Decoding the regulatory genome
Transcription factors implement genomic regulation
What are the gene regulators that control gene expression?
At what genes do these regulators operate?
DNA-protein binding is essential to cellular function
Protein molecules that bind to specific DNA sequences and act as molecular switches to turn genes on or off.

Humans have ~2000 transcription factors.
Protein molecules that bind to specific DNA sequences and act as molecular switches to turn genes on or off.

Humans have ~2000 transcription factors.
Combinatorial control lies at the heart of the complexity and diversity of eukaryotes

(Molecular biology of the gene, 6ed)
Transcriptional regulatory network information will:

• reveal how cellular processes are connected and coordinated

• suggest new strategies to manipulate phenotypes and combat disease
Environment-Specific Regulator Behaviors

<table>
<thead>
<tr>
<th>Global Behavior</th>
<th>Selected Regulator-Gene Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Condition Invariant</strong> (e.g. Leu3)</td>
<td>Environment 1: Leu3&lt;sub&gt;LEU4&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Leu3&lt;sub&gt;ILV2&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Leu3&lt;sub&gt;ILV5&lt;/sub&gt;</td>
</tr>
<tr>
<td><strong>Condition Enabled</strong> (e.g. Msn2)</td>
<td>Environment 1: Msn2&lt;sub&gt;GSY2&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Msn2&lt;sub&gt;CCP1&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Msn2&lt;sub&gt;GPH1&lt;/sub&gt;</td>
</tr>
<tr>
<td><strong>Condition Expanded</strong> (e.g. Gcn4)</td>
<td>Environment 1: Gcn4&lt;sub&gt;ASN1&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Gcn4&lt;sub&gt;ARG8&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Gcn4&lt;sub&gt;BNA1&lt;/sub&gt;</td>
</tr>
<tr>
<td><strong>Condition Altered</strong> (e.g. Ste12)</td>
<td>Environment 1: Ste12&lt;sub&gt;AGA1&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Ste12&lt;sub&gt;STE12&lt;/sub&gt;</td>
</tr>
<tr>
<td></td>
<td>Environment 1: Ste12&lt;sub&gt;MSB2&lt;/sub&gt;</td>
</tr>
</tbody>
</table>
Sample of the Yeast Draft Transcriptional Regulatory Code (~150 regulators)

Chromosome II
Positions 370000:379300

Chromosome IV
Positions 1358800:1366600

Chromosome V
Positions 1359000:1366000
The ENCODE project

• After the Human Genome Project, ENCODE was established (the DECODE name was taken...)
• The National Institutes of Health (NIH) funded project to characterize the function of the human genome
• ~$400M cost, > 32 labs, > 440 scientists
• Mostly based upon the technologies from Tuesday and Today
Discovering the genomic regulatory code
Need you to think critically about claimed results

- How replicable?
- How interpretable?
- What could go wrong?
- At what spatial resolution? (1bp? 50bp? 500bp?)
- How generalizable from “typical best” example to whole genome function
Chromatin Immunoprecipitation (ChIP) sequencing (ChIP-seq) reveals genome-protein interactions
Short Read Mapping

Aligning the short reads to a genome lets us determine their origin.

Short Read Alignment Task

• Given a reference and a set of reads, report at least one “good” local alignment for each read to a genome if the alignment exists

• Reads may match in multiple places (“multimapping”)

• What is “good”? We concentrate on:
  – Fewer mismatches is better
  – Failing to align a low-quality base is better than failing to align a high-quality base

Where are the Oct4 binding events in these data? (x axis genome location, y axis read count observed) (black bar = 1000bp)
Repetitive “blacklisted” regions are typically not considered and are gaps in our knowledge of genomic function.
mES cell Oct4 ChIP Seq displays distinct binding events
The read spatial distribution can be learned.
A mixture model can describe which genomic locations are bound

\[
p(R|\pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n|m)
\]

\[
\sum_{m=1}^{M} \pi_m = 1
\]
EM – no prior
Motif-based positional prior biases the binding event prediction

**Mixture model**

- Possible events: $M$
- Observed reads: $N$

$$p(R \mid \pi) = \prod_{n=1}^{N} \sum_{m=1}^{M} \pi_m p(r_n \mid j), \quad \sum_{m=1}^{M} \pi_m = 1$$

