Genome 541, unit 3
Gene regulation and epigenomics
The central dogma
This unit

• Lecture 1: Transcription factor binding (sequence)
• Lecture 2: Transcription factor binding (genomics assays)
• Lecture 3: Integrating genomics assays
• Lecture 4: Chromatin conformation
Methods focus: the generative framework
Genome 541
Unit 3, lecture 1
Transcription factor binding using sequence
This lecture

- Sequence models of transcription factor binding
- Method: How to evaluate accuracy of a model and use it to make predictions
- Method: The EM algorithm for sequence motifs
- Applications of sequence models
Sequence models of transcription factor binding
  • Method: How to evaluate accuracy of a model and use it to make predictions
  • Method: The EM algorithm for sequence motifs
  • Applications of sequence models
Motifs
Motif (n): a succession of notes that has some special importance in or is characteristic of a composition
Sequence-specific transcription factors drive gene regulation

*Motif (n)*: a recurring genomic sequence pattern
The most common model of sequence motifs is the position-specific frequency matrix (PSFM)

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 \\
A & 0.1 & 0.95 & \ldots \\
C & 0.0 & 0.05 \\
G & 0.8 & 0.0 \\
T & 0.1 & 0.0 \\
\end{array}
\]
Motivating question: How accurately does a PSFM predict the binding of a given transcription factor?
This lecture

- Sequence models of transcription factor binding
  Method: How to evaluate accuracy of a model and use it to make predictions
- Method: The EM algorithm for sequence motifs
- Applications of sequence models
JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!

BUT WE'RE PLAYING MINECRAFT!
...Fine.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!

BUT MIYNECRAFT!

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN GREEN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND A LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

---

News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE

ONLY 5% CHANCE OF COINCIDENCE!

Scientists...
Evaluating accuracy

- Evaluating multiple predictions
- Trading off sensitivity and specificity
There are three main measures of accuracy

<table>
<thead>
<tr>
<th></th>
<th>TF binds</th>
<th>TF doesn’t bind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted TF binding</td>
<td>True positive (TP)</td>
<td>False positive (FP)</td>
</tr>
<tr>
<td>Predicted no TF binding</td>
<td>False negative (FN)</td>
<td>True negative (TN)</td>
</tr>
</tbody>
</table>

- **Recall** = $\frac{TP}{TP + FN}$
- **Precision** = $\frac{TP}{TP + FP}$
- **Specificity** = $\frac{TN}{TN + FP}$
Accuracy-related measures have a lot of synonyms

<table>
<thead>
<tr>
<th>Formula</th>
<th>Names</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP / (TP + FN)</td>
<td>Recall/Detection rate/Specificity/(1 - FPR)</td>
</tr>
<tr>
<td></td>
<td>Sensitivity/True positive rate/Power</td>
</tr>
<tr>
<td></td>
<td>1 - False negative rate</td>
</tr>
<tr>
<td>TP / (TP + FP)</td>
<td>Precision/Positive predictive value/1 - FDR</td>
</tr>
<tr>
<td></td>
<td>1 - False discovery rate</td>
</tr>
<tr>
<td>TN / (TN + FP)</td>
<td>Specificity/True negative rate/1 - FPR</td>
</tr>
<tr>
<td></td>
<td>1 - False positive rate</td>
</tr>
<tr>
<td></td>
<td>1 - Statistical significance level</td>
</tr>
</tbody>
</table>
Evaluating accuracy

- Evaluating accuracy
- Evaluating multiple predictions
- Trading off sensitivity and specificity
There are two ways to control the accuracy of multiple predictions

- Family-wise error rate: The probability of getting one false positive.
- False discovery rate: Fraction of positive predictions that are false
There are two ways to control the accuracy of multiple predictions

- Family-wise error rate: The probability of getting one false positive. Bonferroni correction:

\[
\alpha \leftarrow \frac{\alpha_0}{m}
\]

or

\[
p \leftarrow mp_0
\]

- False positive rate: Fraction of positive predictions that are false. Benjamini-Hochberg procedure:
There are two ways to control the accuracy of multiple predictions

- Family-wise error rate: The probability of getting one false positive. Bonferroni correction:

\[
\alpha \leftarrow \frac{\alpha_0}{m} \quad \text{or} \quad p \leftarrow mp_0
\]

- False positive rate: Fraction of positive predictions that are false. Benjamini-Hochberg procedure:
There are two ways to control the accuracy of multiple predictions

• Family-wise error rate: The probability of getting one false positive. Bonferroni correction:

$$\alpha \leftarrow \frac{\alpha_0}{m} \quad \text{or} \quad p \leftarrow m p_0$$

• False discovery rate: Fraction of positive predictions that are false. Benjamini-Hochberg procedure:

$$\hat{j}^* = \max \left\{ j : p_j \leq \frac{j}{m} \alpha \right\}$$

Number of accepted tests

Ordered p-values
There are two ways to control the accuracy of multiple predictions:

1. **Bonferroni correction**

2. **$\alpha$ threshold**

In the diagram:

- The $p$-value is plotted against the $P$-value rank.
- A Bonferroni correction threshold is indicated at $\alpha = 0.03$.
- An $\alpha$ threshold at $\alpha_0 = 0.3$ is also shown.
There are two ways to control the accuracy of multiple predictions.
There are two ways to control the accuracy of multiple predictions:

1. **p-value rank**
2. **q-value**

- **q-value**: The minimum false-discovery rate threshold at which a test is called significant.

