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

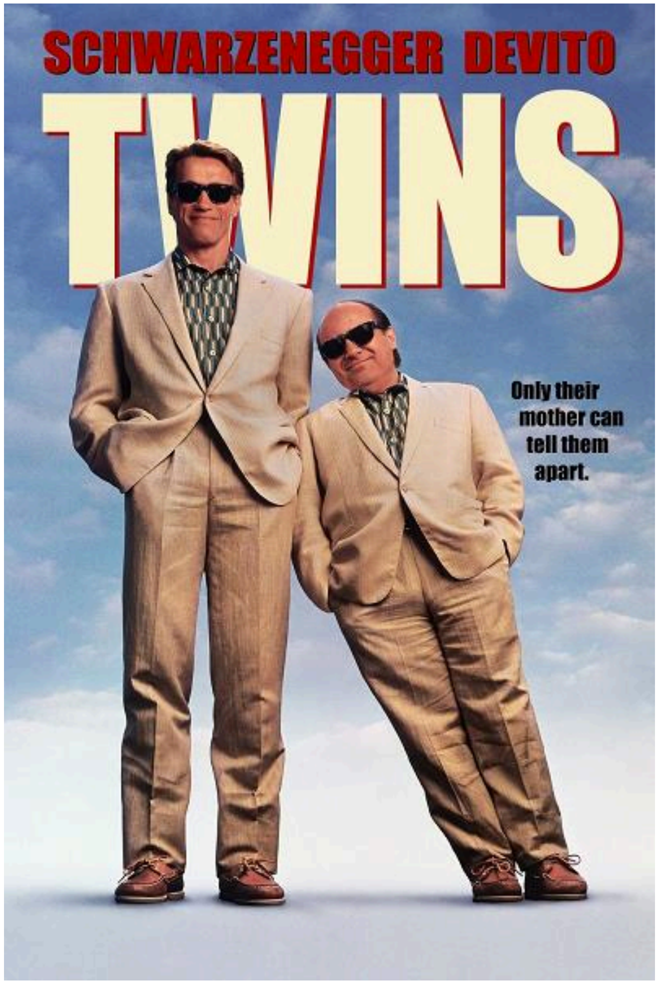
- Mutual information (MI)

- Motif logos

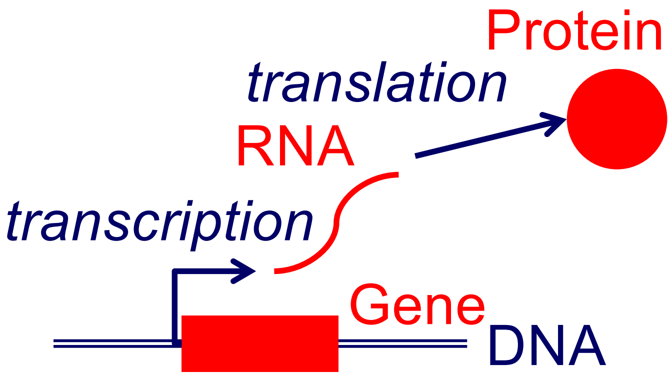
- Using MI to identify cis-regulatory module elements

