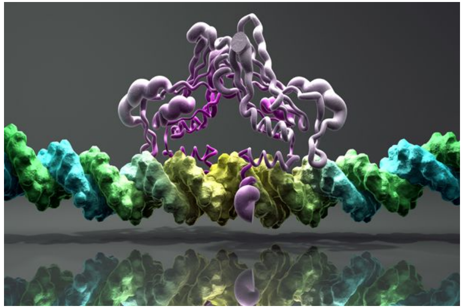
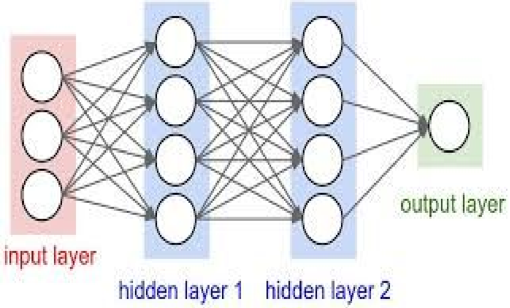
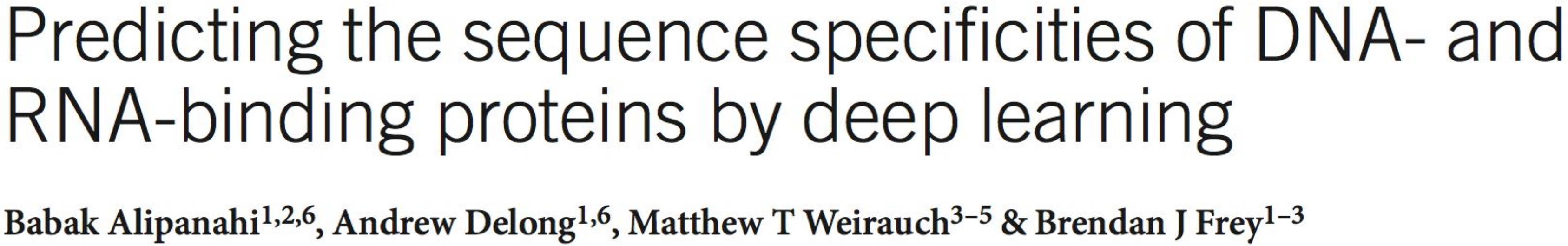


DeepBind





Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning

Babak Alipanahi[1,2,6], Andrew Delong[1,6], Matthew T Weirauch[3–5] & Brendan J Frey[1–3]

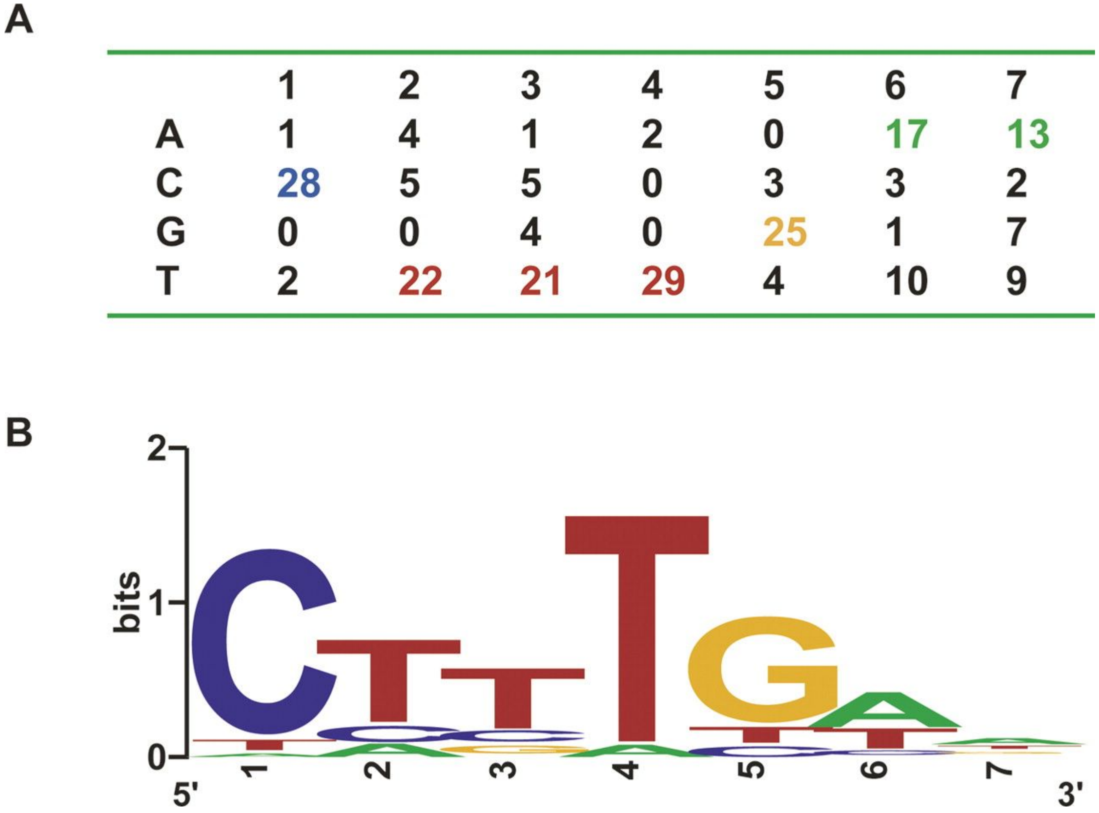
6.874 - Pranam Chatterjee

Why do we care?

