SPEdit: A sequence repository and a random forest classification based analysis platform for industrially relevant thermophilic serine protease

Lilly M. Saleena, Jithin S. Sunny

Department of Biotechnology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, 603203, Kanchipuram, Chennai TN, India

ABSTRACT

Thermophilic enzymes are functionally active at high temperatures which make them suitable for various industrial applications. Amongst them, serine proteases are one of the most widely used enzymes. Although successful protein engineering of this enzyme is being reported, increasing protein sequence information can aid in better rational protein designing endeavours. SPEdit is a web based platform that acts both as a data repository for serine proteases and also aid the analysis of user-provided sequence for the same. As a repository, SPEdit includes enzymes extracted from over 170 thermophilic bacterial genomes that are currently available from NCBI. Additionally, it also houses sequences retrieved from extensive literature survey. Together, SPEdit represents the largest collection of thermophilic serine proteases. It also houses these sequences in the form of a database which enable the users to run blastp against their proteins. This can be particularly useful in identifying novel serine proteases. Further, this platform provides users with a trained model for predicting the thermophilic nature of their serine protease sequences. The classification model was built by employing random forest classifier using all the available serine protease (430) that are reported above. Amino acid percentage and the isoelectric point of every sequence were used as features. Using this data, a classifier was built that can characterize the input serine protease as thermophilic or non-thermophilic. The present accuracy was evaluated to be 0.87. The trained model can be downloaded by the user.

Keywords: Serine protease, Web tool, Thermophilic, BLAST, Random Forest

ISBN: 978-81-947843-4-0; DOI: 10.21467/abstracts.109