Revanth Korrapolu

Documentation

**1. Download two datasets from the Uniprot database (**[**https://www.uniprot.org**](https://www.uniprot.org/)**), (a) all eukaryotic protein sequences and (b) all prokaryotic sequences**

to download all eukaryotic protein sequences:

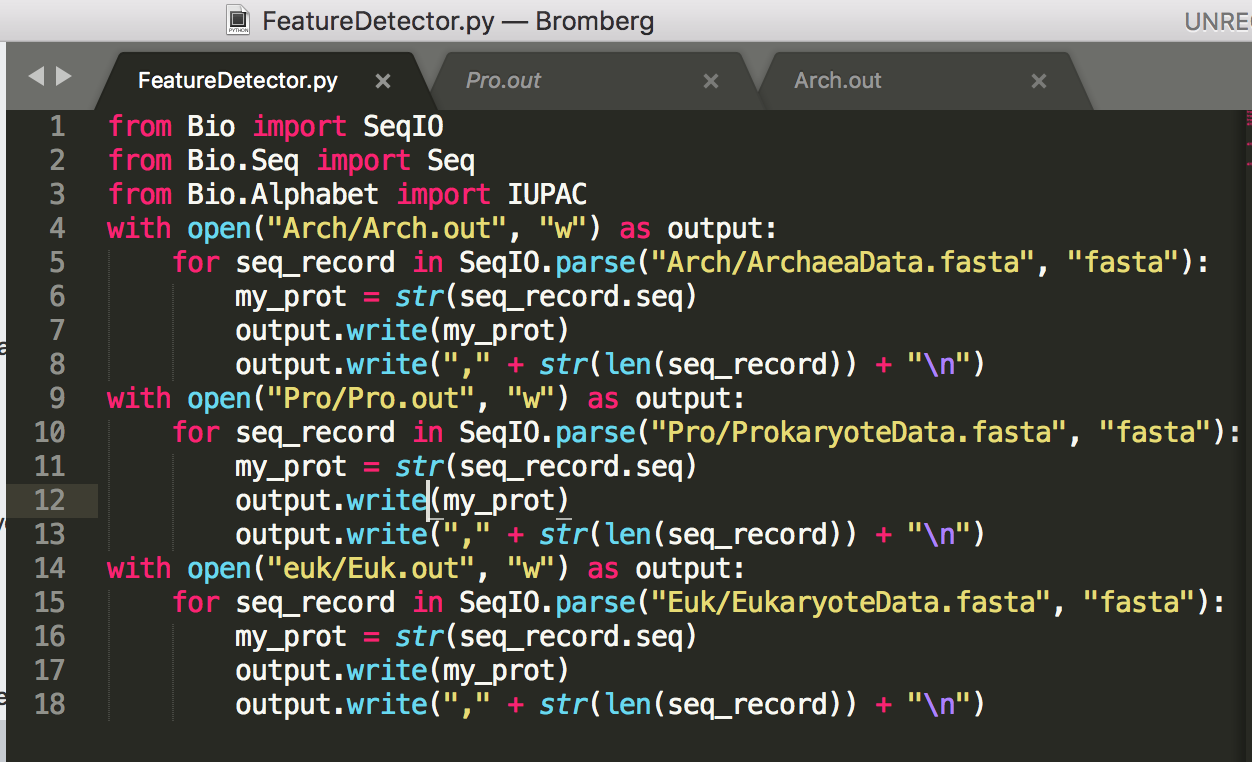
uniparc proteome:(taxonomy:"Eukaryota [2759]")

to download all prokaryotic protein sequences:

uniparc proteome:(taxonomy:"Eukaryota [2759]")

**2. Compute two features for each of those sequences: (1) length and (2) amino acid composition**

Created a simple script using biopython to parse for amino acid composition and sequence length. (See FeatureDetector.py)

****

**3. Create a training set containing all features for both sets of sequences and a class label (0: prokaryotic, 1: eukaryotic)**

Yes: <https://github.com/RevanthK/ML-Protein-Classifier>

The prokaryotic training set is in Pro/Pro.csv

The Archaea training set is in Arch/Arch.csv

The eukaryotic training set is in Euk/Euk.csv

\*the csv’s titled: “Arch\_10.csv” just have 10% of the training data

**4. User weka to load this training set and**

1. **create a pipeline which uses 10-fold cross validation to train a model (appropriate for the input features)**

yes; but struggling to make it scalable. The Weka java applet crashed when using the full training set.

**b) the pipeline should use a balancing filter to make sure the training dataset has an equal amount of class instances**

Using supervised class balancer filter

**c) report model performance**

**d) apply parameter optimization**