variasi Sars-CoV yang memiliki kesamaan satu sama lain dan juga kesamaan atau perbedaan dengan H1N1. Berdasarkan titik waktu diperoleh 3 tricluster tidak memberikan efek karena pola ekspresi gen tiap waktu sama dengan kondisi awal yaitu titik waktu ke-1 dan 17 tricluster diduga memberikan efek paska infeksi. Untuk menilai kualitas hasil tricluster terbentuk dalam penggambaran fungsi biologis dari kumpulan gen pada tricluster dilakukan evaluasi gene ontology (GO). GO adalah sebuah sistem untuk menggambarkan fungsi, biological process, celluler componet gen dan moleculer function dalam berbagai organisme. Dari hasil evaluasi diperoleh sebanyak 20 tricluster yang memiliki keterlibatan dan kaitan kuat dengan setiap konsep GO. Sebanyak 3 tricluster hanya memiliki keterlibatan atau kaitan pada salah satu aspek GO dan 1 tricluster yang memiliki keterlibatan pada semua aspek GO namun hanya pada aspek celuller componet yang memiliki kaitan kuat. Hal ini dapat menjadi acuan bagi peneliti bidang biologi untuk memfokuskan penelitian lebih lanjut dalam pemahaman fungsi biologis pada himpunan tricluster yang memiliki keterlibatan dan kaitan kuat. <hr /> Triclustering analysis is a technique capable of clustering three-dimensional data simultaneously, thus obtaining subspaces of the 3D data consisting of

subsets of observations (genes), attribute subsets (conditions), and context subsets (time). The triclustering analysis conducted in this research utilizes the Î´-Trimax method through a two-way K-means approach. The goal of the Î´-Trimax method is to find triclusters that have minimum values of three-dimensional mean square residu MSR_3D and maximum volume. The two-way K-means approach is used to form an initial population to reduce computational burden and aid in forming better triclusters. This method will be implemented on gene expression data from HAE (Human Airway Epithelial) cultures infected with SARS-CoV, SARS-dORF6, SARS-BatSRBD, and H1N1 viruses. The implementation is carried out through 9 simulations, and the best simulation is obtained with a threshold value of Î´ calculated from MSR of 0.0435, a threshold value of Î»=1.7, resulting in 24 formed triclusters based on the triclustering quality index (TQI) assessment. From the set of triclusters, information regarding the comparison of gene expression patterns between SARS-CoV, SARS-dORF6, SARS-BatSRBD viruses and H1N1 influenza virus is obtained. There are 7 triclusters that exhibit similar gene expression patterns across all conditions, and 8 triclusters that are suspected to have condition differences between various SARS-CoV viruses and the H1N1 virus. Other triclusters also provide information where only certain SARS-CoV variations share similarities with each other or similarities or differences with H1N1. Based on the time points, 3 triclusters show no effect as their gene expression patterns remain the same as the initial condition (time point 1), while 17 triclusters are suspected to have post- infection effects. To assess the quality of the formed triclusters in terms of biological function representation of the gene sets within the triclusters, an evaluation of gene ontology (GO) is performed. GO is a system for describing the functions, biological processes, cellular components, and molecular functions of genes across various organisms. The evaluation method involves the Database for Annotation, Visualization, and Integrated Discovery (DAVID) in calculating p-values. The evaluation results reveal that 20 triclusters have strong involvement and correlation with each GO concept. Three triclusters only exhibit involvement or correlation in one specific aspect of GO, and one tricluster exhibits involvement in all GO aspects, but with a strong correlation only in the cellular component aspect. This information can serve as a reference for researchers in the field of biology to focus further research on understanding the biological functions within tricluster sets that have strong involvement and correlation.</p>