

ABSTRAK

KLASIFIKASI SITUS *UBIQUITINATION* PADA *SEQUENCE PROTEIN* DENGAN ALGORITME *EXTREME GRADIENT BOOSTING*

Oleh

SYELA SEPTANIA

ubiquitination protein merupakan salah satu proses modifikasi pasca translasi (PTM). Identifikasi situs *ubiquitination* merupakan salah satu proses penting karna banyaknya peran yang dimiliki oleh *ubiquitination*. Identifikasi ini tidak hanya membantu untuk memahami mekanisme molekulernya tetapi juga memberikan fakta berharga untuk studi tambahan tentang pengembangan obat karena peran pengaturannya yang kritis. Oleh sebab itu perlu dilakukannya proses klasifikasi untuk identifikasi situs *ubiquitination*. Penelitian ini melakukan proses klasifikasi pada *sequence* protein menggunakan metode *Statistical moment* untuk ekstraksi fitur dan metode *eXtrame Gradient Boosting*(XGBOOST) untuk algoritma klasifikasi. Data yang digunakan merupakan dataset benchmark berupa *sequence* protein yang diakses dari UniProt dan dibuat oleh BMC *Bioinformatics* tahun 2016. Dataset ini terdiri dari dataset 1, dataset 2 dan dataset 3. dataset 1 berjumlah 300 *sequence* protein yang mencakup 150, dataset 2 terdapat 6838 *sequence* protein dan Dataset 3 berisi 12236 *sequence* protein dimana masing-masing setengah dari dataset tersebut adalah data positif dan setengahnya lagi merupakan data negatif. Hasil tertinggi didapatkan pada dataset 1 menggunakan metode pembagian data *10-fold cross-validation* dengan 90 % data *training* dan 10 % data *testing*, yaitu 96,59 % akurasi., 100 % sensitivitas, 93,24 % spesifisitas, dan 93,42 % MCC.

Kata Kunci : Klasifikasi, *Ubiquitination*, *Feature Extraction*, *eXtrame Gradient Boosting* (XGBoost)

ABSTRACT

CLASSIFICATION OF UBIQUITINATION SITES ON PROTEIN SEQUENCES WITH EXTREME GRADIENT BOOSTING ALGORITHM

By

SYELA SEPTANIA

Ubiquitination protein is one of the post-translational modifications (PTM) process. Identification of ubiquitination sites is one of the important processes due to the many roles that ubiquitination has. This identification not only helps to understand its molecular mechanisms but also provides valuable facts for additional studies on drug development due to its critical regulatory role. Therefore, it is necessary to carry out a classification process for identification of ubiquitination sites. This study showed a classification process on protein sequences using the Statistical moment method for feature extraction and the eXtreme Gradient Boosting (XGBOOST) method for the classification algorithm. The data used was a benchmark dataset in the form of protein sequences accessed from UniProt and made by BMC Bioinformatics in 2016. This dataset consisted of dataset 1, dataset 2 and dataset 3. dataset 1 contained 300 protein sequences which included 150, dataset 2 contained 6838 protein sequences and dataset 3 contained 12236 protein sequences where each half of the dataset was positive data and the other half was negative data. The highest results were obtained in dataset 1 using the 10-fold cross-validation data sharing method with 90% training data and 10% testing data, namely 96.59% accuracy, 100% sensitivity, 93.24% specificity, and 93.42% MCC.

Keywords:: Classification, Ubiquitination, Feature Extraction, eXtreme Gradient Boosting (XGBoost)