BindSpace decodes transcription factor binding signals by large-scale sequence embedding

Presenter: Jack Lanchantin

University of Virginia
https://qdata.github.io/deep2Read/

201908

Motivation

- Direct measurement of genome-wide transcription factor (TF) occupancy for all expressed factors in a cell type of interest is currently infeasible outside of large consortium projects
- Therefore, computational prediction of TF binding to sites at regions of accessible chromatin or active histone marks is critically important

Drawbacks of Previous Methods

- Large-scale **in vitro** TF binding experiments provide large amounts of data for training binding models
- However, each experiment is typically summarized as a PWM, which yields near-identical PWMs for closely related TFs
- Previous supervised methods can discriminate accurately between bound and unbound sequences of individual TFs but do not develop a multiclass prediction model that can distinguish between TFs with similar binding signals

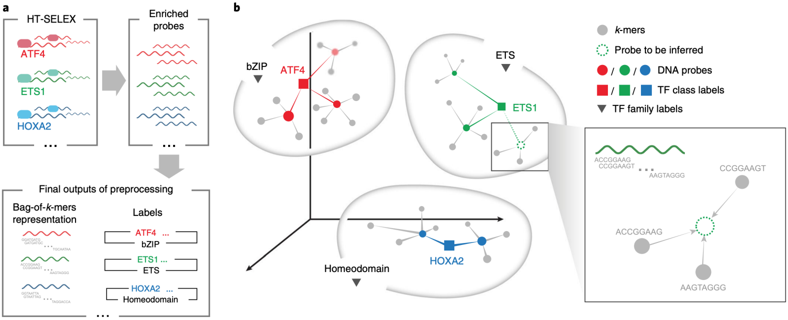
BindSpace

- A multiclass method for joint learning of binding models for hundreds of TFs assayed by HT-SELEX by embedding their bound and unbound DNA sequences and TF labels into a common high-dimensional space
- Adaptation of StarSpace which learns to embed words into a semantic space, in which words with similar meanings embed close to each other
- For multiclass problems, class labels are embedded in the same space

BindSpace

- In BindSpace, k-mers are analogous to words, and TFs and TF families serve as class labels
- BindSpace learns k-mer and label embeddings so that probes embed close to the labels of TFs that bind them and away from other labels
- For in vitro or in vivo TF binding prediction, a test DNA sequence is embedded in BindSpace and assigned the closest TF label

Overview


a
HT-SELEX
ATF4
ETS1
HOXA2
...
Enriched probes
...
Final outputs of preprocessing
Bag-of-k-mers representation
GGATGATG
GATGATGC
TGCAATAA
ACCGGAAG
CCGGAAGT
AAGTAGGG
GGTAATTA
GTAATTAG
TAGGACCA
Labels
ATF4 ...
bZIP
ETS1 ...
ETS
HOXA2 ...
Homeodomain
...
b
bZIP
ATF4
ETS
ETS1
Homeodomain
HOXA2
k-mers
Probe to be inferred
DNA probes
TF class labels
TF family labels
ACCGGAAG
CCGGAAGT
AAGTAGGG
CCGGAAGT
ACCGGAAG
AAGTAGGG

Training Data

- 270 experiments for 243 transcription factors
- **Positive examples**: top 2,000 enriched probes from each experiment (yielding 500,000 positive training sequences)
- **Negative examples**: randomly sampled universal negatives from HT-SELEX probe libraries and non-accessible genomic regions to obtain 500,000 negative training sequences

Evaluation Data

1. Held out HT-SELEX data
2. Independent PBM data sets to test TFs within the same family across in vitro platforms
3. In vivo sites from ENCODE ChIP-Seq
   - Two scenarios for negatives: dinucleotide shuffle from positive samples, and nonbinding regions of accessible chromatin

Sequence and Label Representation

- Each sequence is represented as a bag of 8-mers
- Each bag is associated with both a TF label (e.g., HOXA2) and a TF family label (e.g., Homeodomain) or with a universal negative label

Sequence Representation Details

- Each HT-SELEX probe input sequence s_i is represented by a bag of 8-mers with up to 2 consecutive wildcards (where the wildcard symbol N matches any nucleotide)
- A particular 8-mer is considered a token of s_i if it occurs in either s_i or reverse complement of s_i.

