Predict the target specificity of RNA methylation writer and eraser (a machine learning/classification project)

Yiyou Song | Department of Bioinformatics | Year 3 Undergraduate
Qingru Xu | Department of Bioinformatics | Year 2 Undergraduate

SURF
Summer Undergraduate Research Fellowship

Abstract

N6-Methylated-adenosine (m6A) is determined by a complex protein family, termed‘writers’,‘readers’and‘erasers’. Here, we established m6A writer (METTL3, METTL14 and METTL16) and eraser (FTO and ALKBH5) classifiers using machine learning to predict the correspondingly substrate specificity. Comparable evaluation based on various conditions (feature types, model types and cutoff groups) have been considered and also examined by cross-validation.

Introduction

- Tens of thousands of N6-Methylated-adenosine (m6A) sites are only regulated by small number of regulators (writers, erasers and readers). Compared with previously m6A sites recognition researches, in this study, we we try different machine learning methods for investigating the m6A substrate specificity within one type of regulators, i.e. METTLE3, METTLE14, and METTLE16 for writers; FTO and ALKBH5 for erasers.
- R is a programming language and software environment for statistical computing and graphics, which is widely used among statisticians and data analysis. Caret is a R package using for classification and regression training.
- AUROC (Area under ROC curve) value ranging from 0 to 1 is used to evaluate the performance (the larger, the better).

Method

Feature selection → Feature types performance analysis → Assessment of different models. → Assessment under distinct cutoff group

- **Feature types**

Genome-derived features: using domain knowledge such as the mRNA Transcript topology to describe the m6A sites

Sequence-derived features:

1.chemical encoding

A and G, two rings; C and U, only one ring. → $x_i = \begin{cases} 1 \text{ if } s_i \in \{A, G\} \\ 0 \text{ if } s_i \in \{C, U\} \end{cases}$

A and C , amino group; G and U, keto group. → $y_i = \begin{cases} 1 \text{ if } s_i \in \{A, C\} \\ 0 \text{ if } s_i \in \{G, U\} \end{cases}$

A and U, weak hydrogen bonds; G and C, strong hydrogen bonds. → $z_i = \begin{cases} 1 \text{ if } s_i \in \{A, U\} \\ 0 \text{ if } s_i \in \{C, G\} \end{cases}$

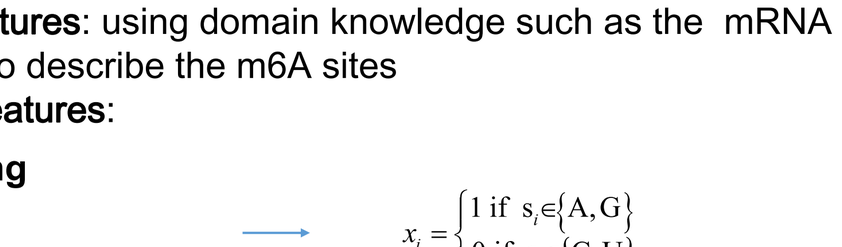
2. nucleotide frequency

$$d_i = \frac{1}{|N_i|} \sum_{j=1}^{l} f(n_j), f(n_j) = \begin{cases} 1, if\ n_j = q \\ 0, other\ cases \end{cases}$$

- **Cutoff groups**

0.6,0.7,0.8 and 0.9 cutoff means the probability of m6A data site to be the true.

Dataset



Result

1. feature selection.

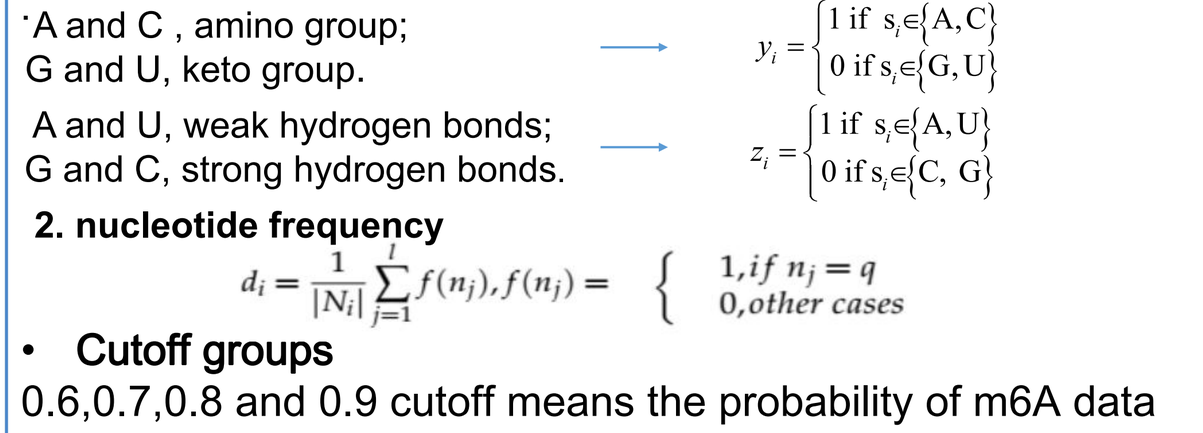


figure 1. Variable selection results for methyltransferase like 3, 14 and 16 with the cut-off >0.6

Black bars represents selected variables while gray bars represents abandoned variables based on full feature. “GC_cont_genes”, “PC_101bp”,“length_gene_full”, “length_UTR3”, “length_gene_ex”, “exon_num” are linked with methyltransferase like 3. “GC_cont_101bp”, “GC_cont_genes”, “PC_101bp”, “length_gene_full”, “length_gene_ex” are the topmost 5 features linked with methyltransferase like 16 and methyltransferase like 14.(**Figure1**)