- Regulatory processes
    - Transcription
    - Alternative Splicing
    - Disease correlation

- Sequence specificity

Position Weight Matrix



	1	2	3	4	5	6	7
A	1	4	1	2	0	17	13
C	28	5	5	0	3	3	2
G	0	0	4	0	25	1	7
T	2	22	21	29	4	10	9

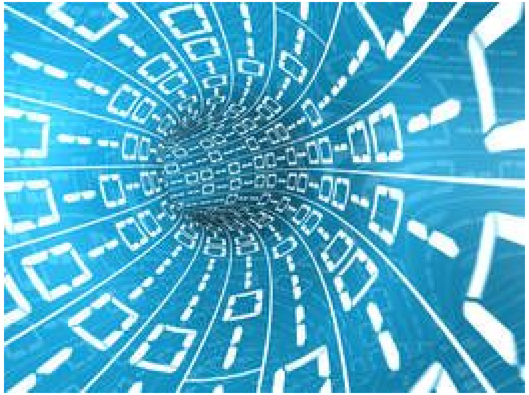
Steps:

1. Get PFM by counting occurrences of each nucleotide at each position.
2. Divide frequency by total # of sequences.
3. Formally, given a set X of N aligned sequences of length i:

$$M_{k,j} = \frac{1}{N}\sum_{i=1}^{N} I(X_{i,j} = k)$$

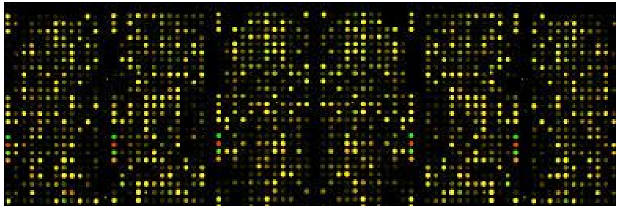
Data Issues

- **Different forms of data**
  - Specifity coefficient
    - Protein Binding Microarrays
    - RNAcompete arrays
  - Ranked Lists of Bound Sequences
    - ChIP-Seq
  - High Affinity Sequence List
    - HT-SELEX
- **Large Quantities of Data**
  - 10,000-100,000 sequences (1 EXPERIMENT)
- **Additional Biases/Limitations**
  - i.e., hyper-ChIPable regions of genome
  - Need to filter



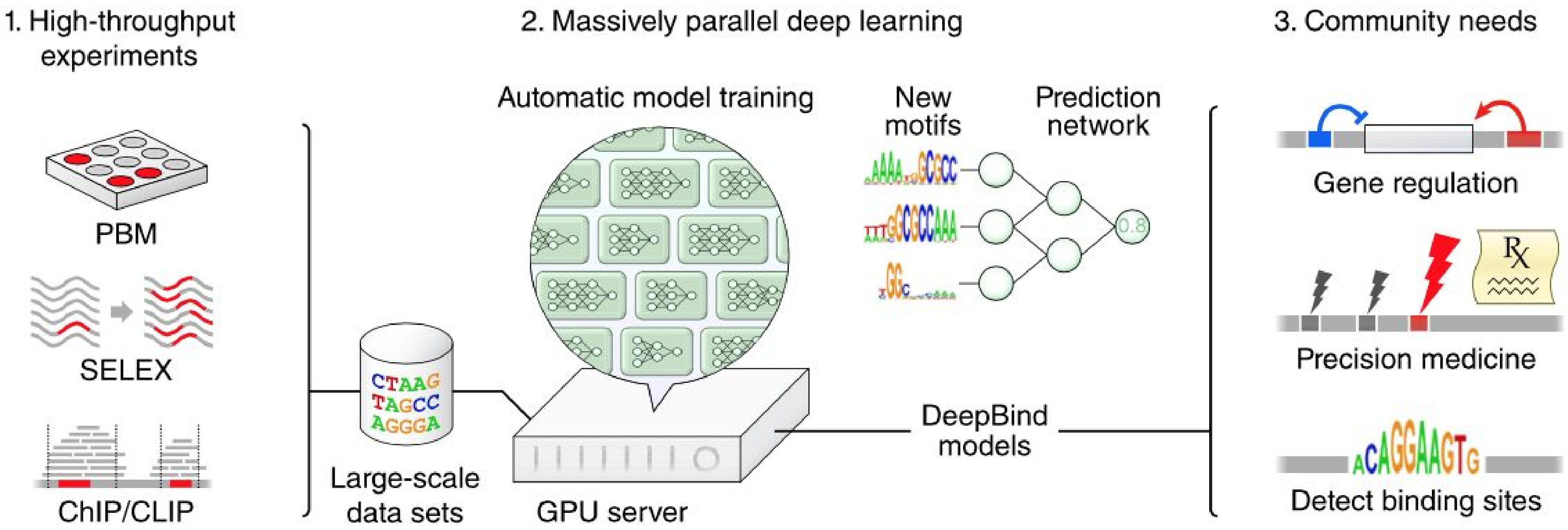
DeepBind Claims

- Apply to both microarray and sequencing data
- Generalize well across technologies
- Tolerate noise and mislabeled data
- Can learn from millions of sequences through parallel implementation on a graphics processing unit (GPU)
- Train models and tune parameters automatically
- Can discover new patterns without location information



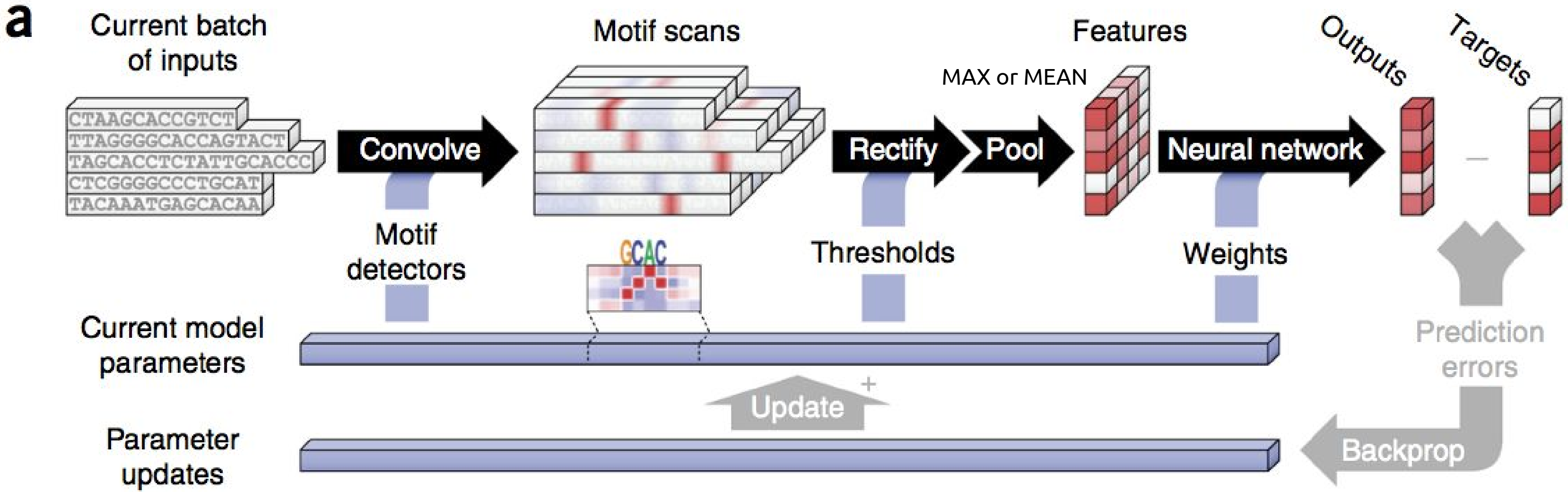
Alipanahi, et al., *Nature Biotechnology,* 2015.