Gene expression and regulation

Identical DNA but different gene expression



Central dogma


Protein

translation

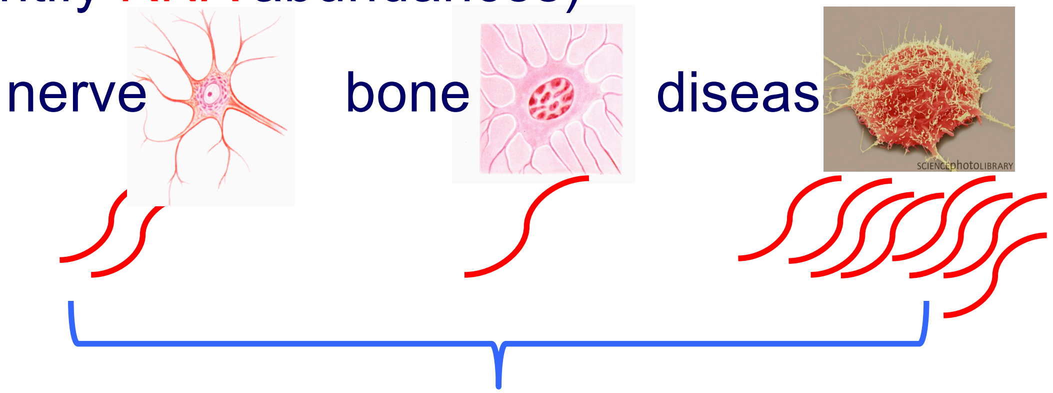
RNA

transcription

Gene

DNA

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


nerve

bone

diseas

Gene function

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?

