Assignment 2
12 Oct 2023

Discriminate α and β type proteins using amino acid composition.

Steps:

1. Download the sequences for alpha (TMH) and beta (TMB)

Part 1: Analysis (3 marks)

2. Compute and tabulate the overall amino acid composition in TMH and TMB (20 values each).
3. Identify the amino acids, which are important for discrimination (use Fisher discriminant ratio). [FDR = $(m_\alpha - m_\beta)^2/(s_\alpha^2 + s_\beta^2)$; m: mean and s: variance; Ref: Bioinformatics Vol. 21, pages 4223–4229; https://sthalles.github.io/fisher-linear-discriminant/]. Include the data in the previous table.
4. Compute the residue pair preference (20x20 matrix) and tabulate the topmost five preferred pairs for alpha and beta. Identify the important pairs based on FDR (five).

Part 2: Discrimination (4 marks)

5. For each sequence in TMH

   (a) Compute the composition

   (b) Compare with overall composition of TMH and compute the absolute deviation and total for the 20 residues

   $$\sigma(\text{TMH}) = \Sigma\, |\text{comp}(x)-\text{comp}(\text{TMH})|$$

   (c) Compare with overall composition of TMB and compute the absolute deviation and total for the 20 residues

   $$\sigma(\text{TMB}) = \Sigma\, |\text{comp}(x)-\text{comp}(\text{TMB})|$$

   (d) If σ(TMH) < σ(TMB), the protein is TMH
   Otherwise, it is TMB

   (e) Correctly predicted TMH are True Positives (TP)

   (f) Wrongly predicted as TMB are False Negatives (FN)

6. Repeat the same with all TMB proteins. In this case,

   (e) Correctly predicted TMB are True Negatives (TN)

   (f) Wrongly predicted as TMH are False Positives (FP)

7. Compute sensitivity, specificity, and accuracy
   Sensitivity = TP/(TP+FN)

Specificity = TN/(TN+FP)
Accuracy = (TP+TN)/(TP+TN+FP+FN)

Tabulate TP, TN, FP, FN, sensitivity, specificity, and accuracy.

9. Take 50% of TMH and 50% of TMB to compute the composition (step 2). For the remaining set of proteins follow steps 5 to 7 to assess the performance.

   Tabulate TP, TN, FP, FN, sensitivity, specificity, and accuracy.

Part 3: Comparison of different features (3 marks)

(a) Use residue pair preference.
(b) Use the combination of amino acids and residue pair preferences.
(c) Use only the important features (amino acid composition, residue pair preference). Topmost “n” features

Discuss the results based on your own interpretation.
Compare with results obtained using machine learning techniques.

Optional (to explore)

10. Change the split in question 9 to 30%, 40%, 60% and 70% and repeat the computation. Tabulate the data.
11. In 5d include a deviation δ (E.g., σ(TMH) + 0.5) estimate the sensitivity, specificity, and accuracy.
12. Use data with non-zero elements.
13. Use correlation rather than deviation.

Deadline: 26 October 2023