ABSTRACT

A basic question in protein science is to which extent mutations affect protein thermostability or not. This knowledge would be particularly relevant for engineering thermostable enzymes. In several experimental approaches, this issue has been serendipitously addressed. The big problem is, this manual experimental will waste high budget because it needs triall and error manual experimental procedure. Triall and error prosedural on the laboratory are used to find the best fits prosedural that will use to make the experimental succes, but needs more time and more budget. It would be therefore convenient providing a computational method that predicts when a given protein mutant is more thermostable than its corresponding wild-type. The purpose of this research is to compare the accuracy data from SVM with the data from experimental manual that was done by Annaluru(2002) and Salazar(2007). SVM (Support Vector Machine) is a machine learning methode that will used in this research as a predictor to predict the enzyme thermostability. BLAST (Basic Local Alignment Tools) are used as a data mining methode to extract data from 14 family protein (Annaluru and Salazar data experimental). From the outputs of the BLAST runs only aligned sequence pairs comprising a mesophilic and a thermophilic protein were selected. Only protein pairs sharing at least 70% of sequence identity were retained (a redundancy reduced dataset). we used SVM L20 (difference of the residue composition in each pair of the dataset) and SVM L400 (difference of the dipeptide composition in each pair of the dataset). When trained and tested on a redundancy reduced dataset (homology >70%), our predictor (SVM) achieves 86% accuracy for SVM L20 (classifies 12 out of 14 experimentally characterized protein mutants with enhanced thermostability from Annaluru and Salazar) .and 79% for SVM L400 (classifies 11 out of 14 experimentally characterized protein mutants with enhanced thermostability from Annaluru and Salazar). This accuracy from SVM are one of the reasonable way to used SVM as a methode to reduce the triall and error procedural because can give the prediction and reduce the failled experimental.

ABSTRAK

Salah satu hal mendasar di dunia science mengenai enzim adalah, penelitian untuk memprediksi pengaruh perpanjangan mutasi terhadap thermostabilitas enzim yang seringkali berkaitan dengan penelitian rekayasa enzime thermostabil. Masalah utama muncul ketika landasan untuk melakukan percobaan secara manual di laboratorium, prosedural pengerjaannya selalu dilakukan secara trial and error. Dana dan waktu yang terbuang untuk mendukung percobaan ini cukup besar, mengingat percobaan masih dilakukan secara manual dan juga faktor keberhasilan yang belum akurat. Hal ini yang mendasari diperlukannya suatu campur tangan secara komputerisasi untuk memberikan prediksi mengenai thermostabil enzim mutant yang diperbandingkan dengan *wild type*-nya(sebelum dimutasi). Tujuan dari penelitian ini adalah untuk memperbandingkan keakurasian data yang diperoleh dari SVM terhadap data experimental yang dilakukan secara manual oleh Annaluru(2002) dan Salazar(2007). SVM (*Support Vector Machine*) adalah suatu metode *Machine Learning* yang digunakan dalam penelitian ini untuk memberikan prediksi stabilitas enzime termutasi. Data set yang digunakan untuk SVM berasal dari data yang di hasilkan sebagai output dari BLAST (*Basic Local Alignment Tools*). BLAST digunakan sebagai metode untuk me-mining data dari ke-14 family protein (Annaluru dan Salazar experimental data). Mining data dari ke-10 protein ditujukan untuk mendapatkan kumpulan protein mutant yang memiliki tingkat homologi >70% dari sekumpulan besar data 10 family protein yang belum jelas tingkat homologinya. Proses SVM dilakukan dua kali, yaitu SVM L20 Untuk residu asam amino yang berbeda dalam setiap pasang data *sequence* dalam data set dan SVM L400 Untuk *sequence* dengan komposisi dipeptida yang berbeda. Hasil akhir, SVM memberikan tingkat keakurasian yang tidak jauh berbeda, yaitu sebesar 86% untuk L20 (SVM berhasil mengidentifikasi 12 dari 14 jenis protein) dan 79% untuk L400 (SVM berhasil mengidentifikasi 11 dari 14 jenis protein). keakurasian SVM yang hampir memprediksi dengan sempurna mengenai thermostabilitas enzime termutasi, dapat digunakan sebagai metode untuk meminimalisir kesalahan dan waktu yang terbuang dalam trial dan error prosedural experimental.