Discovering regulatory sequences from expression data

- Unsupervised clustering
- Information theory
- Optimization
- Non-parametric statistical testing
- Multiple testing
- Overfitting
Transcriptional and post-transcriptional regulation of gene expression
Transcription factor binding sites are ~6-12 bp-long
Genes regulated by the same TF will be co-expressed
Microarray experiment

Several microarray experiments (conditions, time points, treatments)

Microarray Experiments

Genes
Creating co-expression clusters

Unsupervised clustering approaches:
• K-means
• Self-organizing maps
• Hierarchical clustering
K-means clustering

Genes

0
1
2
3
4
Hierarchical clustering
Self-organizing map
Clusters of co-expressed genes

Microarray Experiments

Cluster Index

All Genes on array
A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types

Olivier Elemento,¹,²,³ Noam Slonim,¹,²,³,⁴ and Saeed Tavazoie¹,²,*
¹Lewis-Sigler Institute for Integrative Genomics
²Department of Molecular Biology
Princeton University, Princeton, NJ 08544, USA
³These authors contributed equally to this work.
⁴Present address: IBM Haifa Research Labs, Haifa 31905, Israel.
*Correspondence: tavazoie@genomics.princeton.edu
DOI 10.1016/j.molcel.2007.09.027

SUMMARY

Deciphering the noncoding regulatory genome has proved a formidable challenge. Despite the wealth of available gene expression data, there currently exists no broadly applicable method for characterizing the regulatory elements that shape the rich underlying dynamics. We present a general framework for detecting such regulatory DNA and RNA motifs that relies on directly assessing the mutual information between sequence and gene expression measured specific short DNA sequences and then act to modulate the activity of the RNA polymerase. Transcript stability, localization, and translation are also regulated by proteins and RNAs (e.g., miRNAs), which also bind specific short RNA sequences, generally in 3'UTRs. A comprehensive characterization of these DNA and RNA regulatory elements is a formidable challenge, especially within complex metazoan genomes. Experimental (Gerber et al., 2004; Harbison et al., 2004) and computational approaches are emerging to meet these challenges. Several methods compare the intergenic regions of different genomes, aiming to detect sequence elements that are highly conserved across related species (Elemento and Tavazoie, 2005;
These genes belong to cluster 0
These genes belong to cluster 1
These genes belong to cluster 2
This motif is informative about the cluster indices!

5’ upstream regions

Cluster index
0
0
0
These genes belong to cluster 0
1
1
1
These genes belong to cluster 1
1
1
1
1
These genes belong to cluster 2
...
Mutual Information

\[ I(X ; Y) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \]

\( I(X;Y) \) quantifies the amount of information that a variable \( X \) contains about another variable \( Y \)

Expressed in bits
Motif Expression (Cluster Indices)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>0.27</td>
<td>0.07</td>
<td>0.33</td>
</tr>
<tr>
<td>Present</td>
<td>0.07</td>
<td>0.27</td>
<td>0.00</td>
</tr>
</tbody>
</table>

$I(\text{motif}; \text{expression}) = \sum_{i=1}^{2} \sum_{j=1}^{3} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$

$I(M;E) = 0.34 \text{ bits}$
MI estimator is biased  
(sample size bias)

\[
\begin{align*}
N=10 & \quad X_1=\{1,0,0,1,0,1,0,1,0,1\} \\
& \quad X_2=\{0,0,1,1,0,1,0,1,1,0\} \\
I(X_1;X_2) &=? \\
100,000 \text{ times} &
\end{align*}
\]

\[
\begin{align*}
N=100 & \quad X_1=\{1,0,0,1,\ldots,0,1,0,1\} \\
& \quad X_2=\{0,0,1,1,\ldots,0,1,1,0\} \\
I(X_1;X_2) &=? \\
100,000 \text{ times} &
\end{align*}
\]

\[
\begin{align*}
N=1000 & \quad X_1=\{1,0,0,1,\ldots,0,1,0,1\} \\
& \quad X_2=\{0,0,1,1,\ldots,0,1,1,0\} \\
I(X_1;X_2) &=? \\
100,000 \text{ times} &
\end{align*}
\]

