ABSTRAK

KLASIFIKASI DNA-*BINDING* PROTEIN MENGGUNAKAN ALGORITME *RANDOM FOREST*

Oleh

Ridho Alrafi

Mengklasifikasikan fungsi protein dan struktur dari urutan adalah salah satu tantangan penting untuk kombinasi biologi. Urutan protein masih susah untuk diklasifikasikan karena semakin beragamnya protein yang ditemukan, dan juga banyaknya metode yang bisa digunakan untuk menentukan urutan protein. Pada penelitian ini berfokus terhadap protein pengikat DNA. Protein pengikat DNA memainkan peran penting dalam beberapa bagian besar proses seluler. Oleh karena itu, perlu dilakukannya pengklasifikasian untuk mengidentifikasi protein pengikat DNA berdasarkan urutan protein menggunakan ekstrasi fitur dan random forest. Tujuan dari penelitian ini untuk menganalisa sensitivitas, spesifisitas, akurasi, dan matthew correlation coefficient dari protein pengikat DNA. Dataset protein pengikat DNA diperoleh dari PDB (Protein Data Bank) dengan mencari kata kunci "DNA-binding protein", memiliki jumlah protein 1075 dengan 525 data positif dan 550 data negatif serta memiliki panjang protein terpendek, yaitu 51 amino acids. Dataset dibagi menjadi 2, yaitu 80% data latih 20% data uji dan 90% data latih 10% data uji. Ekstraksi fitur yang digunakan protein descriptor menggunakan R package BioSeqClass versi 1.44.0, yaitu AAIndex, CTD, dan PseAAC, dengan jumlah total 440 fitur. Hasil yang didapatkan diolah kembali menggunakan klasifikasi algoritme random forest dengan mtry 10, 21, dan 42 lalu ntree dipilih secara acak 100, 250, 500, dan 1000. Hasil yang didapatkan paling tinggi didapatkan pada pembagian dataset 90% data latih 10% data uji dengan mtry 42 ntree 1000 sebesar 89.97% sensitivitas, 92.79% spesifisitas, 80.76% MCC, dan 90.42% akurasi. Hasil yang didapatkan menggunakan ekstraksi fitur dan algoritme random forest bisa mengklasifikasikan protein pengikat DNA.

Kata Kunci: Protein pengikat DNA, fitur ekstraksi, klasifikasi, random forest.

ABSTRACT

CLASSIFICATION DNA-BINDING PROTEIN USING RANDOM FOREST ALGORITHM

By

Ridho Alrafi

Classifiying the function and structure of proteins from sequences is one of the most important challenges for combination biology. Protein sequences are still difficult to classify because of the increasing variety of proteins found, as well as the many methods that can be used to determine protein sequences. In this research, we focus on DNA-binding proteins. DNA-binding proteins plays an important role in several major cellular processes. Therefore, it is necessary to classify to identify DNA-binding proteins based on protein sequences using feature extraction and random forest. The purpose of this research was to analyze the sensitivity, specificity, accuracy, and Matthew Correlation Coefficient of DNA-binding protein. The DNA-binding protein dataset was obtained from PDB (Protein Data Bank) by searching the keyword "DNA-binding protein", it has 1075 proteins with 525 positive data and 550 negative data and has the shortest protein length of 51 amino acids. The dataset separation into 2, 80% training data 20% test data and 90% training data 10% test data. Feature extraction used by protein descriptors using R package BioSeqClass version 1.44.0, using AAIndex, CTD, and PseAAC, with a total of 440 features. The results obtained were reprocessed using the random forest algorithm classification with mtry 10, 21, and 42 then ntree was selected at random 100, 250, 500 (default), and 1000. The highest results obtained in the separation of the dataset 90% training data 10% test data with mtry 42 ntree 1000, result 89.97% sensitivity, 92.79% specificity, 80.76% MCC, and 90.42% accuracy. The results obtained using feature extraction and random forest algorithms can classification of DNA binding proteins.

Keywords: DNA-binding protein, feature extraction, classification, random forest.