Overview of DeepBind



Alipanahi, et al., *Nature Biotechnology,* 2015.

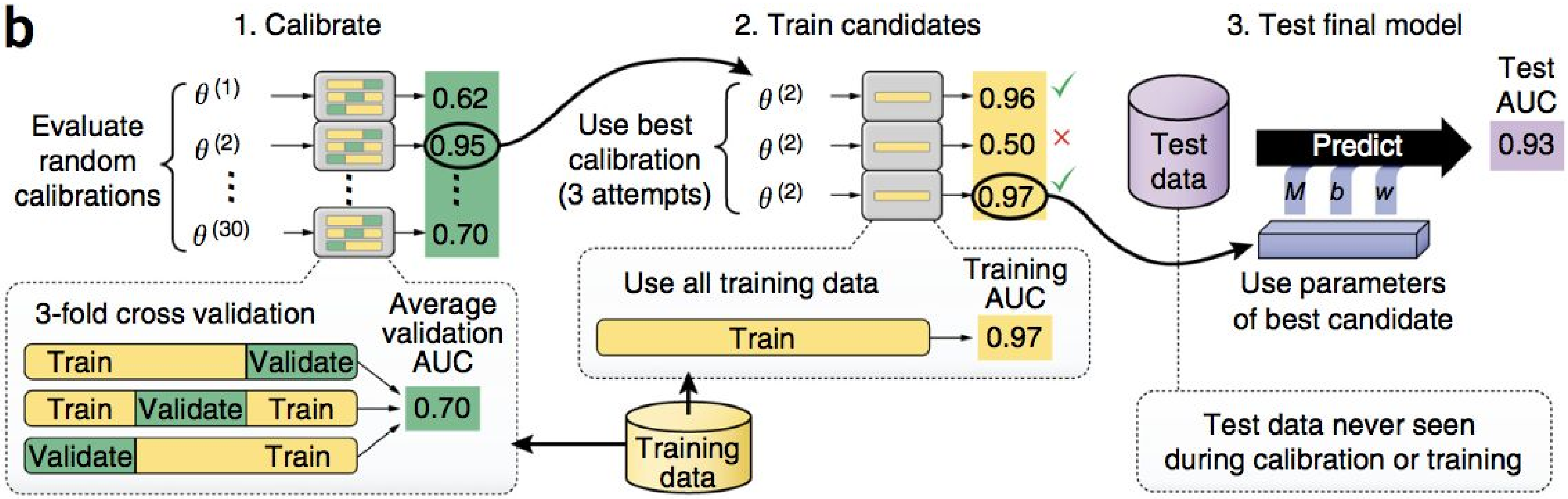
Training Procedure



BINDING SCORE → $\boxed{f(s) = \mathrm{net}_W(\mathrm{pool}(\mathrm{rect}_b(\mathrm{conv}_M(s))))}$

Alipanahi, et al., *Nature Biotechnology*, 2015.

Calibration and Testing Procedure



12 teraba😮of data!!!

Alipanahi, et al., *Nature Biotechnology*, 2015.

Let’s unpack that...

- Thousands of PBM, RNAcompete, ChIP-Seq, and HT-SELEX experiments
- Create 927 DeepBind models
- 538 Transcription Factors
- 194 RNA-binding Proteins (RBPs)

(This took 4+ years, btw)

Alipanahi, et al., *Nature Biotechnology,* 2015.

How well does it work?

- Test on PBM data from DREAM5 TF-DNA Motif Recognition Challenge
- 86 different mouse transcription factors
- 2 array designs (~40,000 probes each)
    - All possible 10-mers, non-palindromic 8-mers (32x)
- Train on probe intensities, predict on held-out test array design

Example Competing Algorithms (26 in total)

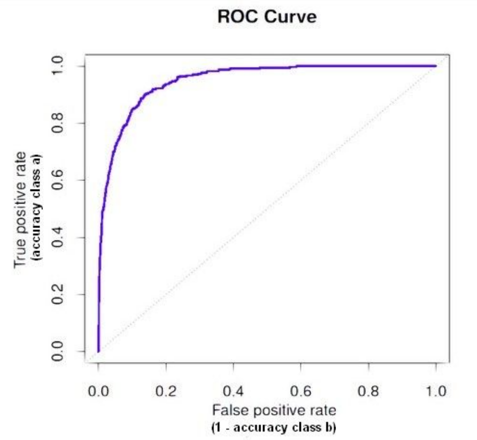
- FeatureREDUCE (biophysical PWM/k-mer)
- BEEML-PBM (weighted regression)
- RankMotif++ (probabilistic)
- PFM models (position frequency matrices)

None of these are deep-learning-based!

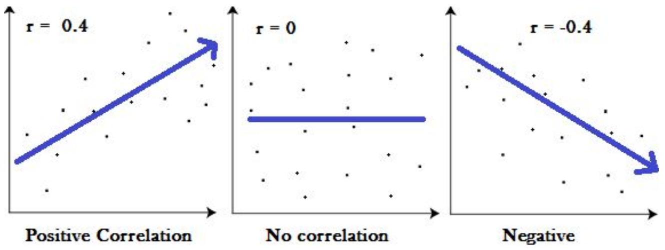
Alipanahi, et al., *Nature Biotechnology,* 2015.

