ABSTRAK

KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Oleh

Suci Hikmawati

Modifikasi protein merupakan protein yang strukturnya mengalami perubahan. Glikosilasi adalah salah satu modifikasi protein yang paling rumit dalam sel eukariotik. Glikosilasi berperan penting dalam fungsi biologis, seperti pengenalan antigen, komunikasi sel-sel, ekspresi gen, dan pelipatan protein. Dalam melakukan identifikasi situs glikosilasi pada sekuens protein dapat dilakukan secara eksperimental, namun merupakan tantangan yang sulit karena memerlukan waktu yang lama dan biaya yang mahal. Beberapa peneliti mengusulkan menggunakan metode komputasi untuk identifikasi situs glikosilasi, sehingga pada penelitian ini dilakukan pengklasifikasian untuk mengidentifikasi situs glikosilasi pada sekuens protein dengan menggunakan metode pendekatan *machine learning* yaitu algoritma *random forest.* Data yang digunakan adalah situs n glikosilasi yang terdiri dari data positif berjumlah 11.601 sekuens dan data negatif 12.160 sekuens. Tahapan yang dilakukan pada penelitian ini, diantaranya : cleaning data yaitu menghapus sekuens yang tidak termasuk ke dalam jenis asam amino, fitur yang berkaitan dengan situs modifikasi pasca translasi di ekstraksi, fitur tersebut terdiri dari *statistical moment* dan fitur *position and composition* dengan total fitur berjumlah 153, pemodelan random forest, hingga pengujian klasifikasi menggunakan 10-fold cross validation dan confusion matrix. Setelah dilakukan beberapa percobaan, diperoleh hasil pengujian klasifikasi random forest dengan akurasi tertinggi sebesar 92,29%, sensitivitas 97,71%, spesifisitas 87,12%, serta MCC sebesar 85,12%.

Kata kunci: klasifikasi, machine learning, glikosilasi, ekstraksi fitur, *random forest.*

ABSTRACT

CLASSIFICATION OF POST-TRANSLATIONAL MODIFICATIONS N-GLYCOSYLATION ON PROTEIN SEQUENCES USING RANDOM FOREST ALGORITHM

By

Suci Hikmawati

Modified protein is a protein whose structure has changed. Glycosylation is one of the most complex protein modifications in eukaryotic cells. Glycosylation plays an important role in biological functions, such as antigen recognition, cell-cell communication, gene expression, and protein folding. In identifying glycosylation sites on protein sequences can be done experimentally, but it is a difficult challenge because it requires a long time and is expensive. Several researchers have proposed using computational methods to identify glycosylation sites, so that in this study a classifier was carried out to identify glycosylation sites in protein sequences using a machine learning approach, namely the random forest algorithm. The data used is site n glycosylation consisting of positive data totaling 11,601 sequences and negative data 12,160 sequences. The stages carried out in this study, including: cleaning data, namely: delete sequences that do not belong to this type of amino acid, features related to post-translation modification sites are extracted, these features consist of statistical moment and position and composition features with a total of 153 features, random forest modeling, to classification testing using 10-fold cross validation and confusion matrix. After several experiments, the results of the random forest classification test were obtained with the highest accuracy of 92.29%, sensitivity of 97.71%, specificity of 87.12%, and MCC of 85.12%.

Keywords: classification, machine learning, glycosylation, feature extraction, random forest.