CSCD 429/529 Data Mining Lab 4 (20 points)

Use **Orange** to predict gene localization

1. **Data Set Description:** This dataset was used in the 2001 kdd cup data mining competition. (http://www.cs.wisc.edu/~dpage/kddcup2001/). There were in fact two tasks in the competition with this dataset, the prediction of the "Function" attribute, and prediction of the "Localization" attribute. **Here we focus on the prediction of "Localization"**. The dataset provides a variety of details about the several genes of one particular type of organism. The main dataset (*Genes_relation.data* and *Genes_related.test*) contains row data of the following form:

   Gene ID, Essential, Class, Complex, Phenotype, Motif, Chromosome Number, Function, ***Localization****.*

   The description of data attributes was given in file *Genes_relation.names*. The first attribute is a discrete variable corresponding to the gene ID (there are 1243 gene values). The remaining 8 attributes consist of discrete variables, most of them related to the proteins coded by the gene, e.g. the "Function" attribute describes some crucial functions the respective protein is involved in, and the "Localization" is simply the part of the cell where the protein is localized.

   In addition to the data of the above form, there are also data files (*Interactions_relations.data* and *Interactions_relation.test*) which contain information about interactions between pairs of genes. **To simplify your work, you may not use the interactions data**.

   The keys are provided in the file keys.txt. Use this file to evaluate the accuracy of your solution.

2. **Task**:

   - Create a new process named **LocalizationPrediction**.
   - Load in the training data.
   - Preprocess the data if necessary.
   - Build a predictive model to predict on attribute "Localization".
   - Apply the model to the test data.
   - Find the performance of the model.

Submission:

You need to submit your report online which should include:

- (5 points) the diagram of the entire process;
- (10 points) the explanations of operators used in building the process;
- (5 points) the performance matrix. **The overall accuracy generated by your solution is expected to be higher than 60%.**