Model Invariant Predictive Features for RNA HOTness in Drosophila

Andrew Chen
Stat 157 Class Project

Abstract:

In this paper, we assess important features, and interactions, for the supervised binary classification problem of predicting High Occupancy Target (HOT) RNAs - RNA targets which bind to many RNA-binding proteins - using CHIP-seq and RNAi data. We used labels from Stoiber et al. and trained logistic regression models, vanilla multilayer perceptrons (MLP), and Iterative Random Forests (iRF) to predict HOT targets, and assessed feature importance on a held-out test set using permutation based feature importance (Logistic, MLP), and stability scores (iRF). We find that two features, caup_10.12 and caup_16.18, are highly important across models, and that the interaction between these two features is highly stable.

Introduction:

Ribonucleoproteins (RNPs) are macromolecules that contain both RNA and RNA-binding proteins. These RNPs participate in a wide array of biological functions, including, but not limited to, splicing, polyadenylation, stability, transportation, localization, and translation [1]. In the paper *Extensive cross-regulation of post-transcriptional regulatory networks in Drosophila*, the authors studied a set of twenty RNA-binding proteins (RBPs) and analyzed both the RNA and protein compositions of these RNPS, and among the several findings in the paper, found that more than two hundred RNAs were High Occupancy Target (HOT) - HOT RNAs interact with more than half of the RNA-binding proteins (RBP) studied in the paper [1].

In this work, we seek to find HOT targets without RBP data, and instead look for signals in existing CHIP-seq and RNAi data for Drosophila. Using both classical regression models (logistic), vanilla Multilayer Perceptions, and iterative Random Forests, we seek to find model-invariant predictive features from CHIP-seq and RNAi assays that are important in classifying RNA HOTness.

Results:

Model Performance

Exploratory Data Analysis with UMAP, does not reveal easily observable clusters (see Fig 1), however all three models examined in this study were able to achieve good accuracy (better than random guessing) on the held-out test set after training (Fig 2). All three models were trained on the same oversampled dataset, and evaluated on the same test-set to allow fair comparisons across models. The original dataset was oversampled to achieve balanced class-representation. Other methods were attempted at achieving balanced class-representation (See Methods).

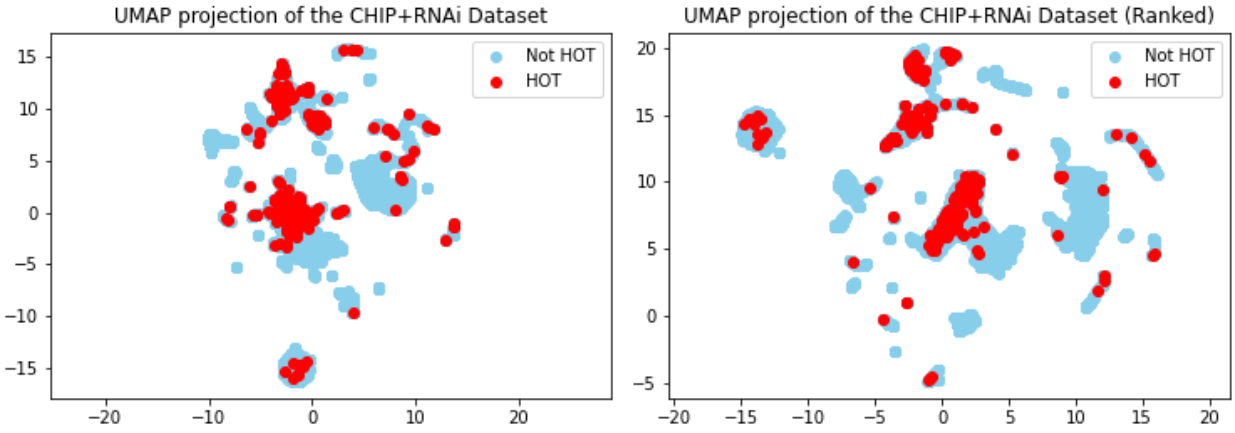

Fig 1. UMAP of CHIP + RNAi Datasets (before oversampling) with and without transforming CHIP Data to order statistics.

	Train accuracy	Test accuracy
Logistic	0.87	0.75
MLP	0.76	0.76
iRF	N/A	0.93

Fig 2. Summary of Model performance. For iRF, the test accuracy of the final weighted Random Forest is reported. The training accuracy for this final Random Forest could not be obtained.

Both the Logistic Regression model and MLP achieve similar accuracy on the held-out test set, whereas the final weighted Random Forest in the iRF algorithm achieves near-perfect accuracy on the test set.

Important Features and Interactions

To evaluate feature importance, we conduct permutation feature importance tests for each feature in the datamatrix (See Methods). For iRF, we report stability scores instead of permutation feature importance tests as stability scores allow us to examine higher-order interactions instead of just marginal feature importances [2]. Since the results of permutation feature importance tests can vary across runs, we average the results of ten permutation importance tests for each feature, and report the coefficient of variation (CV) (See Fig 3).

