Big Data Technology

IS 6733

Group #6 Project

Protein Sequence Analysis

The data is derived from UniProt website (<http://www.uniprot.org/>) which provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

The full data-set was downloaded as an XML file. The entire file was parsed and variables of interest were imported into a CSV file.

The data-set has identified unique amino acid sequences that have certain functions referred to as features.

The goal of our analysis is to build a classification model that will locate these features, and predict their presence in a new protein sequence.

Team Members

Matt Angotti

John Figueroa

Sean Palen

Stacey Walls

Scott Wynd