Metrics

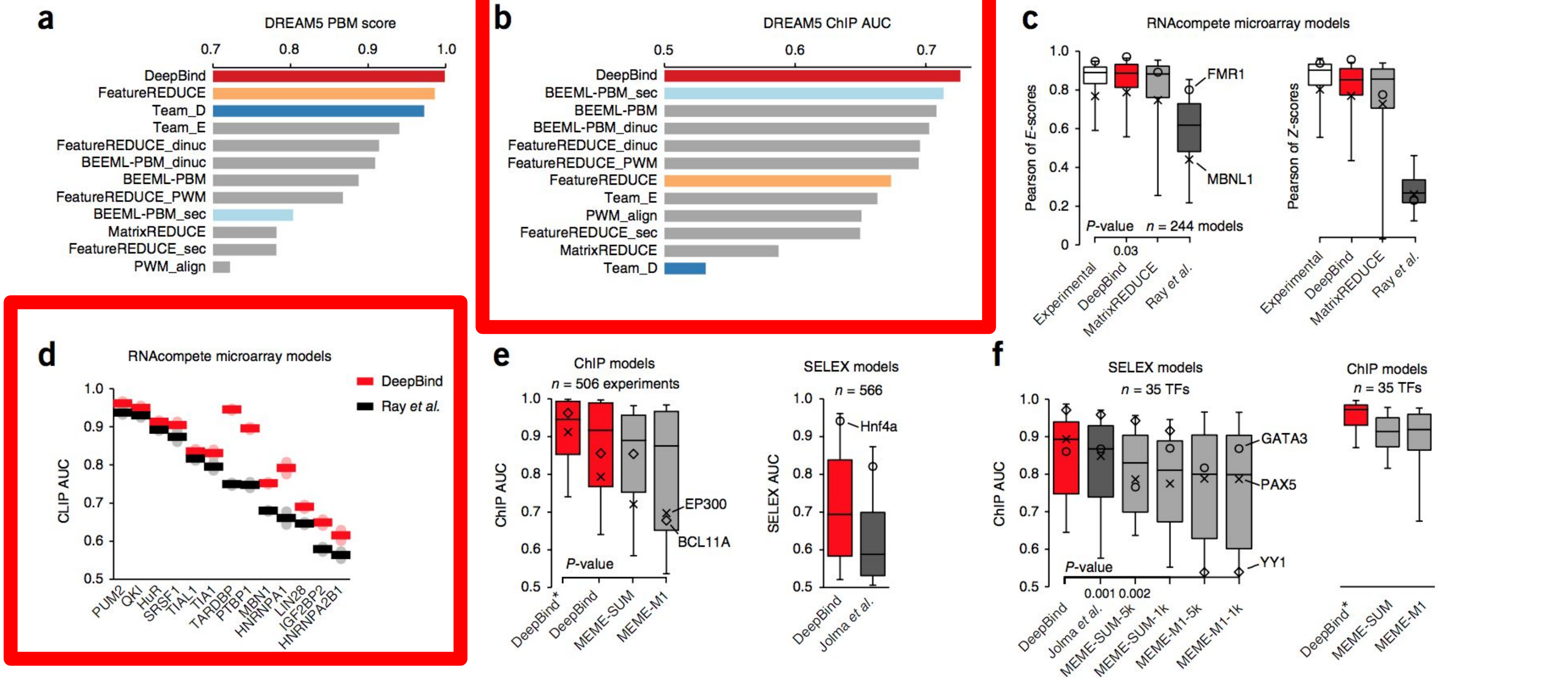
- Area Under Curve (AUC)
  - Measures true positive rate of model as a function of false positive rate (ROC curve)
  - Tells us how good the model identifies actual positives
  - Higher AUC means better performing model



- Pearson Correlation
  - Measures linear correlation between predicted intensity and probe intensities
  - Higher absolute values (maxed at 1), indicate better performing mode.



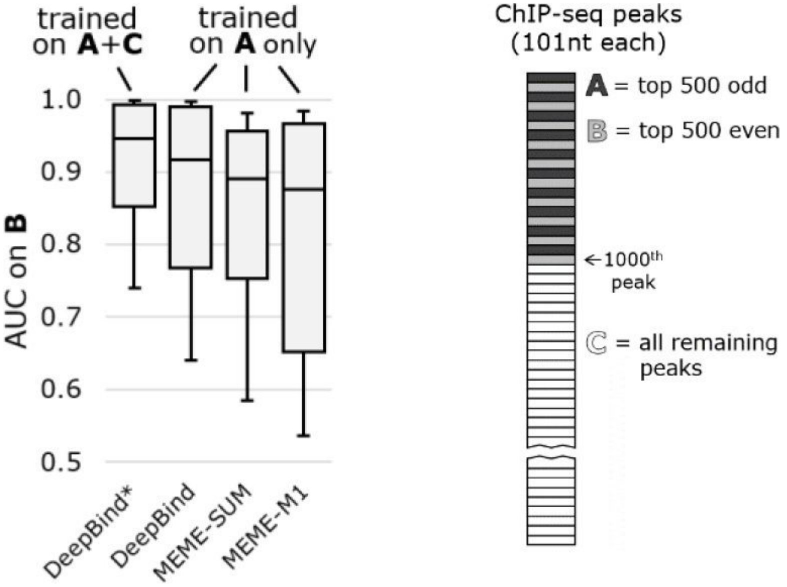
Quantitative Performance Against Other Methods



Alipanahi, et al., *Nature Biotechnology*, 2015.

Do *in vitro* models accurately identify *in vivo* bound sequences?