Of particular interest in Fig 3 are the caup_10.12, and caup_16.18 features. Both Logistic Regression and iRF highlight these as important features. For iRF, caup_10.12 is present in all top ten higher-order interactions, and caup_16.18 is present in four of these interactions. For Logistic Regression, caup_10.12 and caup16.18 are the top two important features. It is not known why important features in the MLP model do not overlap with important features found in the other two models or whether the features identified in MLP are actually significant.

A)

Interaction	Stability Score
caup_10.12-caup_16.18	1.00
caup_10.12	0.93
caup_10.12-Chip.CG14710	0.87
caup_10.12-Chip.Deaf1	0.77
dmrt99B_16.18-caup_10.12-caup_16.18-Chip.Deaf1	0.77
dmrt99B_16.18-caup_10.12	0.67
dmrt99B_16.18-caup_10.12-caup_16.18	0.67
dmrt99B_16.18-caup_10.12-Chip.Deaf1	0.67
caup_10.12-Chip.CG2116	0.63
caup_10.12-caup_16.18-Chip.CG14710	0.60

B)

Logistic			MLP		
Feature	Mean Importance	CV	Feature	Mean Importance	CV
caup_16.18	1.20	0.05	CG9305_16.18	1.02	0.01
caup_10.12	1.08	0.03	Chip.CG1529	1.02	0.01
Chip.CG10147	1.07	0.04	CG32006_12.14	1.02	0.01
Chip.MESR4	1.05	0.02	Chip.bigmax	1.02	0.01
Chip.corto	1.05	0.02	Chip.pfk	1.02	0.01
Chip.pzg	1.05	0.02	CG17181_8.10	1.01	0.01
CG9876_16.18	1.05	0.03	Chip.cg	1.01	0.01
Chip.salr.x	1.04	0.02	Chip.Pdp1.x	1.01	0.01
Chip.eyg	1.04	0.01	Chip.MESR4	1.01	0.01
Chip.gcm2	1.04	0.02	Chip.Ets97D	1.01	0.01

Fig 3. A) Top ten higher-order interactions sorted in descending order by stability score. B) Top ten features for Logistic and MLP models sorted in descending order.

Methods

To achieve balanced class representation, we tried both oversampling the rare class (HOT RNAs), and ensembling models trained on smaller balanced datasets in which the majority class is undersampled for each model. The latter performed poorly on the metric of accuracy for both the training and test data which is why we decided to proceed with oversampling. However, oversampling increases the chance of overfitting, especially to the rare class (each HOT RNA is sampled 70 times on average in the final data set).

Hyperparameter tuning was done for both the logistic regression model, and MLP model with scikit-learn's default cross-validation libraries [3]. For iRF, the hyperparameters chosen in the Github demo (https://github.com/Yu-Group/iterative-Random-Forest) were used except we increased K (the number of iterations for the iterative re-weighting) to 7, and increased n_estimators (the number of decision trees in each re-weighting iteration) to 200. Models and code in this paper can be found on github (https://github.com/andrewdchen/RNAHOTness).

To assess feature importance in both the logistic regression model, and MLP model, permutation based feature importance was conducted. The algorithm is adapted from Fisher et al. [4] and for this analysis can be summarized as follows:

1. Train model p on predictors X_train, and labels Y_train.
    a. Obtain final accuracy on test data. Acc_orig = Acc(p(X_test), Y_test)
2. For each feature j in X_test:
    a. Generate X_perm by permuting column j.
    b. Compute Acc_perm = Acc(p(X_perm), Y_test)
    c. Calculate feature importance (FI) as Acc_orig/Acc_perm
3. Sort features by descending importance.

Feature importances were calculated on the test set to estimate important features in unseen data, and avoid calling important features that are merely a symptom of overfitting.

Discussion

The results of this analysis suggest that caup_10.12 and caup_16.18 are important predictive features in classifying HOT RNAs, and that these features are model-invariant at least between iRFs and logistic regression. Further studies are required to elucidate the connection between these RNAi assays and the HOT RNAs.

In addition, although this analysis shows that permutation based feature importance can be used to identify important features, there are several limitations. If two features are highly correlated, their permutation based importance will be lower than expected as the predictor could always pick the other correlated feature and vice versa. This is accounted for in iRF through the inclusion of Random Intersection Trees (RIT) which preserve important features even if they are correlated. Further, in this analysis permutation based feature importance finds a completely disjoint set of important features for the MLP model which highlights both the importance of looking at multiple models to find robust predictive features, and the need to study why important features in one model can become unimportant in another model.

References

1. Stoiber MH, Olson S, May GE, Duff MO, Manent J, Obar R, et al. Extensive cross-regulation of post-transcriptional regulatory networks in Drosophila. Genome Res. 2015;25: 1692–1702.

2. Basu S, Kumbier K, Brown JB, Yu B. Iterative random forests to discover predictive and stable high-order interactions. Proc Natl Acad Sci U S A. 2018;115: 1943–1948.

3. Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python. Packt Publishing Ltd; 2013.

4. Aaron Fisher, Cynthia Rudin, Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of Machine Learning Research 2019.