

Figures for Assignment 4

Figure 1: Cleaned ITS sequences show broad but comparable length distributions

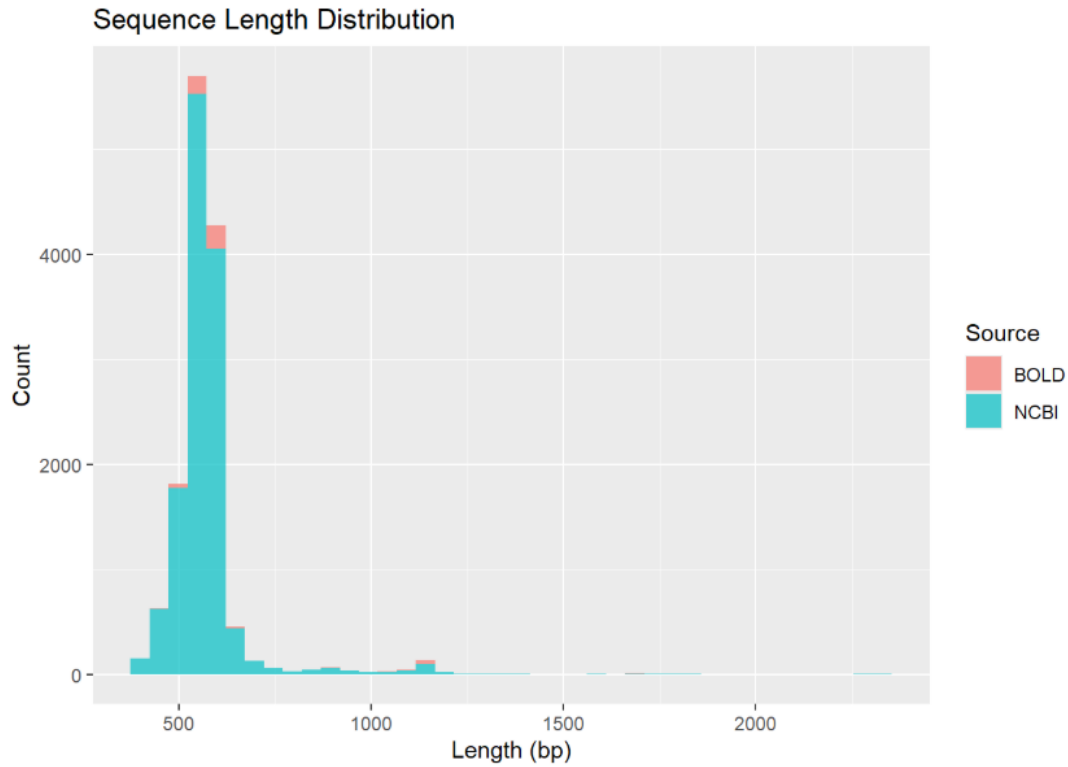


Figure 2 : Variation of GC content between databases could indicate technical biases

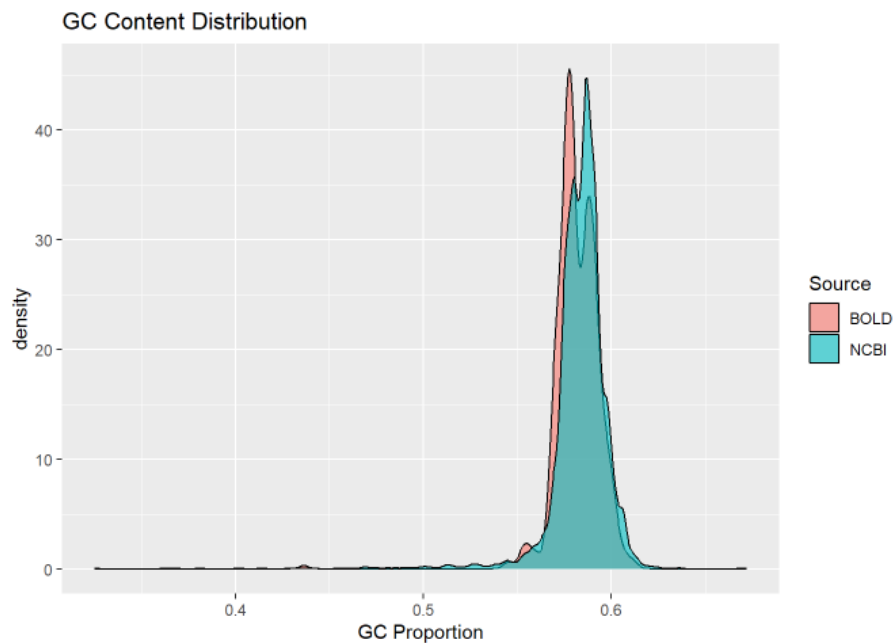


Figure 3: Strong class imbalance in cleaned dataset (H4 confirmed)