**Position specific priors**

- Events are sparse
- Events occurs more likely at motif positions

$$p(\pi) \propto \prod_{m=1}^{M} (\pi_m)^{-\alpha_s + \alpha_m}$$

$\alpha_s$: uniform sparse prior parameter governing the degree of sparseness, $\alpha_s > 0$;
$\alpha_m$: position specific motif-based prior
Mixture model of protein-genome interaction

- $\pi_m$ is fraction of all reads $N$ produced by location $m$
- $M$ is the total number of locations
- $\sum_{m=1}^{M} \pi_m = 1$
- $\gamma_{mn}$ is fractional responsibility of location $m$ for read $n$
- $N_m$ is number of reads produced by location $m$
- $p(r_n|m)$ is probability location $m$ produced read $r_n$ using our shear distribution
Mixture model of protein-genome interaction

- $\pi_m$ is fraction of all reads $N$ produced by location $m$
- $M$ is the total number of locations
- $\sum_{m=1}^{M} \pi_m = 1$
- $\gamma_{mn}$ is fractional responsibility of location $m$ for read $n$
- $N_m$ is number of reads produced by location $m$
- $p(r_n | m)$ is probability location $m$ produced read $r_n$ using our shear distribution

E Step (Step 1)

$$\gamma_{mn} = \frac{\pi_m p(r_n | m)}{\sum_{j=1}^{M} \pi_j p(r_n | j)}$$

M Step (Step 2, then iterate)

$$N_m = \sum_{n=1}^{N} \gamma_{mn}$$

$$\pi_m = \frac{\max(0, N_m - \alpha_s + \alpha_m)}{\sum_{j=1}^{M} \max(0, N_j - \alpha_s + \alpha_m)}$$

$max$ performs component elimination to simplify solution
EM – Sparse prior
Genome-wide Event finding and Motif discovery

ChIP-Seq Reads → DNA Sequences

Biases binding event predictions towards motif positions

Event finding → Motif discovery

G E M

1. Identifies the enriched motifs proximal to binding sites

2. Binding events and explanatory DNA motifs
Sequence logos describe what is bound

\[ S_{b,i} = \]

Logo height of base \( b \) at position \( i \)

\[ f_{b,i} = \text{Fraction of base } b \text{ at position } i \]

\[ I_i = 2 + \sum_{b \in \{A,C,G,T\}} f_{b,i} \log_2 f_{b,i} \]

\[ S_{b,i} = f_{b,i} I_i \]
GEM improves spatial accuracy in binding event prediction
GEM improves the spatial accuracy in resolving proximal binding events.
GEM motif discovery outperforms other methods when detecting motifs in ChIP-Seq data

![Bar chart showing performance comparison of different motif discovery methods across different numbers of motifs. The chart includes GEM, MEME, Weeder, POSMO, ChIPMunk, MDscan, and HMS. The x-axis represents the number of motifs (Top 1 to Top 6), and the y-axis represents the number of successful experiments. The chart indicates that GEM outperforms other methods across all numbers of motifs.]
What discovered genome binding events are significant?
How can we compute the significance of an event in the context of background noise?
Computing the significance of an event

- $N_e$ is the number of reads assigned to an event
- $N_c$ is the scaled number of reads in the same genomic region in the control channel
- $N = N_e + N_c$
- For our null we use the Cumulative Distribution Function for the Binomial distribution with $P = 0.5$ This assumes that reads go to either the IP or control channel with equal probability
- We will need to do multiple hypothesis correction on the total number of events found using the Benjamani-Hochberg FDR method (Lecture 1)

$$p_{value} = \sum_{i=0}^{N_c} \binom{N}{i} P^i (1 - P)^{N-i}$$
The Irreproducible Data Rate (IDR) identifies significant events by consistent replicate ranks

https://biodatascience.github.io/compbio/test/IDR.html
The Irreproducible Data Rate (IDR) fits two mixture components to a rank correspondence function

\[ \Psi \] is % of pairs ranked in upper t% of R1 and R2

Derivative \( \Psi' \) shows decay point between components

Idea – construct mixture model containing two components to model the ranks of the replicates, reproducible and irreproducible.

