Practical 1
06 August 2024

Discriminate α-helical and β-barrel membrane proteins using amino acid composition

Instructions

Upload your program files, output files and the report (naming: "roll no.pdf") in a zipped folder (naming: "roll no.zip").

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

Steps:

1. Go to PDBTM database (http://pdbtm.enzim.hu/)
2. Download alpha helical membrane protein sequences (TMH) in FASTA format
3. Obtain non-redundant sequences using CD-HIT software (40%)
4. Repeat steps 2 and 3 for beta barrel membrane proteins (TMB)

5. Compute and tabulate the overall amino acid composition in TMH and TMB (20 values each).

6. 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^2: variance; Ref: Bioinformatics Vol. 21, pages 4223–4229; https://sthalles.github.io/fisher-linear-discriminant/]. Include the data in the previous table.

7. 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
   $\sigma(TMH) = \Sigma\ |comp(x)-comp(TMH)|$
   (c) Compare with overall composition of TMB and compute the absolute deviation and total for the 20 residues
   $\sigma(TMB) = \Sigma\ |comp(x)-comp(TMB)|$
   (d) If $\sigma(TMH) < \sigma(TMB)$, 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)

8. 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)

9. Compute sensitivity, specificity and accuracy

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

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

11. Change the split in question 9 to 30%, 40%, 60% and 70% and repeat the computation. Tabulate the data. (Optional)

12. In 7d include a deviation δ (E.g., σ(TMH) + 0.5) estimate the sensitivity, specificity and accuracy. (Optional)

Deadline: 12 August 2024