

## Module 5 report

Project title: **Can mRNA stability be used to predict mRNA localization in neurons?**

By : **Marina Chekulaeva**

Prepared by: Aref Hosseini; [aref.hosseini@unibe.ch](mailto:aref.hosseini@unibe.ch); 24.11.2024

---

### Project Overview

Marina Chekulaeva's Module 3 project investigates whether mRNA stability can be used to predict mRNA localization in neurons. Neurons, as highly specialized cells, have unique structural features, including cell bodies and long axons. All mRNA production occurs in the nuclei of these cells, after which the synthesized RNA is transported to other parts of the neuron, such as axons, via motor proteins.

To address the research question, the study examined several key features of mRNAs, including their localization, half-life, and specific RNA motifs (ARE, m6A, tAI, and 3'UTR length). The dataset was analyzed to evaluate associations between these features and to predict mRNA stability and localization.

The study followed these steps:

1. **Feature Association:** Correlations between mRNA half-life and RNA motifs were assessed.
2. **Stability Prediction:** RNA stability was predicted based on features like m6A, ARE, tAI, and 3'UTR length using the machine learning method.
3. **Localization Prediction:** mRNA localization was predicted using a model incorporating RNA stability and other features.
4. **Evaluation:** The results of the predictions were assessed using correlation analysis.

---

### Results

#### Feature Association

- Pearson correlation analysis revealed that m6A and ARE has a pnegative correlation with mRNA half-life, while tAI are positevely correlated with mRNA half-life.

#### Prediction of mRNA Stability

- Using a random forest model, the data was split into training (70%) and testing (30%) sets.
- The model trained on RNA features (m6A, ARE, tAI, and 3'UTR length) yielded a predicted  $R^2=0.23$  applied to the test set.

### **Prediction of mRNA Localization**

- A random forest model was trained and tested to predict mRNA localization using mRNA stability and RNA features (m6A, ARE, tAI, and 3'UTR length).
  - The correlation between the predicted and actual localization was  $R^2=0.18$ .
- 

### **Strengths of the Study**

1. **Novel Research Question:** Investigating the link between mRNA stability and localization in neurons provides valuable insights into cellular mechanisms and expands the understanding of RNA biology.
  2. **Comprehensive Dataset:** The inclusion of various RNA features, such as m6A, ARE, tAI, and 3'UTR length, ensures a holistic approach to understanding RNA behavior.
  3. **Use of Machine Learning:** The application of random forest models demonstrates an innovative approach to exploring relationships in complex biological datasets.
  4. **Quantitative Analysis:** Pearson correlation and  $R^2$  values provide measurable and interpretable results.
- 

### **How to Improve the Study**

1. **Enhance Model Performance:**
  - The  $R^2$  values for both mRNA stability ( $R^2=0.23$ ) and localization ( $R^2=0.18$ ) indicate weak predictive power. Future iterations could benefit from exploring alternative machine learning models, such as gradient boosting (e.g., XGBoost) or neural networks.
  - Feature engineering could include additional RNA characteristics, such as secondary structures or protein-binding data.
2. **Increase Dataset Size:**
  - A larger and more diverse dataset could improve the generalizability of the models and provide better training for machine learning algorithms.
3. **Address Class Imbalance:**

- If certain RNA motifs or localization categories are underrepresented, techniques such as oversampling (SMOTE) or class weighting in the model can be applied.

#### **4. Incorporate Biological Replicates:**

- Variability in RNA assays could be reduced by including biological replicates to improve reliability and confidence in the predictions.

---

## **Conclusion**

This study provides a foundational exploration of the relationship between mRNA stability and localization in neurons using RNA features and machine learning. While the correlations identified and predictive models developed are a promising start, the relatively low  $R^2$  values highlight areas for improvement in model performance and dataset quality. Addressing these limitations could significantly enhance the study's contributions to understanding RNA behavior and its potential applications in neurobiology.