Given: 
- $\alpha_0 = 0.3$
- $q = 0.5$
Evaluating accuracy

• Evaluating accuracy
• Evaluating multiple predictions
Trading off sensitivity and specificity
A Receiver-Operator Characteristic (ROC) curve plots recall against specificity.
A Precision-Recall (PR) curve plots recall against precision.
There are several measures that summarize multiple measures of accuracy

• auROC: Area under the ROC curve
• auPR: Area under the PR curve
• F1 measure = 2*(precision*recall)/(precision+recall)
• “accuracy” = (TP+TN) / (TP+TN+FP+FN)
• Recall at a given FDR
• Recall at a given FPR
How do you handle ties in prediction probability scores?
How do you handle ties in prediction probability scores?
Interpolation in an ROC curve forms a straight line.

Recall

False positive rate

A

B

A+B/2

C
Interpolation in an PR curve forms convex curve
This lecture

- Sequence models of transcription factor binding
- Method: How to evaluate accuracy of a model and use it to make predictions
- Method: The EM algorithm for sequence motifs
- Applications of sequence models
The EM algorithm optimizes latent variable (missing data) models.
The EM algorithm optimizes latent variable (missing data) models

$$P(\text{sequence}|\text{motif starts at position } i, \text{PSFM}) = \prod_{j=i}^{i+L} \text{PSFM}(\text{sequence}_j, j - i) \cdot \prod_{j \notin [i, i+L]} \text{background}(\text{sequence}_j)$$
The EM algorithm optimizes latent variable (missing data) models

\[ P(\text{sequence}|\text{motif starts at position } i, \text{PSFM}) = \prod_{j=i}^{i+L} \text{PSFM}(\text{sequence}_j, j-i) \cdot \prod_{j \notin [i,i+L]} \text{background}(\text{sequence}_j) \]

Want to optimize:

\[ P(X|\theta) = \sum_Y P(X, Y|\theta) \]
The EM algorithm optimizes latent variable (missing data) models

\[
P(\text{sequence} | \text{motif starts at position } i, \text{PSFM}) = \prod_{j=i}^{i+L} \text{PSFM}(\text{sequence}_j, j - i) \cdot \prod_{j \notin [i, i+L]} \text{background}(\text{sequence}_j)
\]

Want to optimize:

\[
P(\mathbf{X} | \theta) = \sum_Y P(\mathbf{X}, \mathbf{Y} | \theta)
\]

Easier to optimize:

\[
\text{E-step: } Q(\mathbf{Y} | \mathbf{X}, \theta) \leftarrow P(\mathbf{Y} | \mathbf{X}, \theta)
\]

\[
\text{M-step: } \theta' \leftarrow \text{argmax}_{\theta'} E_{\mathbf{Y} \sim Q(\mathbf{Y} | \mathbf{X}, \theta)} [\log P(\mathbf{X} | \mathbf{Y}, \theta')]
\]
MEME

One-motif-per-sequence assumption:

\[ Y = \text{position of motif start} \]

```
TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT
```

Subsequence initialization:

```
TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT
```

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.17</td>
<td>0.5</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>4</td>
<td>0.17</td>
<td>0.17</td>
<td>0.5</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>
```
This lecture

- Sequence models of transcription factor binding
- Method: How to evaluate accuracy of a model and use it to make predictions
- Method: The EM algorithm for sequence motifs

Applications of sequence models
TF specificity is driven by co-binding and interactions with nucleosomes

Di-nucleotide PWMs model dependencies between adjacent positions

doi: 10.1534/genetics.112.138685
Shape features capture physical properties of DNA as a function of sequence

doi: 10.1093/nar/gkt437
Classifiers trained on large data sets cell type-specific binding predict very accurately.

CTCF ChIP-seq in Liver

TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT
ATGGCAGAATCACTTTAAAACGTGGCCCCACCACCGTGCACGCCGTGCATTGTCTGTGCATTGTTACTGCGAAATGACTCAACG
CACATCCAACGAATCACCTCACCTACGTGCATCTACACCTTAATCGTGCATCTCATCGCCGAAATGGCATGCTTTTTT

...
A multi-task approach shares representations between factors

Shared feature extraction

CTCF
classifier

NRSF
classifier

Binds

Doesn’t bind

Prediction

Intermediate representation

Input sequence

0.1 -4.2 5.8 ...

TCTCTCTCCACGGCTAATTAGGTGATCATG
A multi-task approach shares representations between factors

doi:10.1038/nmeth.3547
Transcription factors bind in cis-regulatory modules to achieve specificity
Transcription factors bind in cis-regulatory modules to achieve specificity.

Scan for motif co-occurrence.
Position preference provides evidence for motif function

*P. falciparum* cytoplasmic localization motif
Administrivia

- I will post lecture notes after class.

- HW5 out soon: Learning motif PSFMs with EM (due Friday 5/6).

- Please write a one-minute response before you leave.