- 506 ENCODE ChIP-Seq data sets
- *In vivo* laboratory biases
  - Cell-type specificities
  - Nucleosome interactions
  - Chromatin remodeling, etc.
- 137 transcription factors
- Performed better than other non-deep learning methods based on AUC
- Can generalize to other data acquisition methods

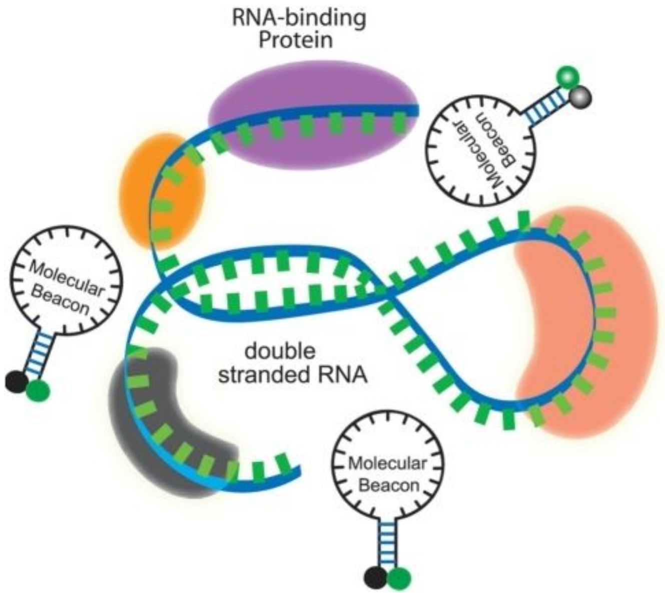


Alipanahi, et al., *Nature Biotechnology,* 2015.

First place goes to….DEEPBIND!

Why are RBPs sequence specifities difficult to predict?

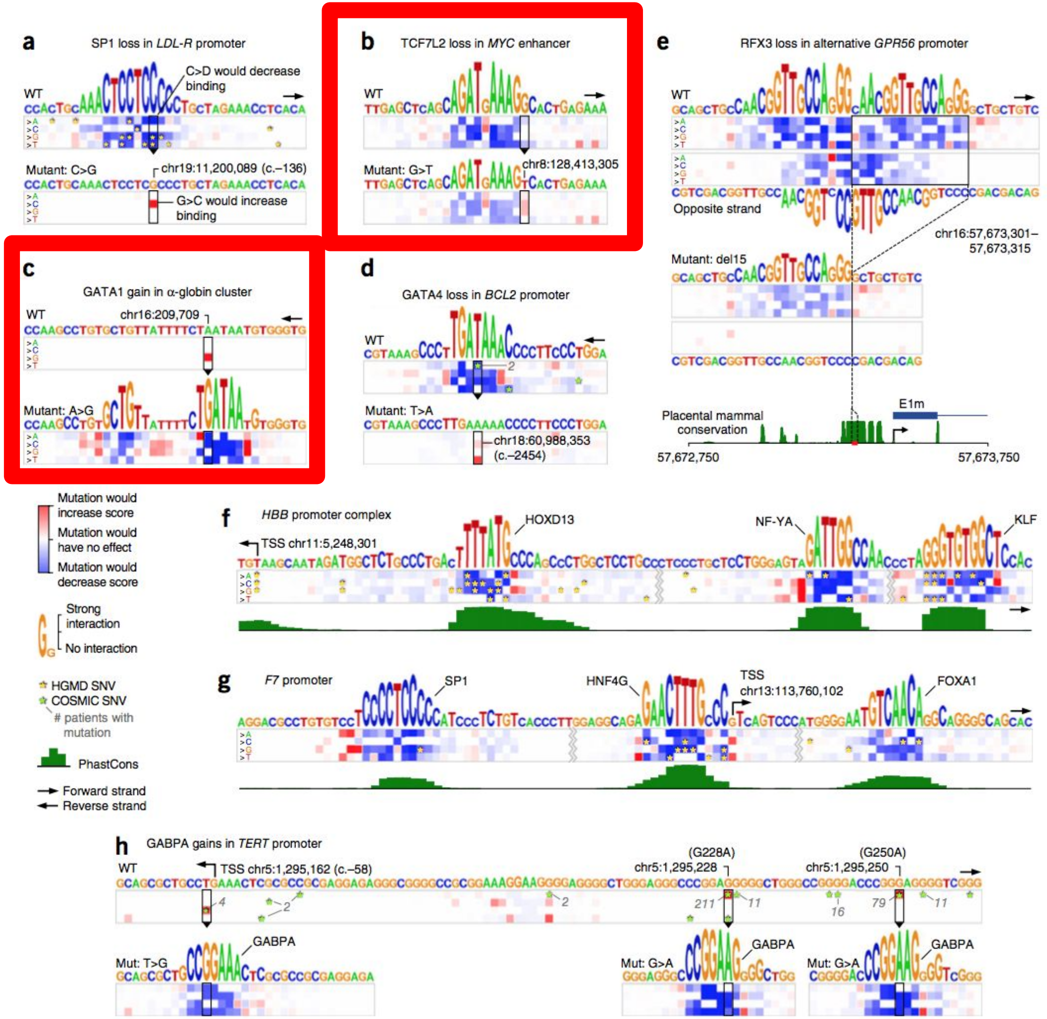
- Usually bind to ssRNA
- More flexible than DNA
- Can fold into stable secondary structures
- Recognition motif is highly flexible
  - Multiple domains neeeded for binding
- RNA structure also affects binding



Rhee, et al., *Nucleic Acids Research* 2008,

Identifying Damaging Genetic Variants

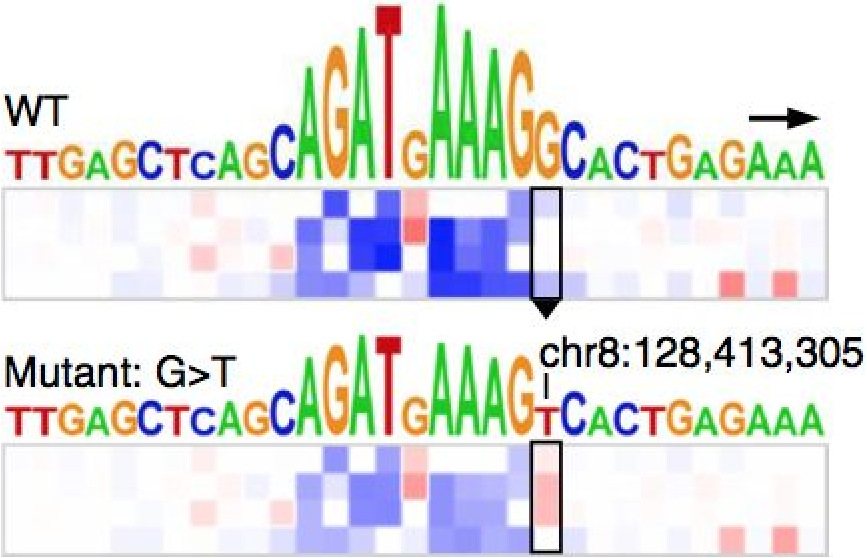
- How to do this?
- MUTATION MAPS!
  - Importance of each base
  - Effect of each mutation on binding score
- Illustrates effect of point mutations on binding affinity



