INTISARI

SELEKSI FITUR BERBASIS ALGORITMA GENETIKA UNTUK PREDIKSI STRUKTUR SEKUNDER PROTEIN

FATHURROCHMAN HABIBIE
19/448703/PPA/05786

Protein merupakan makromolekul penting dalam struktur dan fungsi sebuah sel. Interaksi pada protein bertanggungjawab untuk mengontrol berbagai fungsi vital dalam tubuh, seperti mengaktifkan sistem kekebalan tubuh, mengatur oksigenasi, dan menentukan respon obat. Biasanya, struktur sekunder protein dapat ditentukan melalui metode eksperimental (misalnya, kristalografi sinar-X, NMR). Namun, metode eksperimental sangat mahal, memerlukan waktu yang lama, dan memerlukan prosedur yang kompleks. Oleh karena itu, pendekatan komputasi untuk memprediksi struktur sekunder protein penting dalam bidang biologi.

Struktur sekunder protein dapat ditentukan oleh urutan penyusun asam amino. Biasanya, prediksi struktur sekunder protein menggunakan dua fitur tetap: urutan asam amino dan profil PSSM. Namun, fitur protein tambahan lainnya (misalnya, biofisika, fisikokimia, skor konformasi) dapat meningkatkan akurasi prediksi struktur sekunder protein. Tesis ini berfokus pada seleksi fitur untuk meningkatkan akurasi dalam memprediksi struktur sekunder suatu protein. Dalam tesis ini, pertama-tama kami mengusulkan untuk menggunakan model CNN dan algoritma genetika untuk menemukan subset fitur yang optimal. Kemudian kami melatih model CNN-BLSTM menggunakan fitur yang dipilih dan mencapai akurasi Q8 74,5% pada dataset CB513.

Kata kunci: Seleksi Fitur, Algoritma Genetika, Prediksi Struktur Sekunder Protein, CNN, BLSTM

ABSTRACT

GA-BASED FEATURE SELECTION FOR PROTEIN SECONDARY STRUCTURE PREDICTION

FATHURROCHMAN HABIBIE
19/448703/PPA/05786

Proteins are essential macromolecules for the structure and function of a cell. Interactions of proteins are responsible for controlling various vital functions in the body, such as having a role in activating the immune system, regulating oxygenation, and determining drug response. Usually, the secondary structure of a protein can be determined through experimental methods (e.g., X-ray crystallography, NMR). However, the experimental methods are very expensive, time-consuming, and require a complex procedure. Hence, the computational approach for predicting the secondary structure of a protein is important in the biology field.

The secondary structure of a protein can be determined by its constituent sequence of amino acids. Usually, the protein secondary structure prediction uses two fixed features: amino acid sequences and PSSM profiles. However, other additional protein features (e.g., biophysical, physicochemical, conformation scores) can improve protein secondary structure prediction accuracy. This thesis focuses on feature selection to improve the accuracy in predicting the secondary structure of a protein. In this thesis, we first proposed to use a CNN model and a genetic algorithm to find the optimal subset of features. Then we trained a CNN-BLSTM model using the selected features and achieved 74.5% Q8 accuracy on the CB513 dataset.

Keywords: Feature Selection, Genetic Algorithm, Protein Secondary Structure Prediction, CNN, BLSTM