The IDR of an event is the probability that it belongs to the irreproducible component

The Irreproducible Data Rate (IDR) identifies significant events by consistent replicate ranks

https://ccg.epfl.ch/var/sib_april15/cases/landt12/idr.html
Chance of a motif occurring $x$ times in bound set by chance (hypergeometric)

\[
\binom{N}{x} = \frac{N!}{(N-x)!x!}
\]

\[
p(x) = \frac{\binom{B}{x} \binom{N-B}{S-x}}{\binom{N}{S}}
\]

\[
p_{value}(x) = \sum_{i=x}^{\min(S,B)} p(i)
\]
Transcription Factors Interact
The spatial arrangement of transcription factor binding is critical

The IFN-β enhanceosome

(Vo and Goodman, JBC, 2001)

(Panne, Cell, 2007)

• Single point mutations disable the enhancer
• No major protein-protein interaction
A precise genome wide characterization of *in vivo* spacing constraints between key transcription factors would reveal key aspects of the gene regulation.
GEM reveals transcription factor spatial binding constraints
What are significant spacings?

- Compute average number of motifs in background region [200bp 400bp] and [-400bp -200bp]
- Use Poisson CDF to compute p-value of number of occurrences at each location in [-100bp 100bp]
- Bonferroni correct each p-value (p-value x 201)
- We choose that a corrected p-value is significant at $10^{-8}$
Spatial binding constraints detected from ENCODE ChIP-Seq datasets

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Description</th>
<th>Expts</th>
<th>Constraint pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>K562</td>
<td>leukemia</td>
<td>38</td>
<td>154</td>
</tr>
<tr>
<td>GM12878</td>
<td>lymphoblastoid</td>
<td>29</td>
<td>134</td>
</tr>
<tr>
<td>HepG2</td>
<td>liver carcinoma</td>
<td>21</td>
<td>86</td>
</tr>
<tr>
<td>HeLa-S3</td>
<td>cervical carcinoma</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>H1-hESC</td>
<td>embryonic stem cells</td>
<td>11</td>
<td>19</td>
</tr>
</tbody>
</table>

(355 distinct TF pairs)
Spatial binding constraints detected from ENCODE ChIP-Seq datasets

(PDB: 1FOS)

(PDB: 1NKP)
Cooperative binding: c-Jun/USF-1
Competitive binding: CTCF/Egr-1

Egr-1

PDB:1ZAA

(315)
Collaborative binding: HNF4-α/FOXA1

![Collaborative binding diagram](image)
Binding is not necessarily evolutionarily conserved
Is conservation a good predictor of conserved binding events across species?
Promoter proximal binding is not well conserved in liver (FOXA2, HNF1A, HNF4A, HNF6)

- No Binding: 2160 (54%)
- All Events Conserved: 132 (3%)
- Some Events Conserved: 320 (8%)

- No Events Conserved: 740 (19%)
- No Events Conserved: 536 (13%)
- No Events Conserved: 144 (4%)

D. Odom, R. Dowell      E. Fraenkel, D. Gifford Labs
Nature Genetics, 2007
Evaluating CNNs for binding prediction – what makes a good architecture?
Traditional DNA-protein binding models
One possible learned network structure

<table>
<thead>
<tr>
<th>Input</th>
<th>Lower layers</th>
<th>Middle layers</th>
<th>Higher layers</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motifs</td>
<td>Combination of motifs</td>
<td>Regulatory grammar</td>
<td>Bound/Unbound</td>
<td></td>
</tr>
</tbody>
</table>
CNNs can outperform conventional approaches in modeling DNA-protein binding

Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning
Babak Alipanahi¹,²,₆, Andrew Delong¹,₆, Matthew T Weirauch³⁻⁵ & Brendan J Frey¹⁻³

DeepBind (2015):
One convolutional layer with 16 kernels, maximum pooling window
DeepBind is “shallow learning” compared with other CNNs

Open questions about deep learning for genomics

• What architectures work best to model DNA-protein binding?
• How “deep” should a network be?
• What components of the network contribute most to overall performance?
• Is the optimum network design specific to the task / experiment / TF?
Today’s approach

