Inferring Models of cis-Regulatory Modules using Information Theory

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2019
Colin Dewey
colin.dewey@wisc.edu
Overview

• Biological question
  – What is causing differential gene expression?

• Goal
  – Find regulatory motifs in the DNA sequence

• Solution
  – FIRE (Finding Informative Regulatory Elements)
Goals for Lecture

Key concepts:

• Entropy

• Mutual information (MI)

• Motif logos

• Using MI to identify cis-regulatory module elements
A Common Type of Question

What causes this set of yeast genes to be up-regulated in stress conditions?

Figure from Gasch et al., *Mol. Biol. Cell*, 2000
**cis-Regulatory Modules (CRMs)**

- Co-expressed genes are often controlled by specific configurations of binding sites
Information Theory Background

- **Problem**
  - Create a code to communicate information

- **Example**
  - Need to communicate the manufacturer of each bike
Information Theory Background

• Four types of bikes
• Possible code

<table>
<thead>
<tr>
<th>Type</th>
<th>code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trek</td>
<td>11</td>
</tr>
<tr>
<td>Specialized</td>
<td>10</td>
</tr>
<tr>
<td>Cervelo</td>
<td>01</td>
</tr>
<tr>
<td>Serotta</td>
<td>00</td>
</tr>
</tbody>
</table>

• Expected number of bits we have to communicate:
  2 bits/bike
**Information Theory Background**

- Can we do better?
- Yes, if the bike types aren’t equiprobable

<table>
<thead>
<tr>
<th>Type, probability</th>
<th># bits</th>
<th>code</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{Trek}) = 0.5$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$P(\text{Specialized}) = 0.25$</td>
<td>2</td>
<td>01</td>
</tr>
<tr>
<td>$P(\text{Cervelo}) = 0.125$</td>
<td>3</td>
<td>001</td>
</tr>
<tr>
<td>$P(\text{Serotta}) = 0.125$</td>
<td>3</td>
<td>000</td>
</tr>
</tbody>
</table>

- Optimal code uses $-\log_2 P(c)$ bits for event with probability $P(c)$
Information Theory Background

<table>
<thead>
<tr>
<th>Type, probability</th>
<th># bits</th>
<th>code</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{Trek}) = 0.5$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$P(\text{Specialized}) = 0.25$</td>
<td>2</td>
<td>01</td>
</tr>
<tr>
<td>$P(\text{Cervelo}) = 0.125$</td>
<td>3</td>
<td>001</td>
</tr>
<tr>
<td>$P(\text{Serotta}) = 0.125$</td>
<td>3</td>
<td>000</td>
</tr>
</tbody>
</table>

- Expected number of bits we have to communicate: 1.75 bits/bike

$$- \sum_{c=1}^{\lvert C \rvert} P(c) \log_2 P(c)$$
Entropy

- Entropy is a measure of uncertainty associated with a random variable.
- Can be interpreted as the expected number of bits required to communicate the value of the variable.

\[
H(C) = -\sum_{c=1}^{\left| C \right|} P(c) \log_2 P(c)
\]
How is entropy related to DNA sequences?
Sequence Logos

- Typically represent a binding site
- Height of each character $c$ is proportional to $P(c)$
Sequence Logos

- Height of logo at a given position determined by decrease in entropy (from maximum possible)

\[
H_{\text{max}} - H(C) = \log_2 N - \left( - \sum_c P(c) \log_2 P(c) \right)
\]

# of characters in alphabet

decrease in entropy
Mutual Information

- *Mutual information* quantifies how much knowing the value of one variable tells about the value of another.

\[
I(M;C) = H(M) - H(M \mid C)
\]

\[
= \sum_m \sum_c P(m,c) \log_2 \left( \frac{P(m,c)}{P(m)P(c)} \right)
\]
FIRE
Elemento et al., Molecular Cell 2007

- **Finding Informative Regulatory Elements** (FIRE)