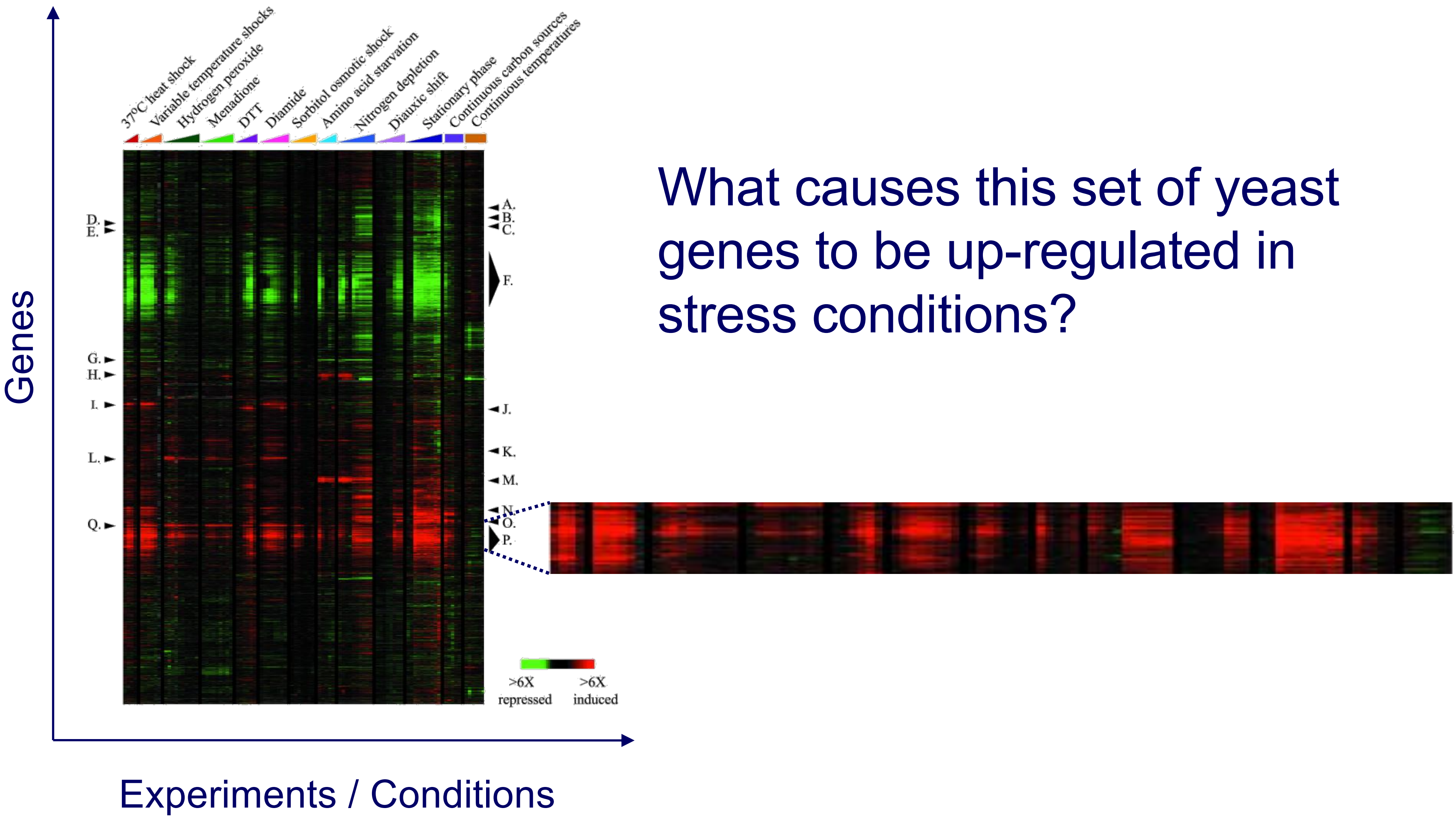
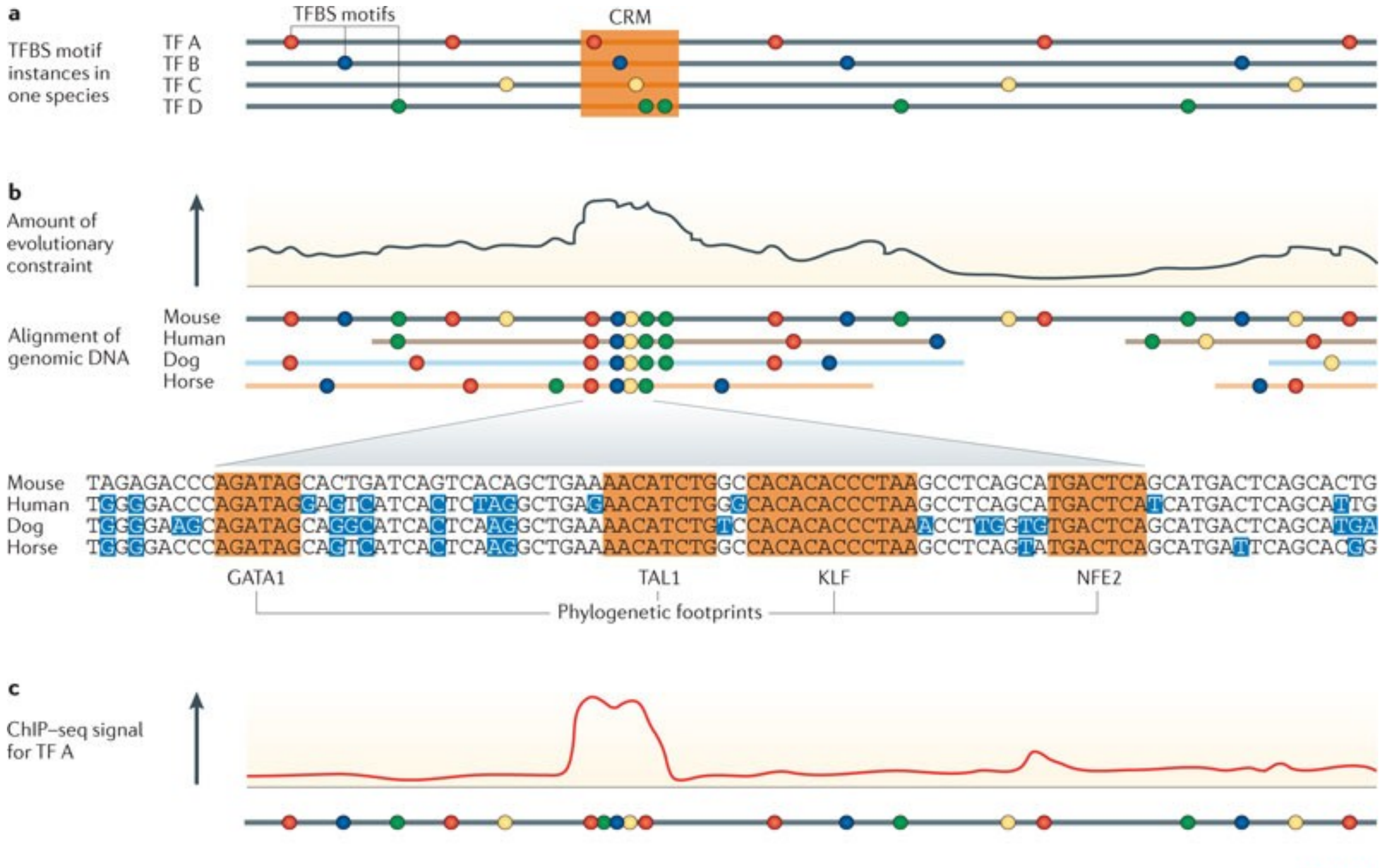


