Analisis Pemilihan Molekul Inhibitor Dipeptidil Peptidase 4 pada Perancangan Obat Diabetes Tipe 2 menggunakan Algoritma K-Modes Clustering dengan Levenshtein Distance = Molecular Selection Analysis of Dipeptidyl Peptidase-4 Inhibitors in The Drug Discovery of Type 2 Diabetes using K-Modes Clustering Algorithm with Levenshtein Distance

Sarah Syarofina, author

Abstrak

Inhibitor dipeptidil peptidase 4 (DPP-4) baru perlu dikembangkan untuk meminimalkan efek samping merugikan yang diakibatkan oleh obat golongan inhibitor DPP-4 yang telah terdaftar. Penelitian ini bertujuan untuk menghasilkan subset molekul inhibitor DPP-4 yang representatif dengan mengaplikasikan algoritma K-Modes clustering dengan Levenshtein distance pada proses clustering dan melakukan analisis pemilihan molekul inhibitor DPP-4 berdasarkan kriteria nilai logP dari aturan Lipinskis Rule of 5. 2053 molekul inhibitor DPP-4 diperoleh dari situs ChEMBL. Clustering dilakukan terhadap fingerprint molekuler inhibitor DPP-4 yang diperoleh dari fitur SMILES (Simplified Molecular Input Line Entry System). Metode MACCS (Molecular Access System) Keys, ECFP (Extended Connectivity Fingerprint) diameter 4 dan 6, dan FCFP (Functional Class Fingerprint) diameter 4 dan 6, digunakan untuk membangun lima dataset fingerprint untuk proses clustering. Prosedur clustering diawali dengan menentukan jumlah klaster dengan menghitung nilai Koefisien Silhouette sebagai metode evaluasi klaster. Penerapan algoritma K-Modes clustering dengan Levenshtein distance pada 2053 molekul inhibitor DPP-4 menghasilkan nilai Koefisien Silhouette maksimal dari dataset MACCS sebesar 0.3947 dengan jumlah klaster 1258. Pemilihan molekul berdasarkan kriteria nilai logP dan aturan Lipinskis Rule of 5 menghasilkan 778 molekul inhibitor DPP-4 dari semua dataset dengan 298 molekul inaktif dan 480 molekul aktif dan nilai logP berkisar antara -1.67 sampai dengan 4.97. <hr />
New dipeptidyl peptidase 4 (DPP-4) inhibitors need to be developed to minimize the adverse side effects caused by registered DPP-4 inhibitor drugs. This study aims to produce a representative subset of DPP-4 inhibitor molecules by applying the K-Modes clustering algorithm with Levenshtein distance in the clustering process and analyzing the selection of DPP-4 inhibitor molecules based on the logP value criteria. 2053 DPP-4 inhibitor molecules obtained from the ChEMBL website. Clustering was carried out on the molecular fingerprint obtained from the SMILES feature. The MACCS Keys, ECFP (diameter 4 and 6), and FCFP (diameter 4 and 6) methods were used to construct fingerprint datasets for the clustering process. The clustering procedure begins by determining the number of clusters by calculating the Silhouette Coefficient value. The application of the K-Modes clustering with Levenshtein distance to 2053 DPP-4 inhibitor molecules resulted in the maximum Silhouette Coefficient value of the MACCS dataset of 0.3947 with the number of clusters 1258. Selection of molecules based on logP value criteria and Lipinskis Rule of 5 resulted in 778 DPP-4 inhibitor molecules. of all the datasets with 298 inactive

molecules and 480 active molecules and the logP value ranged from -1.67 to 4.97.</p>