Transductive SVM Approach to Identifying Crohn’s Disease from Exome Data

To construct a suitable feature set for classification, an initial assessment of each variant sequence element was carried out against the CAGI2011 data. The ability of each variant element to distinguish healthy control from Crohn’s patient exomes was evaluated by calculating a Fisher exact test. A p-value threshold of 5x10-6 was applied, giving a total of 102 features for classification. In addition to labelled CAGI2011 data, CAGI2015 data was included in the training set as unlabelled data, and a label spreading model used for transductive learning with an SVM (RBF kernel).

Onset data was analysed in a similar way, though in this case a point biserial test was used to identify variant sequence features and onset age, giving a total of 57 selected features. Onset age was predicted using a linear model trained with Lasso regression.

Data handling was done using Perl scripts and machine learning using Python/scikit-learn.