Figure from Gasch *et al*., *Mol. Biol. Cell*, 2000

cis-Regulatory Modules (CRMs)

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

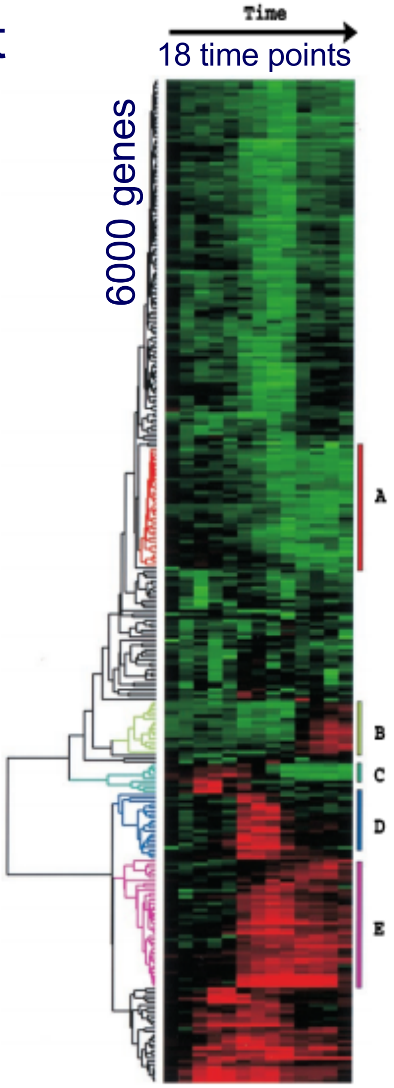

a
TFBS motif instances in one species
TFBS motifs
CRM
TF A
TF B
TF C
TF D
b
Amount of evolutionary constraint
Alignment of genomic DNA
Mouse
Human
Dog
Horse
Mouse TAGAGACCCAGATAGCACTGATCAGTCACAGCTGAAAACATCTGGCCACACACCCTAAGCCTCAGCATGACTCAGCATGACTCAGCACTG
Human TGGGACCCAGATAGGAGTCATCACTCTAGGCTGAGAACATCTGGGCACACACCCTAAGCCTCAGCATGACTCATCATGACTCAGCATTG
Dog TGGGAAGCAGATAGCAGGCATCACTCAAGGCTGAAAACATCTGTCCACACACCCTAAACCTTGGTGTGACTCAGCATGACTCAGCATGA
Horse TGGGACCCAGATAGCAGTCATCACTCAAGGCTGAAAACATCTGGCCACACACCCTAAGCCTCAGTATGACTCAGCATGATTCAGCACGG
GATA1
TAL1
KLF
NFE2
Phylogenetic footprints
c
ChIP–seq signal for TF A

Nature Reviews Genetics volume 13, pages469–483(2012)

