Inferring Models of cis-Regulatory Modules using Information Theory

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2021
Daifeng Wang
daifeng.wang@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
Gene expression and regulation

Central dogma

Gene expression levels (e.g., values to quantify RNA abundances)

Gene function

Gene regulation: mechanisms controlling gene expression levels

Identical DNA but different gene expression

Gene regulation: mechanisms controlling gene expression levels
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
Co-expressed genes have similar functions in single species

A gene co-expression network (relationship) can reveal functional groupings
- Hierarchical clustering, K-means, Gaussian mixture model (GMM), Principal component analysis (PCA), ...

Yeast cell cycle

Eisen et al., PNAS, 1998.

Carlson et al., BMC Bioinformatics, 2006.
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

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

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

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

A | 0.1 | 0.3 | 0.1 | 0.2 | 0.2 | 0.4 | 0.3 | 0.1  
C | 0.5 | 0.2 | 0.1 | 0.1 | 0.6 | 0.1 | 0.2 | 0.7  
G | 0.2 | 0.2 | 0.6 | 0.5 | 0.1 | 0.2 | 0.2 | 0.1  
T | 0.2 | 0.3 | 0.2 | 0.2 | 0.1 | 0.3 | 0.3 | 0.1  

frequency logo  
information content logo
Sequence Logos

• Typically represent a binding site

• Frequency logo: Height of each character \( c \) is proportional to \( P(c) \)

• Information content logo: based on entropy \( (H) \) of a random variable \( (C) \) representing distribution of character states at each position
Sequence Logos

- Height of logo at a given position determined by decrease in entropy (from maximum possible); i.e., information content

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

# 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.

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

\[
= \sum_m \sum_c P(m, c) \log_2 \left( \frac{P(m, c)}{P(m)P(c)} \right)
\]
Correlation vs. Mutual information
**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
Significance test via randomization

- Given an empirical MI value for a motif, \( I \)

- Randomly shuffle cluster labels of genes (or other variables such as expression), and calculate MI

- Repeat shuffling \( N_r \) times and get \( N_r \) MI values

- Pseudo p-value = \( \frac{\text{sum}(I < N_r \text{ MI values})}{N_r} \) to see if it is less than a significance threshold (e.g., \( 1/N_r \))
  - Z-score = \( \frac{(I - \text{mean}(I_{\text{random}}))}{\text{sigma}(I_{\text{random}})} \)

Elemento et al., *Molecular Cell* 2007, Supplement
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, \(n\)

\[\begin{align*}
\text{best} & \leftarrow \text{null} \\
\text{repeat } n \text{ times} \\
& \quad \text{repeat } \text{until convergence (no improvement at any position)} \\
& \quad \quad \text{if score(motif)} > \text{score(best)} \\
& \quad \quad \quad \text{best} \leftarrow \text{motif}
\end{align*}\]

return: \text{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

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

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

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

\[M'\] previous motif \[M\] new candidate motif \[C\] 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

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 \( M_2 \)

<table>
<thead>
<tr>
<th>5' upstream region</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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
Mutual Information for Gene Networks

- Mutual information and conditional mutual information can also be useful for reconstructing biological networks.

- Build gene-gene network where edges indicate high MI in genes’ expression levels.

- Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE).
ARACNE

- Gaussian kernel estimator to estimate mutual information
  - No binning or histograms

- Data processing inequality
  - Prune indirect edges

Margolin et al. *BMC Bioinformatics* 2006