ABSTRACT

METHYLATION PREDICTION IN ARGININE PROTEIN SEQUENCE USING RANDOM FOREST

BY

WIWIT MUDYANINGSIH

One of the popular Post-Translational Modification (PTM) is methylation. Methylation can occur in the amino acid arginine. Arginine methylation carries out and regulates several important biological functions, including gene regulation and signal transduction. Experimental identification of arginine methylation sites is a daunting task because it is expensive and requires more time and energy. Therefore reliable predictions play an important task in predicting quickly and identifying possible methylation sites in proteomes. In the study using the Random Forest method which is one of the data mining techniques for classification, and previously performed feature extraction, feature extraction used is CTD, PseAAc, AA index and QSO by using the protr and bioseqclass packages in the r programming language. and obtained 138 variables. This study has 3 types of experiments, training set in 1:1 ratio, Testing data set and Independent data set. The prediction perform reasonably well with the highest accuracy obtained in the Independent Data experiment that is equal to 98.08%, while for Training Data the ratio of 1: 1 gets 93.76% and Testing gets the lowest accuracy that is equal to 80.32%.

Keywords: *Methylation, Random Forest, Prediction, Arginine, Feature Extraction, Machine Learning*

ABSTRAK

PREDIKSI METILASI PADA *SEQUENCE* PROTEIN *ARGININE* MENGGUNAKAN RANDOM FOREST

OLEH

WIWIT MUDYANINGSIH

Salah satu *Post-Translational Modification* (PTM) yang populer adalah metilasi**.** Metilasi dapat terjadi dalam asam amino arginine**.** Metilasi arginin melakukan dan mengatur beberapa fungsi biologis penting, termasuk regulasi gen dan transduksi sinyal. Identifikasi eksperimental situs *arginine* metilasi adalah tugas yang berat dikarenakan mahal serta memerlukan waktu dan tenaga yang lebih. Oleh karena itu prediksi yang handal memainkan tugas penting dalam memprediksi dengan cepat dan mengidentifikasi kemungkinan situs metilasi di proteomes. Pada penelitian menggunakan metode Random forest yang merupakan salah satu teknik data mining untuk melakukan klasifikasi, serta sebelumnya dilakukan *feature extraction*, *feature extraction* yang digunakan yaitu CTD, PseAAc, AA index dan QSO dengan menggunakan *package* protr dan *bioseqclass* pada bahasa pemrograman R Programming., dan didapatkan 138 variabel. Penelitian ini memiliki 3 jenis eksperimen yaitu data Rasio 1:1, *Testing* dan *Independent* dataset. Prediksi ini berkinerja yang cukup baik dengan didapatkan akurasi tertinggi pada percobaan Data *Independent* yaitu sebesar 98,08 %, sedangkan untuk data Training Rasio 1:1 mendapatkan 93,76 % dan *Testing* mendapat akurasi terendah yaitu sebesar 80,32 %.

Kata kunci: Metilasi**,** Random Forest, Prediksi, *Arginine*, Ekstraksi Fitur, *Machine Learning*