Sequence and Label Representation Summary

- Objective is to learn an embedding for a total of 113,074 entities
  - 112,800 k-mers (all 8-mers with max 2 wildcard)
  - 243 TF labels
  - 30 TF families
  - 1 universal negative label
- All entities are represented in a vector space of dimension d (d=300 in experiments)

BindSpace Framework

- Training examples for BindSpace are structured as left hand side (LHS) right hand side (RHS) pairs
- In BindSpace, the LHS of the ith input is a DNA probe represented by its constituent k-mers $(w_{i,1}, \dots, w_{i,m_i})$ and the RHS consists of the labels associated with this probe $(l_{i,1}, \dots, l_{i,n_i})$

Embedding Sequences and Labels

- The embedding of the LHS of the ith example is induced by the embedding of all constituent k-mers as follows:

$$\mathrm{lhs}_i = \frac{1}{m_i^p} \sum_{j=1}^{m_i} w_{i,j} \tag{1}$$

- Similarly, the embedding of the RHS of the example is induced by the embedding of all its associated labels:

$$\mathrm{rhsP}_i = \frac{1}{n_i^p} \sum_{j=1}^{n_i} l_{i,j} \tag{2}$$

Negative Samples

- To compute the loss associated with this example, we randomly sample K examples with labels different from example i and compute the RHS associated with each:

$$\text{rhs}\, N_{i,k} = \frac{1}{n_k^p} \sum_{j=1}^{n_k} l_{k,j} \tag{3}$$

Hinge Loss

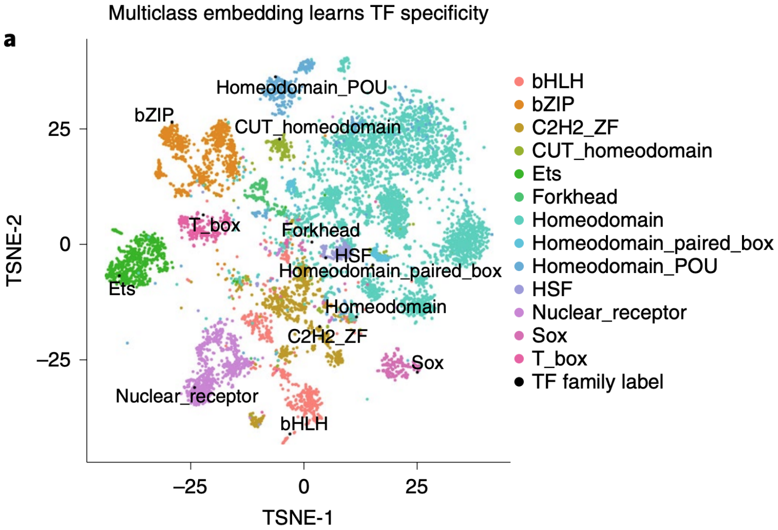
- The loss function for a given positive example with one random negative is:

$$\mathrm{Err}_{ik} = \mathsf{max}\left(0, \mathsf{margin} - \mathsf{lhs}_i \cdot \mathsf{rhs}\, P_i + \mathsf{lhs}_i \cdot \mathsf{rhs}\, N_{i,k}\right) \tag{4}$$

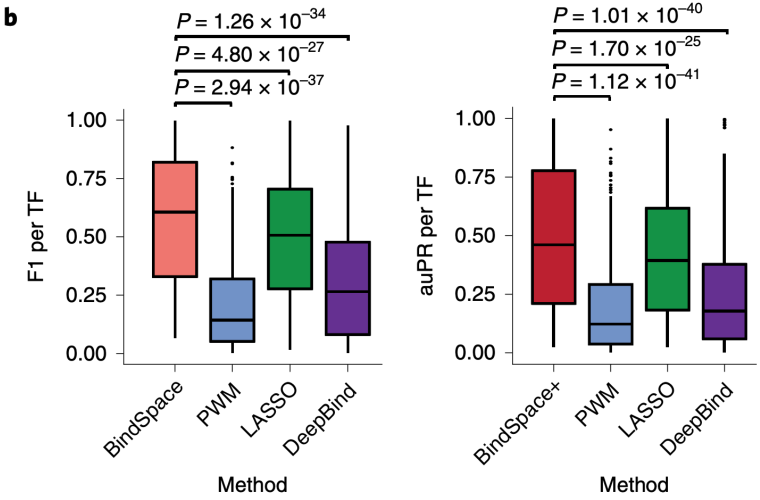
- The total loss associated with example i using K negative samples is:

$$\mathsf{Err}_i = \frac{1}{K}\sum_{k=1}^{K} \mathsf{max}\left(0, \mathsf{margin} - \mathsf{lhs}_i \cdot \mathsf{rhs}\, P_i + \mathsf{lhs}_i \cdot \mathsf{rhs}\, N_{i,k}\right) \tag{5}$$