• Use a framework to systematically benchmark CNN architectures on genomics tasks
• Analyze the contribution of different network components
• Explore if the optimum architecture is task-specific
• Evaluate training data requirements
Systematic benchmarking is important

• Task should be meaningful
  – Real vs. artificial sequences (DeepBind): motif discovery
    • Simple
    • Learn motif from similar nucleotide background
    • Not generalizable to classify real bound sequences
Systematic benchmarking is important

• Task should be meaningful
  – *Real vs. artificial sequences (DeepBind): motif discovery*
  – *Bound motif vs. unbound motif: motif occupancy*
    • Hard
      • Forces the model to learn better and higher-level sequence determinants
Systematic benchmarking is important

- Task should be meaningful
- Balance the number of positive and negative samples
- Control any artificial bias, location of the motif in the sample
- Conclusion should be the consensus across diverse TF ChIP-seq experiments (we use 690 from ENCODE)
CNNs have three important architectural dimensions to vary:

- More layers can capture higher level features.
- More convolution kernels better capture feature diversity.
- Smaller pooling window retains more location information.

More convolution kernels better capture feature diversity.

Smaller pooling window retains more location information.

More layers can capture higher level features.
## CNN architectures compared

<table>
<thead>
<tr>
<th>Our Name</th>
<th>More Conv. Kernels</th>
<th>Deeper</th>
<th>Smaller pooling size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1layer (DeepBind)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1layer_64motif</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1layer_128motif</td>
<td>✓✓✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1layer_local_win9</td>
<td>-</td>
<td>-</td>
<td>✓✓✓</td>
</tr>
<tr>
<td>1layer_local_win3</td>
<td>-</td>
<td>-</td>
<td>✓✓✓</td>
</tr>
<tr>
<td>2layer</td>
<td>-</td>
<td>✓✓✓</td>
<td>-</td>
</tr>
<tr>
<td>3layer</td>
<td>-</td>
<td>✓✓✓</td>
<td>-</td>
</tr>
<tr>
<td>2layer_local_win3</td>
<td>-</td>
<td>✓✓✓</td>
<td>✓✓✓</td>
</tr>
<tr>
<td>3layer_local_win3</td>
<td>-</td>
<td>✓✓✓</td>
<td>✓✓✓</td>
</tr>
</tbody>
</table>
Baseline model reproduces DeepBind
Simple models are best for a motif discovery task

- More convolutional kernels help model motif diversity
- Smaller pooling size, more layers monotonically decrease performance
  - possibly because most determinants are low-level (motifs) and position-independent
Depth improves performance in a **motif occupancy task**

- AUC decreases for all architectures
- More convolutional kernels help model the motif diversity
- Smaller pooling size slightly decreases the performance
- Deeper networks have slightly better performance
  - There are more high-level determinants that can be better modeled by deeper layers, consistent with the task design
Observed performance is experiment-specific

<table>
<thead>
<tr>
<th>Performance relative to 1layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>More layers/smaller pooling window</td>
</tr>
<tr>
<td>Simple structures</td>
</tr>
</tbody>
</table>

- **Small** variance in performance
- **Huge** variance in performance

ChIP-seq experiments
More complex networks require more training data.
Variance increases with fewer training examples

Performance on motif discovery task

- Baseline
- + loc. info
- + layers

80,000 training examples  20,000 training examples  5,000 training examples
CNNs can outperform conventional methods

- CNNs outperform conventional methods with the right structure
- The optimum structure is different from that in computer vision
- Different biological tasks and data yield different conclusions
- Understanding the problem at hand and comparing different structures is important to design a good CNN model for biology applications (http://cnn.csail.mit.edu)
"Pioneer Factors" can have directional effects
**PIQ: algorithm to predictively model TF binding from DNase-seq + Sequence**

**Input:**
Scan 1331 motifs

**Modeling:**

**Predictions:**
Pioneer TFs have identifiable profiles

Computational pioneers

Per-base Chromatin Opening Index

KAISO Nrf1 Zfp161 NFYA

Computational non-pioneers

Per-base Chromatin Opening Index

E2F Klf7 ETS CREB

Oct4 RXR:RAR DR5 FoxA Myc

Distance from Motif (+/-200 bp)
In vitro reporter assays recapitulate computational predictions

