

ABSTRAK

PREDIKSI GLIKOSILASI PADA N-, C- DAN O- DALAM PROTEOM MANUSIA MENGGUNAKAN FITUR SELEKSI MRMR DAN ALGORITME *SUPPORT VECTOR MACHINE.*

Oleh

NAURAH NAZHIFAH

Tubuh manusia mengandung ribuan protein. Pada proses pembentukan protein banyak mengalami modifikasi pasca translasi. Salah satu hasil dari modifikasi pasca translasi adalah glikosilasi. Glikosilasi adalah proses penggabungan glukosa pada struktur protein. Pada tubuh manusia glikosilasi dapat dilihat dari 3 kategori, yaitu N-glikosilasi, O-glikosilasi dan C-glikosilasi. Untuk memahami mekanisme dan peran fungsional glikosilasi, yaitu dengan cara memprediksi substrat dari situs glikosilasi tersebut. Pendekatan komputasi merupakan salah satu cara untuk memprediksi situs glikosilasi tersebut, yaitu menggunakan algoritme *Support Vector Machine* (SVM). Algoritme ini telah banyak digunakan untuk prediksi dan pengklasifikasian. Pada penelitian ini menggunakan 2 jenis data yaitu data *Independent* dan data *Benchmark*. Fitur yang digunakan merupakan fitur ekstraksi yang menghasilkan 90 dimensi dan fitur seleksi yang menggunakan *Maximum Redundancy Minimum Relevance* (MRMR) sebanyak 25, 50 dan 75 kolom. Pengujian klasifikasi SVM menggunakan *5-fold cross validation* dan *confusion matrix*. Hasil akurasi tertinggi terletak pada penggunaan fitur seleksi MRMR sebanyak 75 kolom. Pada Data *Independent* N akurasi terbesar sebesar 86,66% pada kernel *Sigmoid*, sedangkan untuk data *Independent* C akurasi sebesar 87,5% pada kernel *Sigmoid* dan untuk data *Independent* O akurasi sebesar 89,31% berada di kernel RBF. Pada data *Benchmark* N akurasi terbesar sebesar 70,54% pada kernel RBF, sedangkan untuk data *Benchmark* C akurasi terbesar sebesar 95,06% dan untuk data *Benchmark* O terdapat di kernel RBF dengan akurasi terbesar yaitu 92,64%.

Kata Kunci: Glikosilasi, MRMR, *Post Translation Modification*, *Support Vector Machine*.

ABSTRACT

PREDICTION OF GLYCOSYLATION IN N-, C- AND O- IN HUMAN PROTEOM USING MRMR SELECTION FEATURES AND SUPPORT VECTOR MACHINE ALGORITHM.

By

NAURAH NAZHIFAH

The human body contains thousands of proteins. In the process of protein formation, there are many post-translational modifications. One result of post-translational modification is glycosylation. Glycosylation is the process of combining glucose in protein structures. In the human body, glycosylation can be seen from 3 categories, namely N-glycosylation, O-glycosylation and C-glycosylation. To understand the mechanism and functional role of glycosylation by predicting the substrate of the glycosylation site. The computational approach is one way to predict the glycosylation site, using the Support Vector Machine (SVM) algorithm. This algorithm has been widely used for prediction and classification. This study uses 2 types of data, namely Independent data and Benchmark data. The features used are feature extraction which produces 90 dimensions and feature selection using Maximum Redundancy Minimum Relevance (MRMR) of 25, 50 and 75 columns. The SVM classification test uses 5-fold cross validation and confusion matrix. The highest accuracy result lies in the use of the 75 column MRMR selection feature. In Independent data N, the highest accuracy is 86.66% in the *Sigmoid* kernel, while for Independent data C, the accuracy is 87.5% in the *Sigmoid* kernel and for Independent data O, the accuracy is 89.31% in the RBF kernel. In Benchmark data N, the highest accuracy is 70.54% in the RBF kernel, while for Benchmark data C the highest accuracy is 95.06% and for Benchmark data O is in the RBF kernel with the largest accuracy, which is 92.64%.

Keywords: Glycosylation, MRMR, Post Translation Modification, Support Vector Machine