a
SP1 loss in LDL-R promoter
WT
C>D would decrease binding
CCACTGCAAACTCCTCCCCCTGCTAGAAACCTCACA
Mutant: C>G
chr19:11,200,089 (c.–136)
CCACTGCAAACTCCTCGCCCTGCTAGAAACCTCACA
G>C would increase binding
b
TCF7L2 loss in MYC enhancer
WT
TTGAGCTCAGCAGATGAAAGGCACTGAGAAA
Mutant: G>T
chr8:128,413,305
TTGAGCTCAGCAGATGAAAGTCACTGAGAAA
c
GATA1 gain in α-globin cluster
WT
chr16:209,709
CCAAGCCTGTGCTGTTATTTTCTAATAATGTGGGTG
Mutant: A>G
CCAAGCCTGTGCTGTTATTTTCTGATAATGTGGGTG
d
GATA4 loss in BCL2 promoter
WT
CGTAAAGCCCTTGATAAACCCCTTCCCTGGA
2
Mutant: T>A
CGTAAAGCCCTTGAAAAACCCCTTCCCTGGA
chr18:60,988,353 (c.–2454)
e
RFX3 loss in alternative GPR56 promoter
WT
GCAGCTGCCAACGGTTGCCAAGGAACGGTTGCCAAGGCTGCTGTC
CGTCGACGGTTGCCAACGGTTGCCAACGGTCCGCGACGACAG
Opposite strand
chr16:57,673,301–57,673,315
Mutant: del15
GCAGCTGCCAACGGTTGCCAAGGCTGCTGTC
CGTCGACGGTTGCCAACGGTCCGCGACGACAG
Placental mammal conservation
E1m
57,672,750
57,673,750
Mutation would increase score
Mutation would have no effect
Mutation would decrease score
Strong interaction
No interaction
HGMD SNV
COSMIC SNV
# patients with mutation
PhastCons
Forward strand
Reverse strand
f
HBB promoter complex
TSS chr11:5,248,301
HOXD13
NF-YA
KLF
TGTAAGCAATAGATGGCTCTGCCCTGACTTTTATGCCCAGCCCTGGCTCCTGCCCTCCCTGCTCCTGGGAGTAGATTGGCCAACCCTAGGGTGTGGCTCCAC
g
F7 promoter
SP1
HNF4G
TSS chr13:113,760,102
FOXA1
AGGACGCCTGTGTCCTCCCCTCCCCATCCCTCTGTCACCCTTGGAGGCAGAGAACTTTGCCGTCAGTCCCATGGGAATGTCAACAGGCAGGGGCAGCAC
h
GABPA gains in TERT promoter
WT
TSS chr5:1,295,162 (c.–58)
(G228A)
chr5:1,295,228
(G250A)
chr5:1,295,250
GCAGCGCTGCCTGAAACTCGCGCCGCGAGGAGAGGGCGGGGCCGCGGAAAGGAAGGGGAGGGGCTGGGAGGGCCCGGAGGGGGCTGGGCCGGGGACCCGGGAGGGGTCGGG
4
2
2
211
11
16
79
11
GABPA
Mut: T>G
GCAGCGCTGCCGGAAACTCGCGCCGCGAGGAGA
Mut: G>A
GGGAGGGCCCGGAAGGGGCTGG
Mut: G>A
CGGGGACCCGGAAGGGGTCGGG

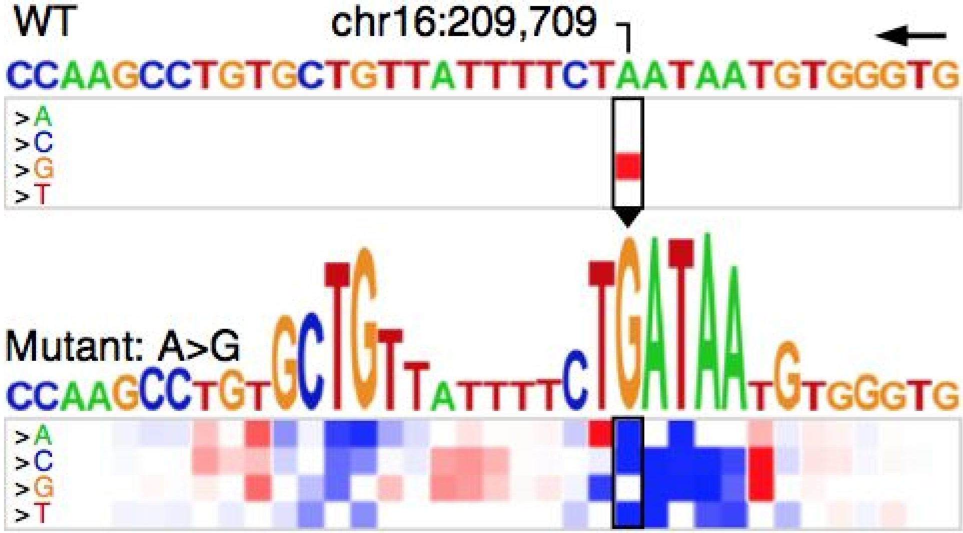
Alipanahi, et al., *Nature Biotechnology,* 2015.

Mutation in MYC Enhancer Weakens TCF7L2 Binding Site


WT
TTGAGCTCAGCAGATGAAAGGCACTGAGAAA
Mutant: G>T
chr8:128,413,305
TTGAGCTCAGCAGATGAAAGTCACTGAGAAA

Alipanahi, et al., *Nature Biotechnology,* 2015.

