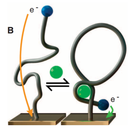
Prediction of Partially Structured DNA Aptamer Libraries

Sharon Newman, Ryan Chen, Teresa Noyola
{newmans, rjc45, tnoyola} @ stanford.edu

Background

- Aptamers are short synthetic DNA sequences that bind to specific targets—a function that is highly useful in applications ranging from sensors to therapeutics[1]



Ex Sequences

Seq 1: CAGTA
Seq 2: CTCTA
Seq 3: CCGTG
Seq 4: CGGTA

Consensus: CNGTA

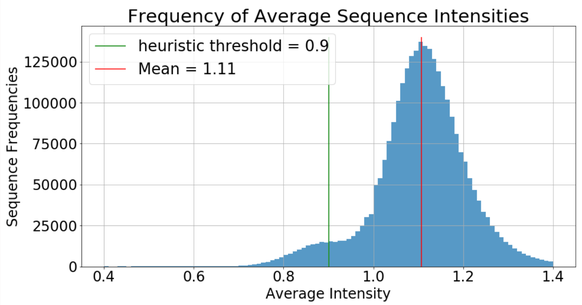
- Sequences are discovered by using random DNA libraries and filtering through multiple rounds of selection akin to natural evolution.
- Goal: Explore the possibility of conserved DNA motifs in good binding sequences.

Dataset

- There are 4.6 Million raw 30bp sequences mapped to intensity value indicating negative binding capability (Soh lab)[2]

Seq 1: CAGTAGTCATG = 1.45 ABU (bad binder)
Seq 2: CTCTAGTCATG = 0.74 ABU (good binder)

- **Pre-processing:** Intensity bandpass, >3 replicate seqs, normalization, rm seqs w/ large std.



Histogram of post-processing sequences mapped to intensity

Sparse Feature Based Classification

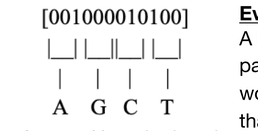


Figure 1: Vectorization of example sequence 'TCA'.

Evaluation metric

A lower weight for a character at a particular index suggests that it would be more favorable to insert that character into the consensus sequence at that index

Remarks

- Based on the ridge regression model, it is unfavorable to have G's at lower indices. It is overall more favorable to have A's and T's, and less favorable to have G's and C's.
- The non-regularized model is likely learning to consider the frequency of all bases at a particular index -- given the enormous dataset, these frequencies are likely equal for each base at every index

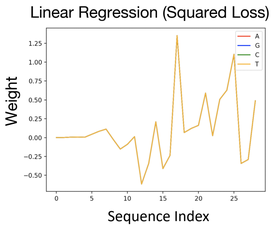


Figure 2: Weights learned by linear least squares model.

$$Loss_{ridge} = (y - \hat{y})^2$$

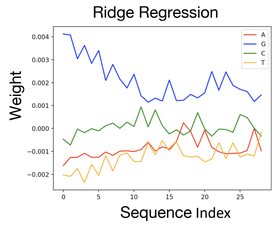


Figure 3: Weights learned by linear least squares model with l2 regularization.

$$Loss_{ridge} = (y - \hat{y})^2 + \sum_{j=1}^{n} w_j^2$$

Error Analysis

The mean squared error of both regression models was 0.01, so the percent error for most predicted values is on the order of 1%

K-gram Features with SVMs

- k-gram features given presence and absence of k-grams as a sparse frequency vector [3]
- Bad binders classified as y=0, good binders as y=1 during preprocessing
- Training set was resampled to have a more even balance of good and bad binders
- Evaluation - confusion table for results

	A	AT	ATC	ATCG	T	TC	TCG	C	CG	G	CGT	CGTA	GT	GTA	TA	TAG	TAGC	AG	AGC	GC
ATCG	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
CGTA	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0
TAGC	1	0	0	0	1	0	0	1	0	1	0	0	0	0	1	1	1	1	1	1

Figure 1: Example feature vector for sequence 'ATCG' showing sparse frequency vector of each possible k-gram from sequence

	Actual Good	Actual Bad
Predicted Good	43	47
Predicted Bad	2304	7617

Figure 2: Confusion table showcasing results

Remarks:

- Of the 2.4 million data points, only 140k were "good binders"
- Sparse frequency vector created with reference to [3]
- Consistent with existing sequence classification SVM results at ~82%

Results:

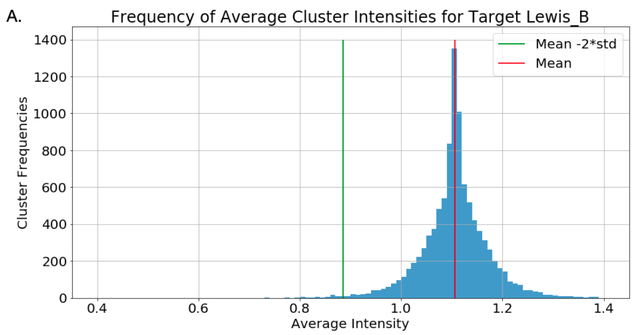
- Classification Rate - 76.5%
- Still classifies most aptamers as "bad binders" but not all weights are 0

Clustering via sequence distance

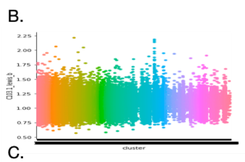
- Mini-batch K-means: clusters based on feature similarities, removing heuristic bias
- Explored clusters with average intensities two std below data mean
- Larger clusters led to increased number of clusters below mean, with similar sequences

- Objective function: $min \sum_{x \in X} ||f(C,x) - x)||^2$
- Number of clusters, k: 3 to 20,000
- Distance metric: euclidean distance

A.



B.



C.

Cluster #:	1894	C103.1_lewis_b
0.884514	ATCTTTTATATGGGTTGTATGTCTTGATCA	
1.180879	ATCTTTTATATGGGTTGTATGTCTTGATCA	
0.942643	ATCTTTTATATGGGTTGTATGTCTTGATCA	
1.107034	ATCTTTTATATGGGTTGTATGTCTTGATCA	
0.953722	ATCTTTTATATGGGTTGTATGTCTTGATCA	
0.770335	ATCTTTTATATGGGTTGTATGTCTTGATCA	
Cluster #:	4410	C103.1_lewis_b
0.866667	GTTCGCGGGGCTCCACTAGATAATAATCGT	
1.083056	GTTCGCGGGGCTCCACTAGATAATAATCGT	
1.069767	GTTCGCGGGGCTCCACTAGATAATAATCGT	
1.028571	GTTCGCGGGGCTCCACTAGATAATAATCGT	
0.727829	GTTCGCGGGGCTCCACTAGATAATAATCGT	
1.144737	GTTCGCGGGGCTCCACTAGATAATAATCGT	

Conclusion

- Using different feature extractors and interpretable models, we have shown the potential existence of DNA binding motifs in datasets like ours.

Future Works

- Extract explicit conserved DNA binding motifs from our models
- Design scalable feature extractors that encode for sequence dependencies and patterns
- Explore evolutionary algorithms more akin to experimental work

References

[1] Dunn M. *Nature Reviews Chemistry. 2017.* 1:10 0076
[2] Yoshikawa, A. In preparation. 2019.
[3] Xing Z., Pei J., Keogh E. A Brief Survey on Sequence Classification. SIGKKD Explorations. 2010.