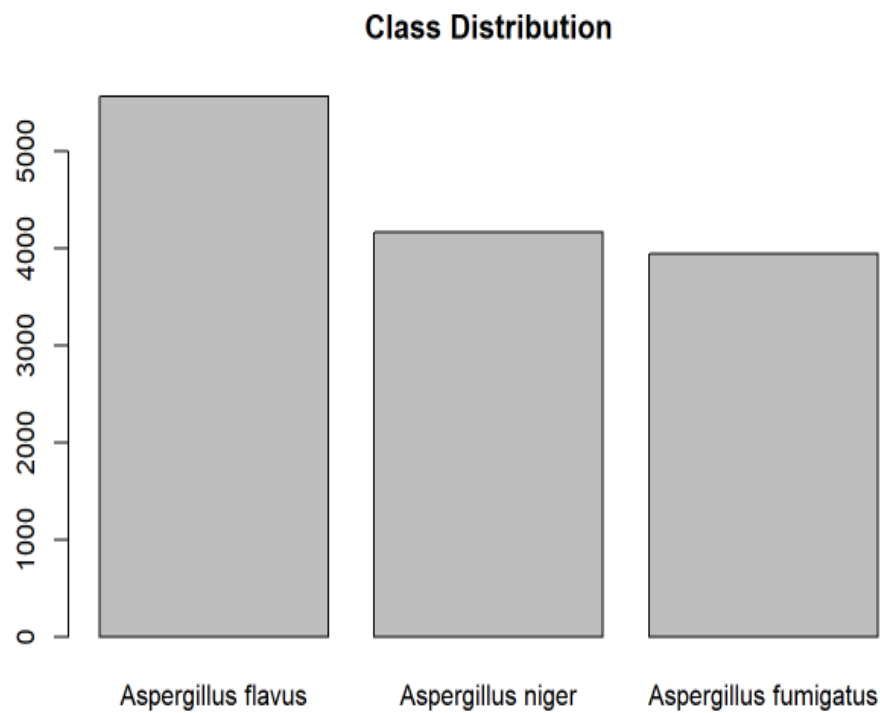


Figure 4: PCA – 3 mers shows partial clustering but overlapping profiles among species (H1 + H2).

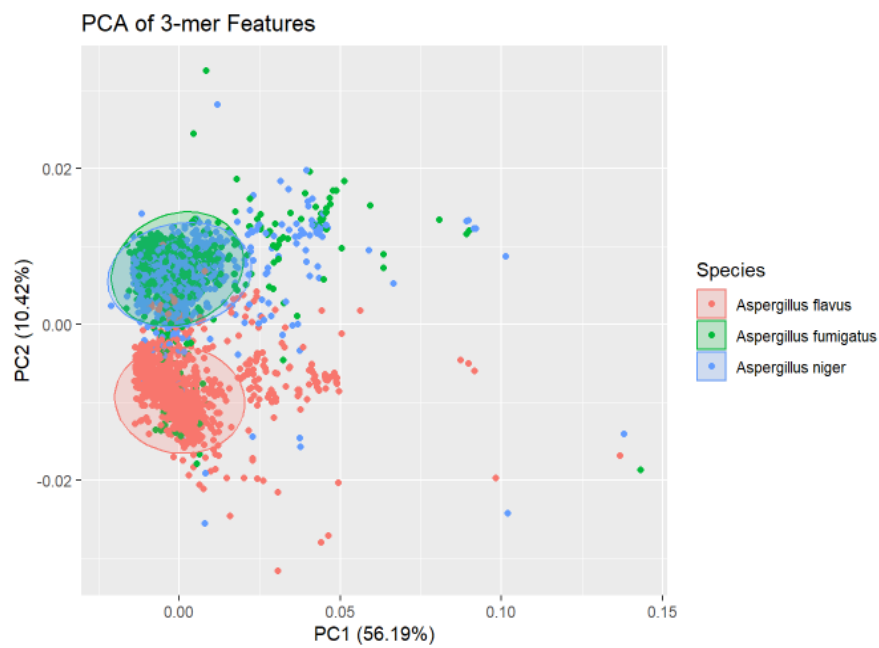


Figure 5: Top 20 Most Important 3-mers for Aspergillus Species Classification (Random Forest Mean Decrease Gini)

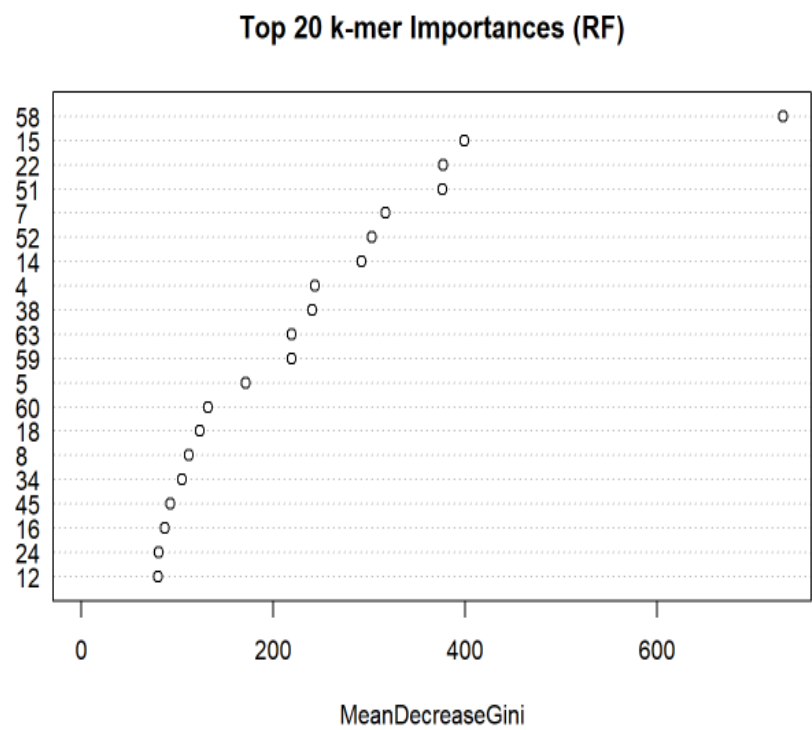


Figure 6: Random Forest + SVM + XGBoost Performance Comparison

