Identification and analysis of thermophilic protein sequences in compost sequencing data

Amanda Rodrigues[1], Aline Maria da Silva[1], João Carlos Setubal[1]

1 USP

Abstract

Typical aerobic composting is a self-heating process in which microbial metabolism drives the temperature above 50?^{o}C, followed by sustained high temperatures between 60-80?^{o}C, and then followed by gradual cooling of the compost pile. Composting is widely considered as a promising source of novel thermostable enzymes, particularly those related to biomass degradation. Protein thermostability is an important aspect of protein biochemical and biotechnological research. Commonly, for chemical reactions, high temperature could increase reaction activity and decrease reaction time. Various mechanisms of thermostability were discussed in the literature, and many authors pointed to changes in amino acid composition as one of clearest manifestations of thermal adaptation. Indeed, it is well-known that enhanced thermostability is reflected in specific trends in amino acid composition. In this sense, it is important to develop a validated method that can predict the stability of a given protein from its primary sequence. In this work, machine learning techniques combined with amino acid composition and dipeptide composition were used to discriminate between thermophilic and non-thermophilic proteins from compost sequencing data. In order to obtain a reliable dataset, three different techniques (artificial neural network, random forest and SVM) were combined to discriminate the proteins sequences. The combination of these methodologies has achieved a specificity of 97% for the test dataset composed of thermophilic and non-thermophilic sequences with experimental evidence extracted from Uniprot. From the sequences classified as thermophilic, sequences related to the degradation of biomass will be selected for phylogenetic and completeness analyzes. Interesting sequences will be synthesized and heterologously cloned.

Funding: CAPES, Fapesp