Nature Reviews | Genetics

Co-expressed genes have similar functions in single species

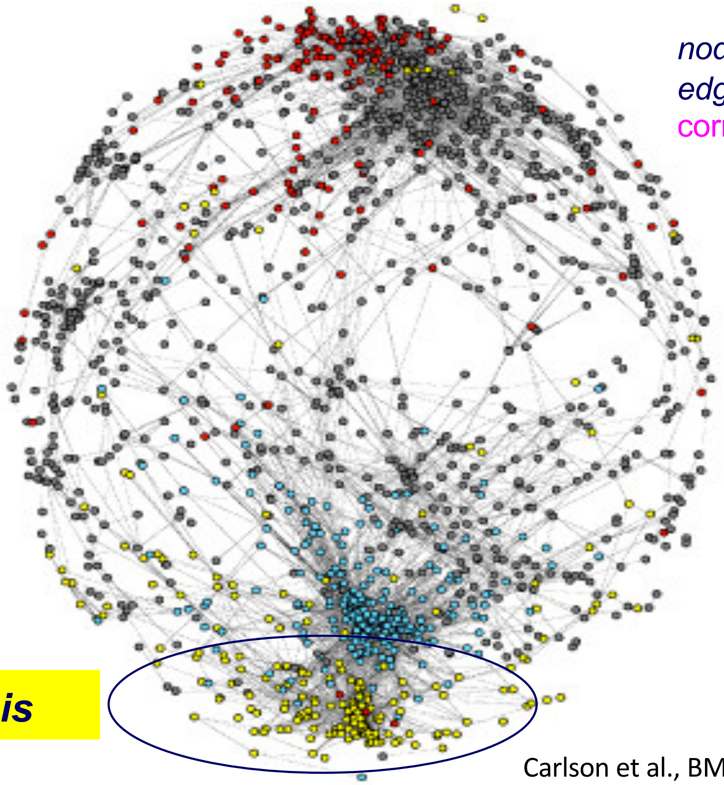
Yeast cell cycle



Eisen et al., PNAS, 1998.

A gene co-expression network (relationship) can reveal ***functional groupings***

- Hierarchal clustering, K-means, Gaussian mixture model (GMM), Principal component analysis (PCA), …



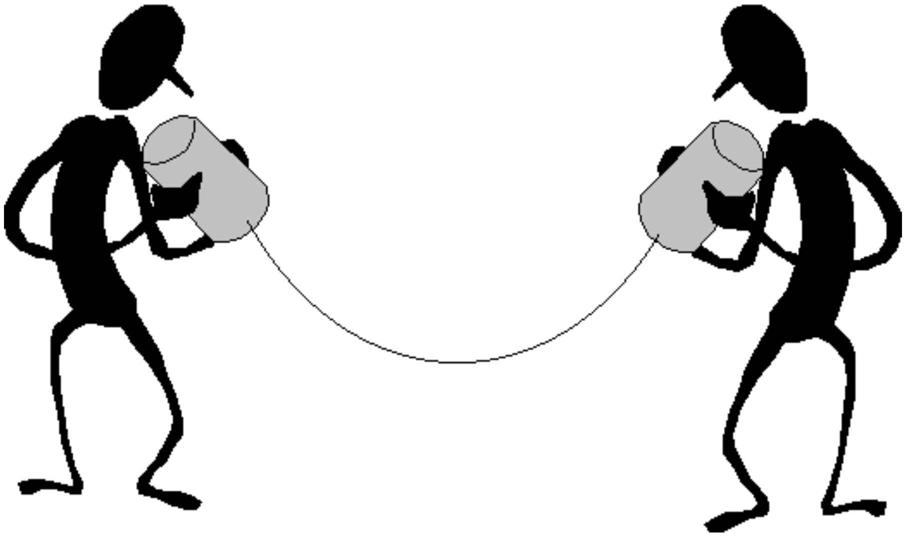
node: gene
edge: expression correlation

Protein synthesis

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

Type	code
Trek	11
Specialized	10
Cervelo	01
Serotta	00

- 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

Type, probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serotta}) = 0.125$	3	000

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

Information Theory Background

Type, probability	# bits	code
$P(\text{Trek}) = 0.5$	1	1
$P(\text{Specialized}) = 0.25$	2	01
$P(\text{Cervelo}) = 0.125$	3	001
$P(\text{Serotta}) = 0.125$	3	000

