Exploring Machine Learning Algorithms to predict protein subcellular localization

Pedro Martins[1], Luka A Clarke[2], Hugo Botelho[2], Margarida D Amaral[2], Francisco M Couto[1]

[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2]University of Lisboa, Faculty of Sciences, BioISI - Biosystems & Integrative Sciences Institute, Lisboa, Portugal

Genes encode proteins which are located to various organelles inside cells in order to perform crucial functions: nucleus, endoplasmic reticulum, mitochondria, etc. To predict the subcellular localization of a given gene product (i.e., where the protein codified by the gene is located) is a particularly helpful information to be included in its functional annotation. Since proteins located in particular intracellular compartments share certain common features, Machine Learning (ML) algorithms are useful for the purpose of class prediction.

The goal of this study was to predict and assign one of the nineteen subcellular locations to 726 human genes involved in CFTR (cystic fibrosis transmembrane conductance regulator) traffic, a protein which cause the genetic disease Cystic Fibrosis, when mutated.

Using the Ensembl IDs of these genes, it was possible to create a dataset with a variety of gene-level and protein-level information extracted from the Ensembl and SWISS-PROT databases. Thus, the training set was constituted by the following features: chromosome number, GC content, GO Terms, amino acid composition, protein length, and secondary structure information; as well as the localization, which is the target for prediction (for proteins with multiple location annotations, we chose one based on the number of times each localization is annotated to that protein).

The ML methods used in this study were: NaiveBayes, BayesNet, SMO, PART and J48 classifiers. The ten-fold cross validation was used to train and test classifiers.

The Bayesian methods NaiveBayes and BayesNet obtained an accuracy of 61% and 65%, respectively, while J48 (decision trees) and PART (classification rules) methods obtained

an accuracy of 70%. Finally, the SMO (support vector machines) algorithm got the highest accuracy, correctly classifying 78% of instances.

Work supported by UID/MULTI/04046/2013 centre grant (to BioISI)

Keywords: Subcellular Location, Machine Learning, NaiveBayes classifier, BayesNet classifier, J48 classifier, PART classifier, SMO classifier.

Preference for presentation: Poster
Location: University of Minho
Author for Correspondence: pedroalmeidamartins@gmail.com