ABSTRACT

Dikson (00000004809)

UTILIZATION OF MACHINE LEARNING FOR ESTABLISHING MODEL OF ENDOXYLANASE OPTIMAL ACTIVITY
Thesis, Faculty of Science and Technology (2018).

(xiii + 35 pages; 6 tables; 9 figures; 4 appendices)

Endoxylanase is a type of enzyme belonging to Glycosyl Hydrolase 11 family capable of hydrolyzing 1,4-β-D-xyloside bonds in between xylose residues within xylan chain. Endoxylanase has been studied for its properties and optimal activities, as well as applied for various kinds of industrial purposes. In search for new varieties of endoxylanase, their optimal activities are studied by conventional means in wetlab, which may require large amounts of time and funding. Machine learning is offered as a cheaper, faster alternative of predicting endoxylanase optimal activity based on existing data from UniProtKB and BRENDA, using logistic regression algorithm as it suits biological data like enzyme sequences for classification. Pre-optimization model yields 67% accuracy for optimal temperature and 93% for kingdom while post-optimization model yields 73% accuracy for optimal temperature and 93% for kingdom. However, cross-validation indicates a wide variation of accuracy yielded by said model, which suggests the model to be overfitting. This was possibly caused by limited amounts of data, or the data remain to be too complex to establish a general model. Further actions needed to refine the algorithm and data processing, even if machine learning is proven capable of analyzing endoxylanase amino acid sequences and predicting their optimal activities.

Keywords: enzyme, endoxylanase, machine learning.

References: 74 (1959 - 2018).

ABSTRAK

Dikson (00000004809)

PENGGUNAAN PEMBELAJARAN MESIN DALAM PEMBANGUNAN MODEL AKTIVITAS OPTIMAL ENZIM ENDOXILANASE
Tugas Akhir, Fakultas Sains dan Teknologi (2018).

(xiii + 35 halaman; 6 tabel; 9 gambar; 4 lampiran)

Endoxilanase merupakan enzim yang berasal dari keluarga Glikosil Hidrolase 11 yang mampu menghidrolisis ikatan 1,4-β-xilosida diantara residu xilosa dalam rantai xilan. Endoxilanase telah dipelajari akan sifat dan aktivitas optimal, dan diterapkan dalam beragam jenis kebutuhan industri. Dalam pencarian varietas endoxilanase baru, aktivitas optimal enzim dicari secara konvensional di laboratorium basah, yang mungkin memerlukan waktu dan dana yang besar. Pembelajaran mesin ditawarkan sebagai alternatif yang murah dan cepat untuk memprediksikan aktivitas optimal endoxilanase berdasarkan dari data pada basis data UniProtKB dan BRENDA, menggunakan algoritma regresi logistik yang dapat digunakan untuk data biologis seperti urutan asam amino enzim untuk klasifikasi. Model pra-optimisasi menghasilkan tingkat akurasi 67% untuk suhu optimal dan 93% untuk kingdom, sementara model pasca-optimisasi menghasilkan tingkat akurasi 73% untuk suhu optimal dan 93% untuk kingdom. Namun demikian, hasil validasi silang menandakan variasi yang luas pada tingkat akurasi yang dihasilkan oleh model tersebut, yang menandakan model mengalami *overfitting*. Hal tersebut diduga diakibatkan oleh jumlah data yang terbatas, ataupun data tersebut masih terlalu rumit untuk algoritma untuk membangun model yang umum. Tindakan lebih lanjut perlu dilakukan untuk memperbaiki algoritma dan pengolahan data, walaupun pembelajaran mesin terbukti mampu untuk menganalisis urutan asam amino enzim endoxilanase dan memprediksikan aktivitas optimalnya.

Kata kunci: enzim, endoxilanase, pembelajaran mesin.

Referensi: 74 (1959 - 2018).