Predicting RNA-Binding Preferences for Proteins

Millie Dwyer, Julie Helmers, Shasha Lin, Lihan Yao

New York University Center for Data Science, Fall 2017

Project Objectives

- Learn low dimensional representations of RRMs to determine proteins with similar binding preferences
- Apply deep learning methods to different representations of RRM sequences

Introduction

RNA Recognition Motifs (RRMs) are substructures of RNA-binding proteins that specify how the protein-RNA binding will occur. For the majority of identified RRMs, it is unknown which RNA sequences they bind to, out of tens of thousands of possibilities. However, there are many RRMs that are homologous, meaning that they share sequence patterns as well as binding behaviors. If it is possible to generate a representation of these similarities and relate this representation to binding preferences, it may be easier to tell how a newly identified RRM will behave. This approach would be faster and less expensive than experimentally determining behavior for individual RRMs.

Data

RRMs are represented as sequences of amino acids from a vocabulary of 22 letters. These sequences are at maximum around 90 letters long.

- Unlabeled data: PFAM - This dataset is used for training unspervised models, containing 99,000 aligned RRM sequences.
- Labeled Data: For approximately 200 RRM-containing proteins, the distribution of binding affinity over RNA sequences is known. Biologists have identified that RRM sequences with at least 70% sequence similarity have nearly identical binding preferences, which expands the labeled data to 719 samples.

Data Samples

Both labeled and unlabeled data are obtained in sequence and aligned formats. Aligned formats are generated externally from the Pfam[**pfam**] protein family database, using their Hidden Markov Model [**HMM**].

- Aligned RRM Sequence:

```
>T117424||RNCMPT00259_RRM__0-----TPSTNVFINY--IP---------P-
--RF-------------------------T---------E-----QD--------
--L-R---N-------------------------L-------CS-Q---Y-----
---------G-----------------E--I-IS-------------------
----------S----K-------------------------------------
-----------------------------------------------------
-----------------------IM------------------------------
-----------------------------------------------I-NL-
--E---------TG----QSKCFG---F-----------V-----------K-
---------F---------R-------E--------L-----------S-----
-Q---------A--H---------A---------A--------I-QA----------I-
---D---G-----M-------SIGN--------K------R--LLAKYAESQE-----
```

- Unaligned RRM Sequence:

```
>T117424||RNCMPT00259_RRM__0
TPSTNVFINY1PPRFTEQDLRNLCSQYGEIISSKIMINLETGQSKCFGFVKFRELS
QAHAAIQAIDGMSIGNKRLLAKYAESQE
```

Methods

We determined our representations from the hidden states in autoencoders, focusing on the following approaches:

- Seq2Vec: Previous standard for RRMsimilarity prediction. Applies Doc2Vec methods to biological sequences, using a sliding window of width 3.
- CNN+LSTM: Inspired by an image captioning architecture, this model combines a ResNet "encoder" and LSTM "decoder" to learn hidden representations of the RRMs.

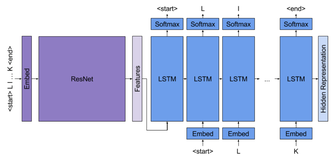

Figure 1: CNN and LSTM Architecture, adapted from `https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning`

- Character Level Autoencoder

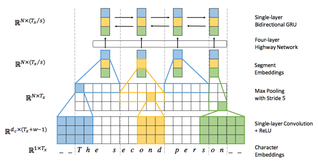

Figure 2: Network Architecture, adapted from **Cho et. al.**

Similarity Evaluation

Similarity regression is a method used to evaluate the predictive quality of our low dimensional representations. It is based on Affinity Regression (Pelossof et. al.), where a weight matrix is learned to translate a representation to its likely binding. Our method of similarity regression differs slightly in its use of training hidden representations, allowing for a more direct comparison between hidden states for similar binding affinities.

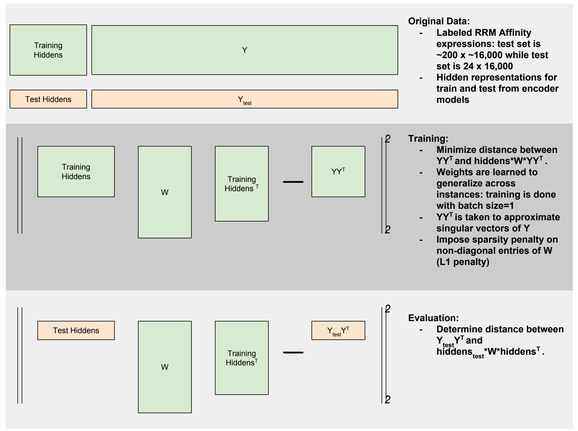

Figure 3: Similarity Regression

Results

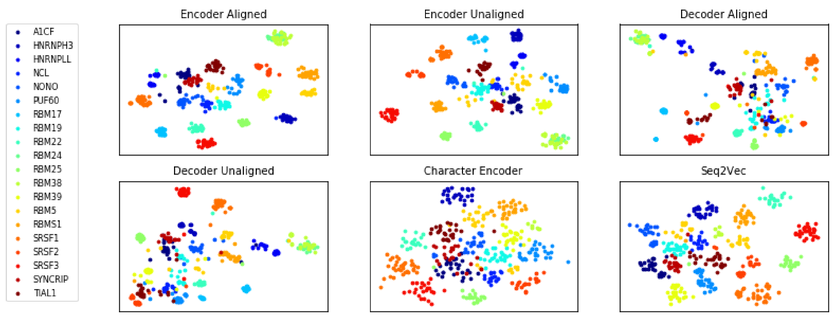

Figure 4: tSNE visualization for all learned representations

Model	Embedding Dim	Alignment	Training Score	Validation Score
Character Level AE	100	aligned	0.215	25572.64
CNN+LSTM	128	aligned	0.240	26675.57
seq2vec	128	aligned	0.373	26714.56
CNN+LSTM	128	unaligned	0.243	26710.64

Table 1: Similarity Regression results

Conclusion

- All 3 unsupervised models are able to learn meaningful representations of RRMs, either with original sequences or on top of aligned ones, as demonstrated by clustering of different RBP types in tSNE visualizations.
- Methods from natural language processing (NLP) can often be directly borrowed to solve problems in protein sequence analysis. The success of this project can be an indicator that more methods in deep learning and NLP can be applied to biology domains.
- We did not see a strong difference in performance between aligned and unaligned sequences for TSNE projections or similarity regression scores. While a highly useful technique in a large range of biology problems, HMM based sequence alignment may not be a necessity for analysis with deep learning models.

Future Work

- Train models on entire protein sequences, and learn RNA Binding Protein (RBP) preference without extracting RRMs first.

Appendix

While preprocessing for aligned data input, we eliminated positions that are less than 1% populated by letters. Graphs below shows the population rate for each positions we kept and removed as a result of this decision.

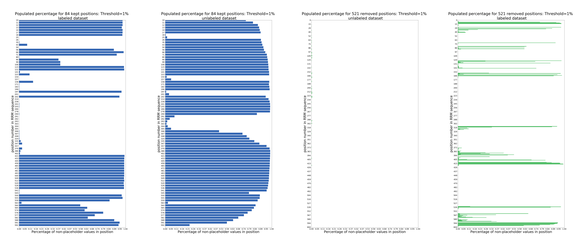

Figure 6: Rate populated by letters by sequence position

Acknowledgements

We would like to thank Professor Quaid Morris and Alexander Sasse for their guidance and assistance.