Avg MI=0.09  
Avg MI=0.007  
Avg MI=0.0009
MI estimator is biased (sample size bias)

• Can correct for sample size bias, e.g. Slonim et al, 2002 ... slow ... not very precise ... not necessary if:

• Keep sample size the same so that we can compare MI values

• Estimate how large an MI value is compared to expected MI
Algorithm for finding informative motifs
motif representations

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Search space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degenerate code</td>
<td>good</td>
<td>large</td>
</tr>
<tr>
<td>[AC]CGATGAG[TC]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words (k-mers)</td>
<td>acceptable</td>
<td>small</td>
</tr>
<tr>
<td>GCGATGAG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Motif Search Algorithm

## $k$-mer $\text{MI}$

<table>
<thead>
<tr>
<th>$k$-mer</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTCATCG</td>
<td>0.0618</td>
</tr>
<tr>
<td>TCATCGC</td>
<td>0.0485</td>
</tr>
<tr>
<td>AAAATTT</td>
<td>0.0438</td>
</tr>
<tr>
<td>GATGAGC</td>
<td>0.0434</td>
</tr>
<tr>
<td>AAAAATT</td>
<td>0.0383</td>
</tr>
<tr>
<td>ATGAGCT</td>
<td>0.0334</td>
</tr>
<tr>
<td>TTGCCAC</td>
<td>0.0322</td>
</tr>
<tr>
<td>TGCCACC</td>
<td>0.0298</td>
</tr>
<tr>
<td>ATCTCAT</td>
<td>0.0265</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>ACGCGCG</td>
<td>0.0018</td>
</tr>
<tr>
<td>CGACGCG</td>
<td>0.0012</td>
</tr>
<tr>
<td>TACGCTA</td>
<td>0.0011</td>
</tr>
<tr>
<td>ACCCCCT</td>
<td>0.0010</td>
</tr>
<tr>
<td>CCACGGC</td>
<td>0.0009</td>
</tr>
<tr>
<td>TTCAAAA</td>
<td>0.0005</td>
</tr>
<tr>
<td>AGACGCG</td>
<td>0.0004</td>
</tr>
<tr>
<td>CGAGAGC</td>
<td>0.0003</td>
</tr>
<tr>
<td>CTTATTA</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

**Highly informative**

- MI = 0.081
- MI = 0.045
- MI = 0.040

**Not informative**

...
Optimizing k-mers into more informative degenerate motifs

which character increases the mutual information by the largest amount?
Optimizing \( k \)-mers into more informative degenerate motifs

\[
\text{ATCC}[C/G]TACA}
\]

5' upstream regions

Cluster Indices

0 0 0 0
1 1 2 2
0 0 0 0
0 0 0 0
2 2 2 2

A/C
T/C
C/G
A/C/G
A/T/C
C/G/T

ATCC[C/G]TACA
Mutual information

Optimization iterations (visited motif positions)
A schematic view of the optimization process

The exhaustive list of k-mers, represents a finite coarse-grain sample within this general space.
A schematic view of the optimization process

All **degenerate code motifs** represent a finite, more fine-grained sample within this general space.
A schematic view of the optimization process
Is a given motif more informative than expected by chance?
$I(\text{motif} ; \text{expression}) = \sum_{i=1}^{2} \sum_{j=1}^{3} P(i,j) \log \frac{P(i,j)}{P(i)P(j)}$
Motif Expression (Cluster Indices)

\[ I(\text{motif} \; ; \text{expression}) = \sum_{i=1}^{2} \sum_{j=1}^{3} P(i,j) \log \frac{P(i,j)}{P(i)P(j)} \]
Maximum of 10,000 expression-shuffled mutual information values
P-value: probability of obtaining by chance a result at least as extreme as observed result

\[ P(X \geq x) \]

Maximum of 10,000 expression-shuffled mutual information values

\[ P < 10^{-4} \]
Why non-parametric test?

