Inferring Models of cis-Regulatory Modules using Information Theory

BMICS 776
www.biostat.wisc.edu/bmi776/
Spring 2015
Colin Dewey
cdewey@biostat.wisc.edu
• 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

the key concepts to understand are the following

• Entropy

• Mutual information (MI)

• Motif logos

• Using MI to identify CRM elements
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>Serrota</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>$P($Trek$) = 0.5$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$P($Specialized$) = 0.25$</td>
<td>2</td>
<td>01</td>
</tr>
<tr>
<td>$P($Cervelo$) = 0.125$</td>
<td>3</td>
<td>001</td>
</tr>
<tr>
<td>$P($Serrota$) = 0.125$</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>$P(\text{Serrota}) = 0.125$</td>
<td>3</td>
<td>000</td>
</tr>
</tbody>
</table>

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

\[
- \sum_{c=1}^{\left| C \right|} P(c) \log_2 P(c)
\]
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

\[ H(C) = - \sum_{c=1}^{\left| C \right|} P(c) \log_2 P(c) \]

entropy function for binary variable
How is entropy related to DNA sequences?
Sequence Logos

- Typically represent a binding site

- Height of each character $c$ is proportional to $P(c)$
• 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
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

- **Given** a set of sequences grouped into clusters
- **Find** motifs, and relationships, that have high mutual information with the clusters
- (also can do this 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

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

$m=0, 1$ 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

3. generalize each $k$-mer into a motif

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
Generalizing a Motif in FIRE

given: $k$-mer, $n$

$best \leftarrow null$
repeat $n$ times
repeat $m$ times
    $motif \leftarrow k$-mer
repeat
    $motif \leftarrow $ GeneralizePosition($motif$)  // shown on previous slide
until convergence (no improvement at any position)
if score($motif$) > score($best$)
    $best \leftarrow motif$

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

Figure from Elemento et al. *Molecular Cell* 2007
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)

\[
\begin{array}{c|c}
C & S \\
\hline
0 & 0 \\
0 & 0 \\
0 & 1 \\
0 & 1 \\
1 & 0 \\
1 & 1 \\
1 & 0 \\
1 & 1 \\
2 & 1 \\
2 & 1 \\
2 & 1 \\
\end{array}
\]

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 overrepresented in a sequence’s cluster

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 overrepresented; similarly for \( M_2 \)

\[
\begin{array}{c|c|c|c|c|c|c|c|c|c}

& & & & & & & & & \\
\hline
5' upstream region & \( M_1 \) & \( M_2 \) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
& & & & & & & & & \\
\end{array}
\]
Discussion of CRM Finding Methods

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

- in contrast to many motif-finding approaches, both CRM methods take advantage of *negative* sequences

- FIRE returns all informative motifs/relationships found, whereas the Noto & Craven approach returns single discriminative model