2. Assessment of one model’ s performance on cross-validation under different feature type.

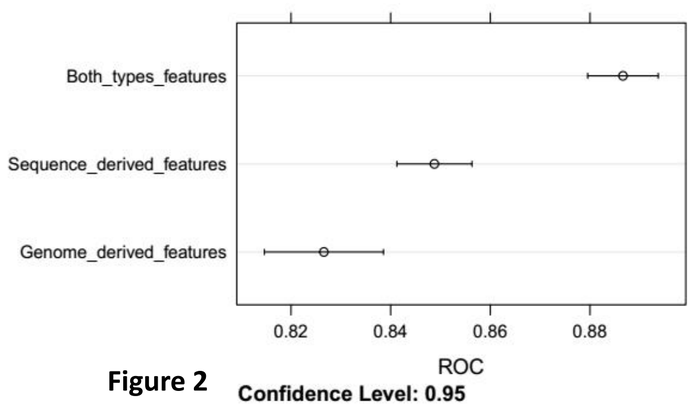


Figure 2

For eraser, the dot represents the mean of the 5-fold cross validation sampling, and the extended line define the 95% confidence interval (mean±1.96SD) (**Figure2**). As a result, all of these trainings are feasible especially for the features including both types.

3. Assessment of different models in distinct cutoff condition.

Except for the overall encouraging performance and better result on GBM and SVM using both types features, there is no difference under lower cutoff conditions (0.6, 0.7, and 0.8) (**Figure 3**). Yet the greatest performance are achieved on 0.9 cutoff group which means the most real true m6A motif data are more beneficial for classifiers to predict better.

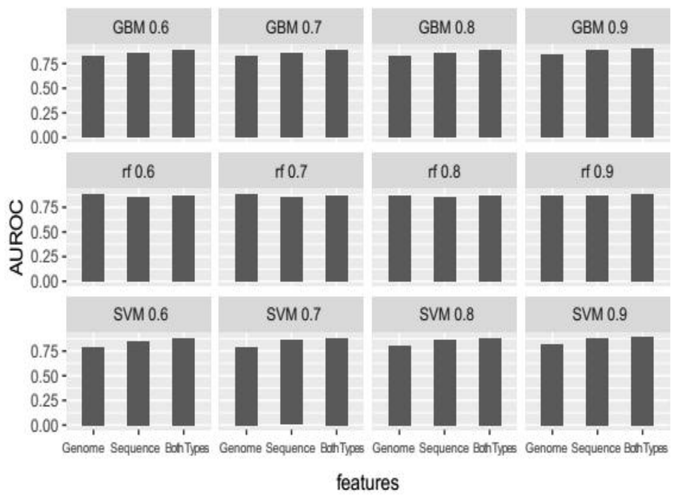


Figure 3

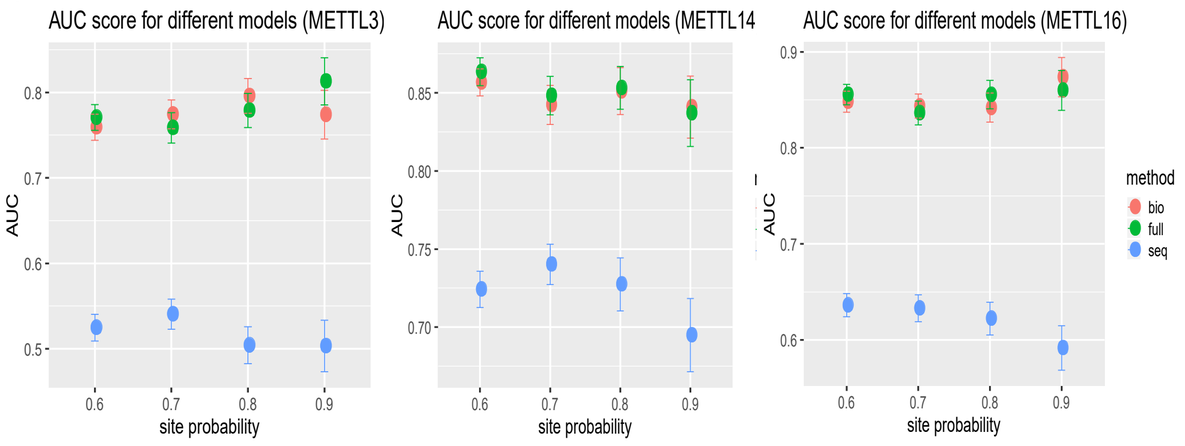


Figure 4

AUROC of identifying methyltransferase like 3 from methyltransferase like 14/16 (error bar: 1 standard deviation),AUROC of identifying methyltransferase like 14 from methyltransferase like 3/16 (error bar: 1 standard deviation) and AUROC of identifying methyltransferase like 16 from methyltransferase like 3/14 (error bar: 1 standard deviation). **Figure 4** showed that the AUROCs of models based on both biological feature and full feature (combination of both biological and sequence feature) are much higher than that based on sequence feature.

Discussion

1. All eraser predictors under various conditions achieved acceptable performances.Binding genome-derived features and sequence-derived features generally performs better than any of the individual type.
2. As for writer predictors, models based on biological and full feature is better than models based on sequence feature. Models baseed on sequence feature received acceptable performance for METTL14 only.
3. The highest 0.9 cutoff group representing the most real true m6A motif data are more beneficial for classifiers to recognize and predict better.