We don’t know what the null distribution of mutual information is like ... depends on sample size, etc.

Null Distribution of T-statistic

Null Distribution of information values

So we estimate it using simulation.
Yeast stress gene expression program (Gasch et al, 2000)

- 173 microarray conditions
- ~ 5,500 genes
- 80 co-expression clusters
- Runtime ~ 1h (standard PC)
17 motifs in 5’ upstream regions
6 motifs in 3’UTRs
PAC is under-represented in cluster 13 (p<1e-5)

PAC is highly over-represented in cluster 66 (p<1e-20)

**Motifs**

<table>
<thead>
<tr>
<th></th>
<th>5’</th>
<th>3’ UTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAC</td>
<td>CTCATC</td>
<td>GA</td>
</tr>
<tr>
<td>RRPE</td>
<td>AAAATTT</td>
<td></td>
</tr>
<tr>
<td>Puf4</td>
<td>UGAU</td>
<td></td>
</tr>
</tbody>
</table>

**Expression Clusters**

- Total # of genes
- # of genes with motif
- # of genes in cluster

**P-value of over-representation of a motif in a cluster of genes**

\[
P(X \geq i) = \sum_{x=i}^{\min(s_1,s_2)} \left( \frac{s_1}{x} \right) \left( \frac{N - s_1}{s_2 - x} \right) \left( \frac{N}{s_2} \right)
\]

Hypergeometric distribution
17 motifs in 5’ upstream regions
6 motifs in 3’UTRs

How many of these motifs are false positives?
Where do false positives come from?

• Multiple hypothesis testing (k-mers)

• Overfit by the optimization procedure
# Motif Search Algorithm

## k-mer | MI
---|---
CTCATCG | 0.0618
TCATCGC | 0.0485
AAAATTT | 0.0438
GATGAGC | 0.0434
AAAAATT | 0.0383
ATGAGCT | 0.0334
TTGCCAC | 0.0322
TGCCACC | 0.0298
ATCTCAT | 0.0265
... | ...
ACGCACGCG | 0.0018
CGACGCG | 0.0012
TACGCTA | 0.0011
ACCCCCT | 0.0010
CCACGCG | 0.0009
TTCAAAA | 0.0005
AGACGCG | 0.0004
CGAGAGC | 0.0003
CTATTATA | 0.0002

**Highly informative**

- MI = 0.081
- MI = 0.045
- MI = 0.040

**Not informative**
Does the algorithm overfit motifs to the expression?

Cluster 0: 112 genes
Cluster 1: 132 genes

Cluster 0: 112 genes
Cluster 1A: 66 genes

Cluster 0: 112 genes
Cluster 1B: 66 genes

Randomly split cluster 1 into cluster 1A and 1B

Best motif, MI=0.38 bits
\[\text{CT}\text{CC}[\text{AG}]\text{[ACT]}\text{AC}[\text{AG}][\text{CT}]\]

MI=0.29 bits when evaluated on this dataset

MI=0.301 bits when evaluated on this dataset

Best motif, MI=0.33 bits
\[\text{[ACG]}[\text{CT}]\text{CC}[\text{AG}][\text{CT}][\text{AG}]\text{C}[\text{AC}]\]
Estimating the false discovery rate

- Run motif discovery algorithm (k-mers+optimization) on random expression profile

- Count how many motifs we get

- Repeat a large number of times, calculate average number of motifs
17 motifs in 5’ upstream regions
6 motifs in 3’UTRs

~ 0.05 “motifs” when shuffling the gene labels of the clustering partition
Entropy

\[ H(X) = - \sum_{x} P(x) \log P(x) \]

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>H(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>1 bit</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>H(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.8</td>
<td>0.72 bits</td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
<th>H(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>0 bits</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>
Mutual Information

\[ I(X ; Y) = H(Y) - H(Y \mid X) \]

Uncertainty about \( Y \)

the amount of uncertainty remaining about \( Y \) after \( X \) is known