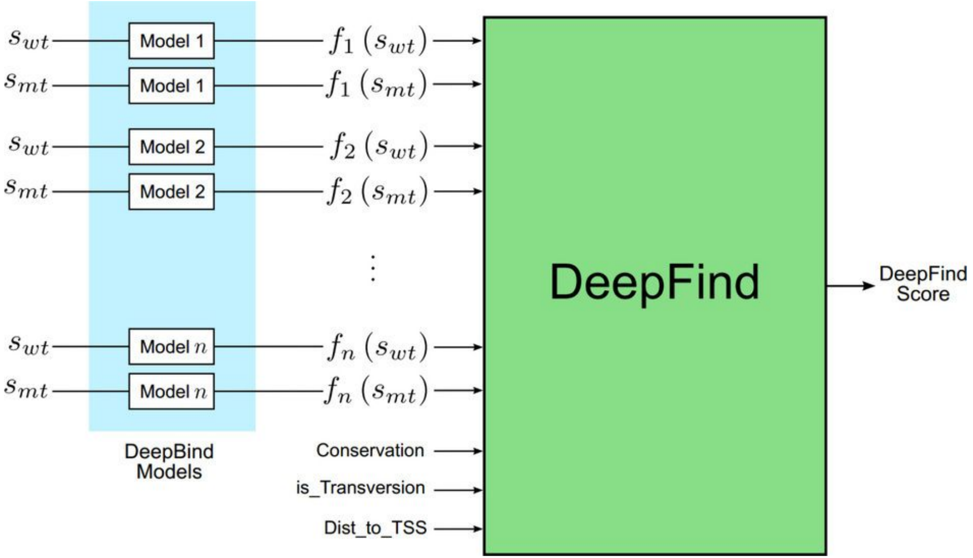
SNP in Globin Cluster Creates GATA1 Binding Site


WT
chr16:209,709
CCAAGCCTGTGCTGTTATTTTCTAATAATGTGGGTG
>A
>C
>G
>T
Mutant: A>G
CCAAGCCTGTGCTGTTATTTTCTGATAATGTGGGTG
>A
>C
>G
>T

Alipanahi, et al., *Nature Biotechnology,* 2015.

DeepFind: an aggregate model

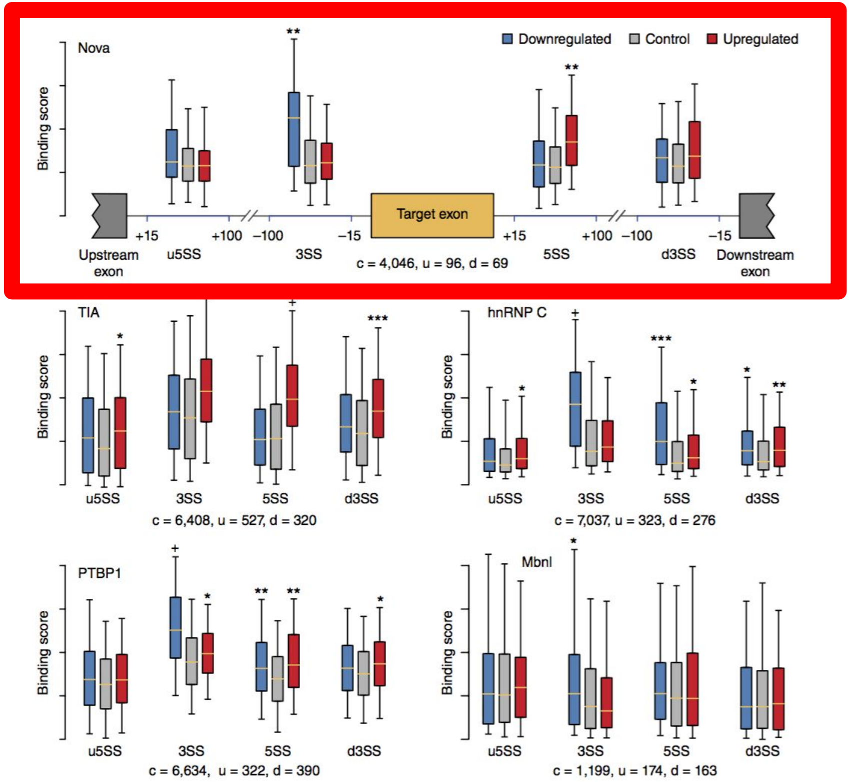
- What's the point? To provide collective contexts.
- I.e., true TF binding sites are likely to be located with other TF binding sites
- AUC ~ 0.76
- Predicts deleterious SNVs in promoters



Alipanahi, et al., *Nature Biotechnology,* 2015.

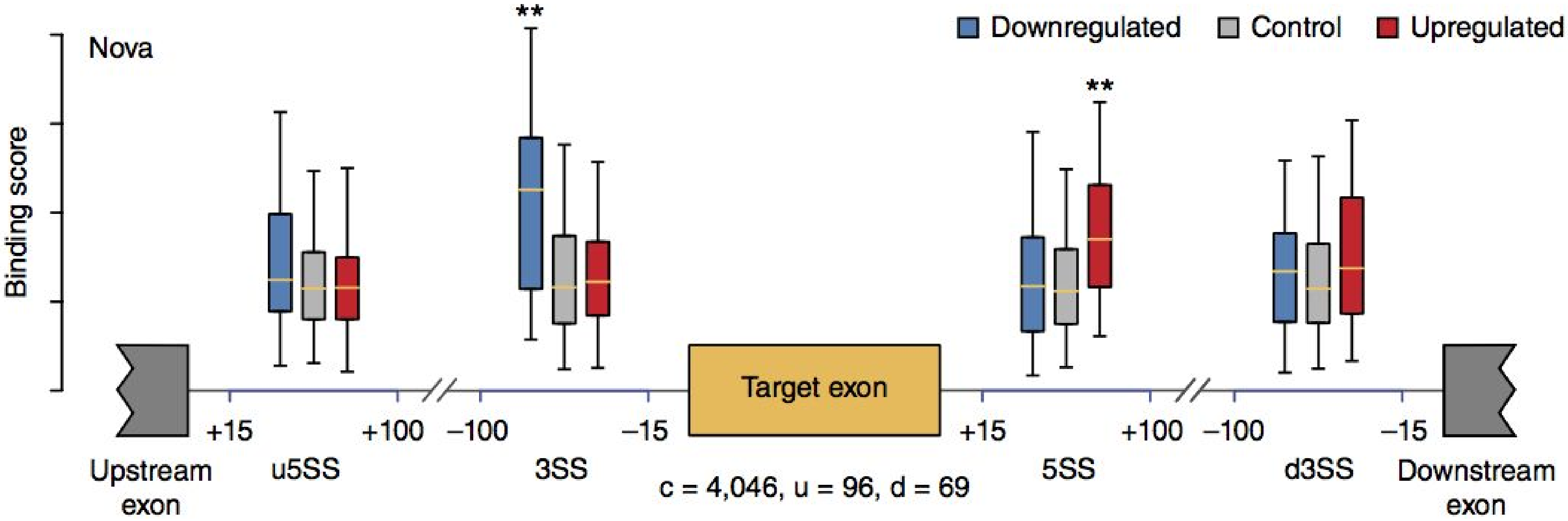
One more application: Alternative Splicing