- **Given** a set of sequences grouped into clusters

- **Find** motifs, and relationships, that have high *mutual information* with the clusters

- Applicable when sequences have continuous values instead of cluster labels
Mutual Information in FIRE

• We can compute the mutual information between a motif and the clusters as follows

\[
I(M; C) = \sum_{m=0}^{1} \sum_{c=1}^{|C|} P(m, c) \log_2 \frac{P(m, c)}{P(m)P(c)}
\]

\( m=0, 1 \) represent absence/presence of motif

\( c \) ranges over the cluster labels
Finding Motifs in FIRE

• Motifs are represented by regular expressions; initially each motif is represented by a strict $k$-mer (e.g. TCCGTAC)

1. Test all $k$-mers ($k=7$ by default) to see which have significant mutual information with the cluster label

2. Filter $k$-mers using a significance test to obtain motif seeds

3. Generalize each motif seed

4. Filter motifs using a significance test
Key Step in Generalizing a Motif in FIRE

- Randomly pick a position in the motif
- Generalize in all ways consistent with current value at position
- Score each by computing mutual information
- Retain the best generalization

A or G in this position

```
TCC[AG]TAC


TCC[ACGT]TAC  TCC[AGT]TAC  TCC[CGT]TAC
```
Generalizing a Motif in FIRE

given: $k$-mer, $n$

$best \leftarrow null$

repeat $n$ times

    motif $\leftarrow k$-mer

repeat

    motif $\leftarrow \text{GeneralizePosition}(motif)$  // shown on previous slide

until convergence (no improvement at any position)

if $\text{score}(motif) > \text{score}(best)$

    $best \leftarrow motif$

return: $best$
Generalizing a Motif in FIRE: Example

Figure from Elemento et al. *Molecular Cell* 2007
Avoiding Redundant Motifs

• Different seeds could converge to similar motifs

\[
\begin{align*}
\text{TCCGTAC} & \quad \rightarrow \quad \text{TCCCTAC} \\
\text{TCC[CG]TAC} & \quad \rightarrow \quad \text{TCC[CG]TAC}
\end{align*}
\]

• Use mutual information to test whether new motif is unique and contributes new information

\[
\frac{I(M;C \mid M')}{I(M;M')} > r
\]

\(M'\) previous motif \quad \(M\) new candidate motif \quad \(C\) expression clusters
Characterizing Predicted Motifs in FIRE

• Mutual information is also used to assess various properties of found motifs
  – orientation bias
  – position bias
  – interaction with another motif
Using MI to Determine Orientation Bias

\[ I(S; C) \]

- \( C \) indicates cluster
- \( S=1 \) indicates motif present on transcribed strand
- \( S=0 \) otherwise (not present or not on transcribed strand)

Also compute MI where \( S=1 \) indicates motif present on complementary strand
Using MI to Determine Position Bias

\[ I(P; O) \]

\( P \) ranges over position bins
\( O=0, 1 \) indicates whether or not the motif is over-represented in a sequence’s cluster

<table>
<thead>
<tr>
<th>( P )</th>
<th>( O )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Only sequences containing the motif are considered for this calculation.
Using MI to Determine Motif Interactions

$I(M_1;M_2)$

$M_1=0$, $1$ indicates whether or not a sequence has the motif and is in a cluster for which the motif is over-represented; similarly for $M_2$.
Motif Interactions Example

Yeast motif-motif interactions
White: positive association
Dark red: negative association
Blue box: DNA-DNA
Green box: DNA-RNA
Plus: spatial co-localization
Discussion of FIRE

• FIRE
  – mutual information used to identify motifs and relationships among them
  – motif search is based on generalizing informative $k$-mers

• Consider advantages and disadvantages of $k$-mers versus PWMs

• In contrast to many motif-finding approaches, FIRE takes advantage of *negative* sequences

• FIRE returns all informative motifs found