T-SNE

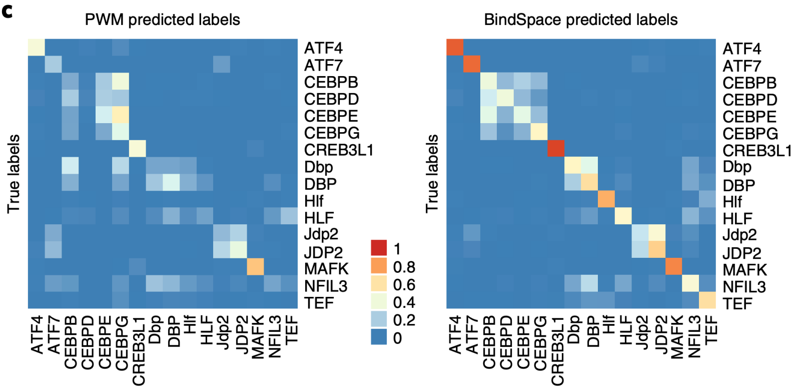

Multiclass embedding learns TF specificity
a
TSNE-2
TSNE-1
25
0
–25
–25
0
25
Homeodomain_POU
bZIP
CUT_homeodomain
T_box
Forkhead
HSF
Homeodomain_paired_box
Ets
Homeodomain
C2H2_ZF
Sox
Nuclear_receptor
bHLH
● bHLH
● bZIP
● C2H2_ZF
● CUT_homeodomain
● Ets
● Forkhead
● Homeodomain
● Homeodomain_paired_box
● Homeodomain_POU
● HSF
● Nuclear_receptor
● Sox
● T_box
● TF family label

HT-Selex Held Out Results



Multi-Class Confusion Matrix for TFs in bZIP family

Motifs of TFs in the bZIP family are similar



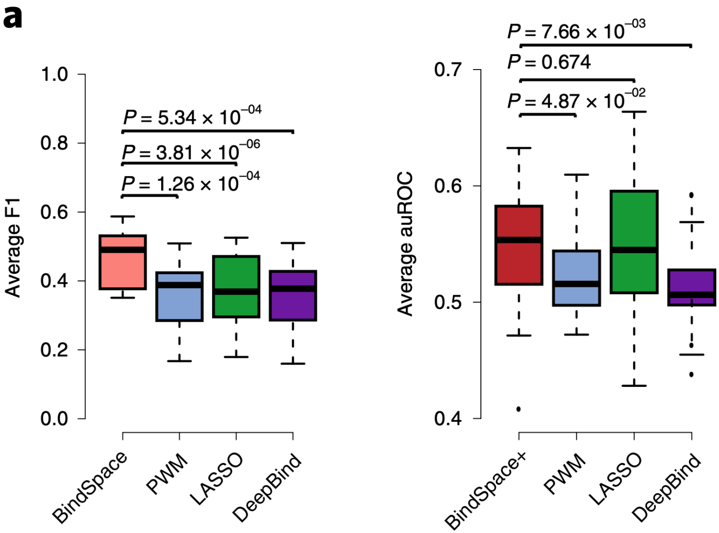
Evaluation Data from ENCODE

- TF binding versus nonbinding at chromatin accessible regions in a given cell type
- Processed publicly available ATAC-seq data and used ENCODE ChIP-seq data for 17 TFs in K562 and 11 TFs in GM12878 that had sufficient overlap with ATAC-seq peaks

Evaluation Data from ENCODE

- BindSpace significantly outperformed all competing methods on K562 by F1 score, and significantly outperformed LASSO on GM12878, but was not significantly bettern than PWM and DeepBind
- There was no significant difference between methods in terms of auPR

Distinguishing between paralogous (from the same family) TF binding sites in vivo - from ENCODE Data



Conclusion

- Train on HT-Selex, test on ENCODE
- Outperforms PWM and LASSO on multi-class outputs