Using a Tol2 based GFP reporter, we confirm finding that these pioneers create new enhancers.
Pioneers appear to be conserved between human/mouse
Pioneer TFs have identifiable profiles

Computational pioneers

KAISO  Nrf1  Zfp161  NFYA

E2F  Klf7  ETS  CREB

Computational non-pioneers

Oct4  RXR:RAR DR5  FoxA  Myc

Distance from Motif (+/-200 bp)
Certain pioneer TFs are directional

- We define asymmetry index as the expected change between left and right sides in (squared) chromatin opening index score

- Biological validation by testing both motif orientations
Certain pioneer TFs are directional

Orienting the motif direction in the reporter recapitulates expected directional behaviors.
Settler factors follow pioneer factor binding and loss of pioneer binding causes chromatin to return to a closed state.

Pioneers (chromatin opening and dependent) are rare and distinct, while there exists a class of chromatin dependent, but non-opening factors.
Loss of pioneer binding causes chromatin to return to closed states.
Validate pioneer/settler model via dominant-negative competition assay

- Construct pioneer DBD protein that retains no pioneering function
- Induction of DBD protein competes for genomic binding, reducing local chromatin accessibility settlers rely on
- Compare proximal chromatin openness
- Compare ChIP levels for neighboring settler binding

![Diagram showing competition between pioneer and settler binding sites.](image-url)
Dominant negative pioneers reduce proximal DNase HS

We created dominant negative versions of the NFYA and Nrf1 pioneers and measured DNase accessibility at native NFYA and Nrf1 sites after induction of dominant negatives.
Dominant negative pioneers reduce proximal binding of c-Myc
Predicting chromatin accessibility
How genome sequence determines cell-type specific chromatin accessibility

Can we predict chromatin accessibility directly from DNA sequence?

DNase-seq data across a 100 kilobase window
(Chromosome 14 K562 cells)

Motivation –

1. Understand the fundamental biology of chromatin accessibility
2. Predict how genomic variants change chromatin accessibility
Can we discover DNA “code words” encoding chromatin accessibility?

- The DNA “code words” encoding chromatin accessibility can be represented by k-mers (k <= 8)
- K-mers affect chromatin accessibility locally within +/- 1 kb with a fixed spatial profile
- A particular k-mer produces the same effect wherever it occurs
Chromatin accessibility arises from interactions, largely among pioneer TFs.
The Synergistic Chromatin Model (SCM) is a K-mer model

~40,000 K-mers in model
~5,000,000 parameters
543 iterations * 360 seconds / iteration * 40 cores
= ~ 90 days
Training on K562 DNase-seq data from chromosomes 1 – 13 predicts chromosome 14 (black line)

KMM $R^2$ 0.80

Control $R^2$ 0.47
SCM predicts accessibility data from a NRF1 binding site.
SCM outperforms contemporary models at predicting chromatin accessibility from sequence (K562)
Accessibility contains cell type specific and cell type independent components (11 cell types, Chr 15-22)
SCM models have similar predictive power for other cell types

- Cerebellum
- Frontal cerebrum
- Epithelial adenocarcinoma
- Fibroblast GM03348
- Fibroblast AG20443
- Frontal cortex
- GM10248
- GM12878
- H7ES
- HEK293T

Correlation on held out data
SCM model trained on ES data performs better on shared DNase hot spots (Chr 15 – 22)
We created synthetic “phrases” each of which contains k-mers that are similar in chromatin opening score.

Claim 3 – CCM Models are accurate for synthetic sequences
Single Locus Oligonucleotide Transfer
>6,000 designed phrases into a chromosomal locus

<table>
<thead>
<tr>
<th>Heterochromatin locus</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>% alleles with phrase integration</td>
<td>35</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td># unique integrations</td>
<td>350,000</td>
<td>150,000</td>
<td>150,000</td>
</tr>
</tbody>
</table>
Predicted accessibility matches measured accessibility
FIN