A machine learning approach to brain region classification

Lissur Azevedo Orsine, Adriano Barbosa da Silva

Laboratório de Biodados, Universidade Federal de Minas Gerais

Abstract

The knowledge of the similarities and particularities of brain regions contributes to understand the hemisphere-specific and holistic organ's biology. The brain can be studied based on different criteria, such as: electrical activity, connectivity, cell types and evolutionary origin. The emergence of RNA-Seq made also possible to characterize the brain according to patterns of gene expression. In this context, we asked ourselves the following question: is it possible to discriminate brain regions based on gene expression? To try to answer this question, we applied a machine learning methodology to the transcriptome of different brain regions. RNA-Seq experiments of two distinct brains containing 22, 318 genes over 10 distinct brain regions were downloaded from the Allen Brain Atlas. The anatomical structures were: frontal lobe (FL), parietal lobe (PL), occipital lobe (OL), temporal lobe (TL), insula (Ins), cingulate gyrus (CgG), parahippocampal gyrus (PHG), striatum (Str), globus pallidus (GP) and cerebellar cortex (CbCx). For the classification, eight machine learning algorithms were applied against the test dataset: Logistic Regression (LR), K-Neighbor Classifier (KNN), Gaussian NB (NB), SVC, Linear SVC (LSVC), Random Forest Classifier (RFC), Decision Tree Regressor (DTR) and Gradient Increase Classifier (GBRT). The best performance algorithm was selected according to 3-fold cross-validation training accuracy score and therefore used in the subsequent analyses. The most satisfactory performance classifier was LR with accuracy of 0.6 (against 0.4 from KNN, 0.5 from NB, 0.1 from SVC, 0.5 from SVC, 0.5 from LSVC , 0.5 from RFC, 0.1 from DTR, and 0.2 from GBRT). The prediction results of the confusion matrix showed that Str was correctly classified in 100% of cases, while OL in approximately 90% of cases, and FL, PL and TL around 30% of cases. It was also possible to observe that PL and TL present a high percentage of reciprocal exchange: around 20% of PL samples were predicted to belong to TL, and approximately 41% of TL samples were classified as PL. Looking at the F1 score, the best ranked brain region was again Str (F1 score = 1.0), followed by FL (0.53), OL (0.42), PL (0.40) and TL (0.30). Finally, the area under the ROC curve ranged from 0.89 (for PL) to 1.0 (for Str). As perspectives, we plan to run the same pipeline for the higher anatomical levels, as well as to compare gene lists per brain region found through the machine learning approach with those obtained from other methodologies.

Funding: CAPES Biologia Computacional