- AS generates transcriptional diversity
- RBPs regulate splicing
- Binding scores at exon junctions regulated by splicing regulators
- Consistent with experimental CLIP-seq data and known binding profiles of RBP’s

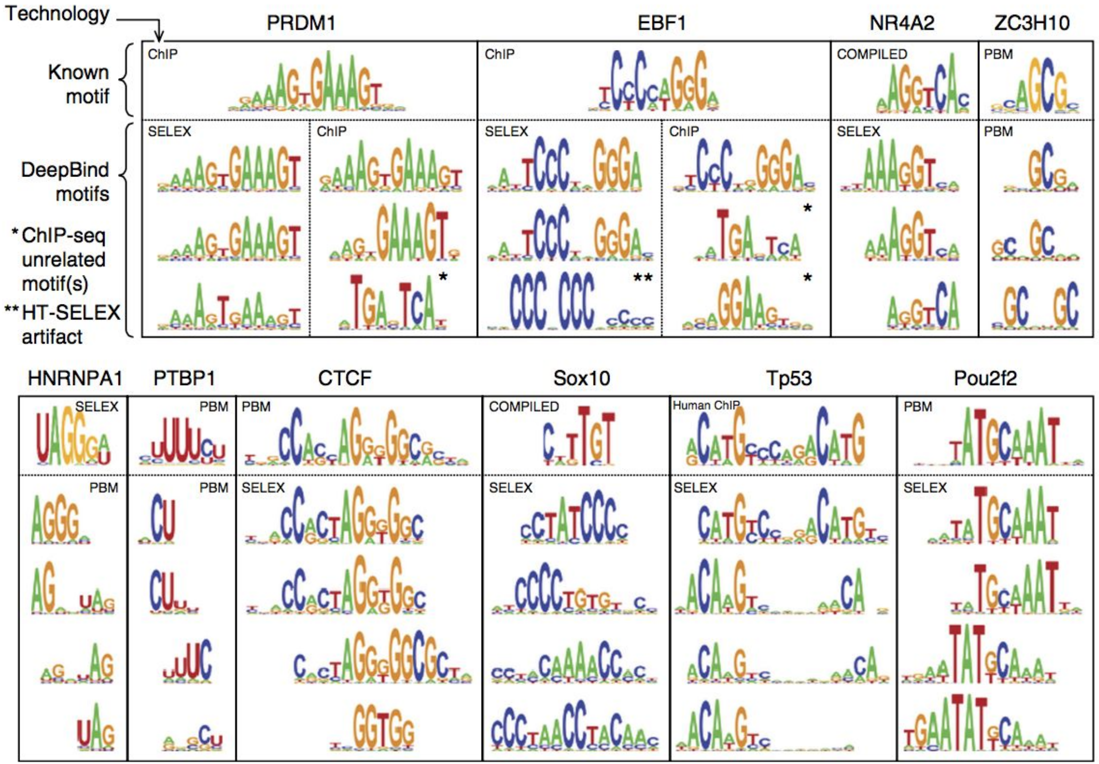


Alipanahi, et al., *Nature Biotechnology*, 2015.

Prediction of Nova Regulation Mechanism



DeepBind Motif Learning



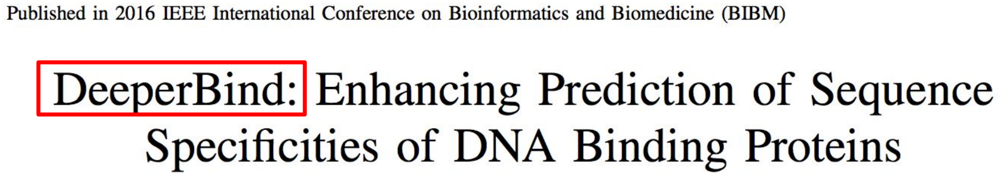
Key Takeaways

- **GOAL:** given regions experimentally determined to be bound by proteins, what is the model describing bound sequences?
- Sequences/Binding Scores -> CNN -> binding scores for novel sequences
- Generates weighted ensembles of PWM's and mutational maps
- ~600 different DeepBind models generated
- Identified RNA-binding sites involved in splicing regulation
- Identified disease-associated variants that affect TF binding

CHECK IT OUT YOURSELF: http://tools.genes.toronto.edu/deepbind/

Shortcomings and Future Work

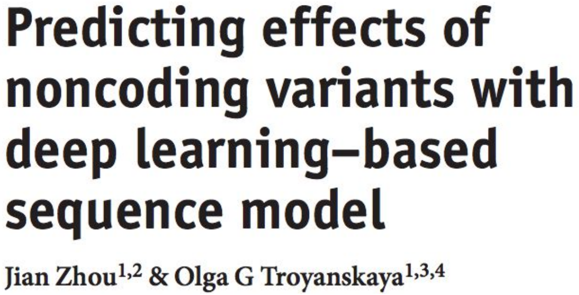
- Comparisons with only non-deep learning models
- Not much better than non-deep learning models
- Assumes one motif in each probe
- Non-coding factors/variants ignored
- Does not account for positional dynamics of probe sequences -> DeeperBind
- How about epigenetic regulation of binding to sequences? -> DeepSEA



Published in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins

Cool name bro



Predicting effects of noncoding variants with deep learning–based sequence model

Jian Zhou[1,2] & Olga G Troyanskaya[1,3,4]

Any Questions?