ABSTRAK

KLASIFIKASI GLIKOSILASI PADA SEQUENCE PROTEIN LISIN MENGGUNAKAN METODE *RANDOM FOREST*

Oleh

Silfia Fitriyana

Glikosilasi pada sequence protein lisin proses pengklompokkan protein lisin yang terglikosilasi ke dalam kelompok-kelompok berdasarkan jenis glikosilasi yang terdapat pasa protein lisin. Klasifikasi glikosilasi pada sequence protein lisin dengan menggunakan metode random forest, tujuan dari penelitian ini untuk menganalisa sensitivitas, spesifisitas, akurasi, dan *matthew correlation coefficient* (MCC). Dataset protein lisin diperoleh dari data benchmark memiliki jumlah protein 620 dengan 214 data positif dan 406 data negatif serta memiliki panjang protein 15 sequence. Dataset dibagi menjadi 2, yaitu 90% data latih dan 10% data uji. Ekstraksi fitur yang digunakan protein descriptor menggunakan R package BioSeqClass versi 1.44.0, yaitu AA *Index*, CTD, dan PseAAC, dengan jumlah total 60 fitur. Hasil yang didapatkan diolah kembali menggunakan klasifikasi algoritme random forest dengan mtry 4, 8, dan 16 lalu ntree dipilih secara acak 250, 500, 750, 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 lisin.

Kata Kunci : Glikosilasi, protein lisin, fitur ekstarksi, klasifikasi, random forest

ABSTRACT

GLYCOSYLATION CLASSIFICATION OF LYSINE PROTEIN SEQUENCE USING THE RANDOM FOREST METHOD

By

Silfia Fitriyana

Glycosylation in the lysine protein sequence is the process of grouping glycosylated lysine proteins into groups based on the type of glycosylation present in the lysine protein sequence. Classification of glycosylation in lysine protein sequences using the random forest method. The purpose of this study was to analyze the sensitivity, specificity, accuracy, and Matthew correlation coefficient (MCC). The lysine protein dataset obtained from benchmark data has a total protein of 620 with 214 positive data and 406 negative data and has a protein length of 15 sequences. The dataset is divided into 2, namely 90% training data and 10% test data. Feature extraction used protein descriptor using R package BioSeqClass version 1.44.0, namely AA Index, CTD, and PseAAC, with a total of 60 features. The results obtained were reprocessed using the random forest classification algorithm with mtry 4, 8, and 16, and then ntree was randomly selected 250, 500, 750, and 1000. The highest results obtained were obtained in the distribution of the dataset 90% training data 10% test data with mtry 42 ntree 1000 of 89.97% sensitivity, 92.79% specificity, 80.76% MCC, and 90.42% accuracy. The results obtained using feature extraction and random forest algorithms can classify lysine proteins.

Keywords: *Glycosylation, lysine protein, extraction features, classification, random forest*