Practical 2

13 August 2024

Discriminate α-helical and β-barrel membrane proteins using machine learning techniques

Instructions

1. Upload your report in pdf format with the naming "roll no_p2.pdf".

2. Do not copy, it will be checked for plagiarism.

Steps:

1. Install Weka (https://waikato.github.io/weka-wiki/downloading_weka/)
2. Prepare the input file in .arff format using amino acid composition (see example)
3. Open the file using Weka
4. Classify them using different machine learning techniques (select any 10 from different classes/sub-classes)
5. Use training set and compute sensitivity, specificity and accuracy from the confusion matrix.
6. Repeat with 5-fold, 10-fold, 20-fold and 66% split cross-validations. Select the best method from the performance.
7. Keep 70%, 60% and 50% of the data as training and use others as test set to evaluate the performance of the best method.
8. Evaluate the importance of each residue by eliminating each residue (repeat 20 times). See the decrease in performance.
9. Analyze the same using only one residue at the time.
10. Construct a balanced dataset (243 each for alpha and beta) and obtain the results with 5-fold cross validation.
11. Tabulate and discuss the results.

Example weka file:

@RELATION trans_membrane

@ATTRIBUTE A numeric

@ATTRIBUTE G numeric

@ATTRIBUTE V numeric

@ATTRIBUTE L numeric

@ATTRIBUTE S numeric

@ATTRIBUTE class {alpha,beta}

@DATA

1, 0.76, 0.45, 0.5, 0.81, alpha

0, 0.16, 0.14, 0, 0.38, beta

0.25, 0.72, 0.25, 0.5, 0.67, beta

0.5, 0.55, 0.59, 1, 0.67, alpha

Deadline: 19 August 2024