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

$$-\sum_{c=1}^{|C|} P(c)\log_2 P(c)$$

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

$$H(C) = -\sum_{c=1}^{|C|} P(c)\log_2 P(c)$$

entropy function for binary variable

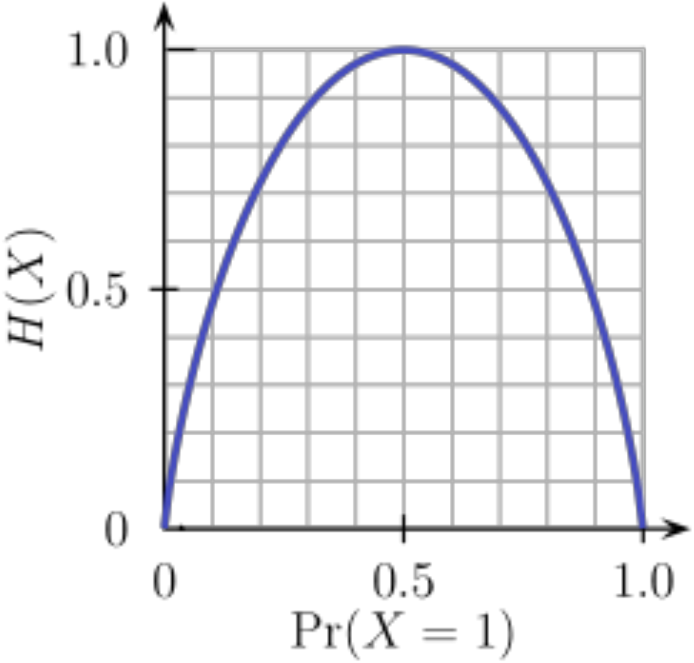


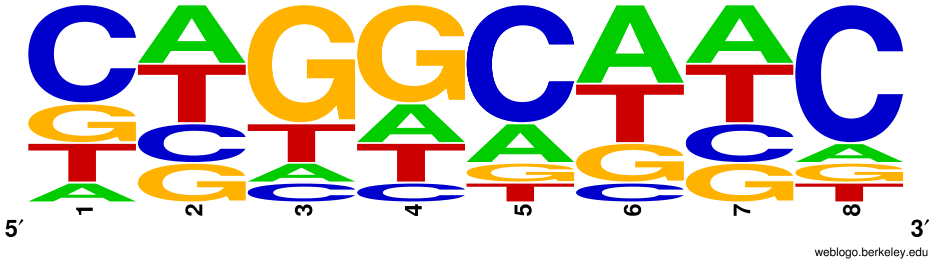
Image from Wikipedia

How is entropy related to DNA sequences?

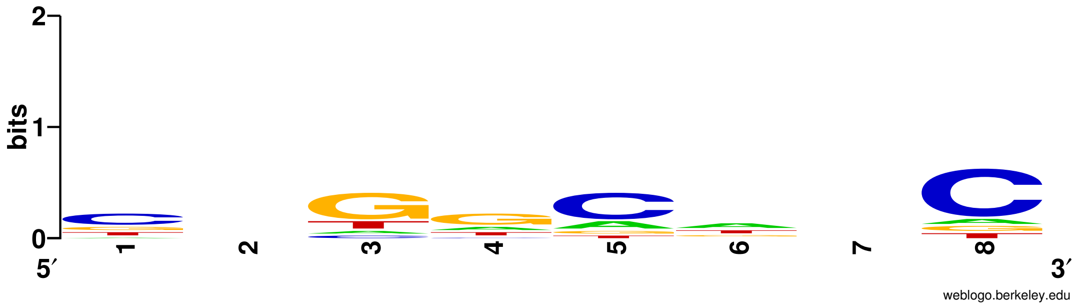
Sequence Logos

	1	2	3	4	5	6	7	8
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

⇩ or ⇩

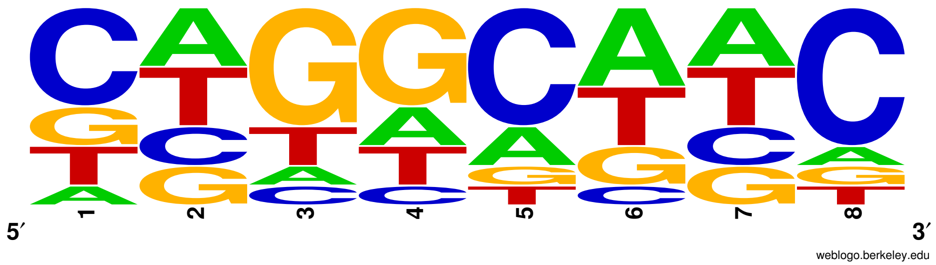


frequency logo



information content logo

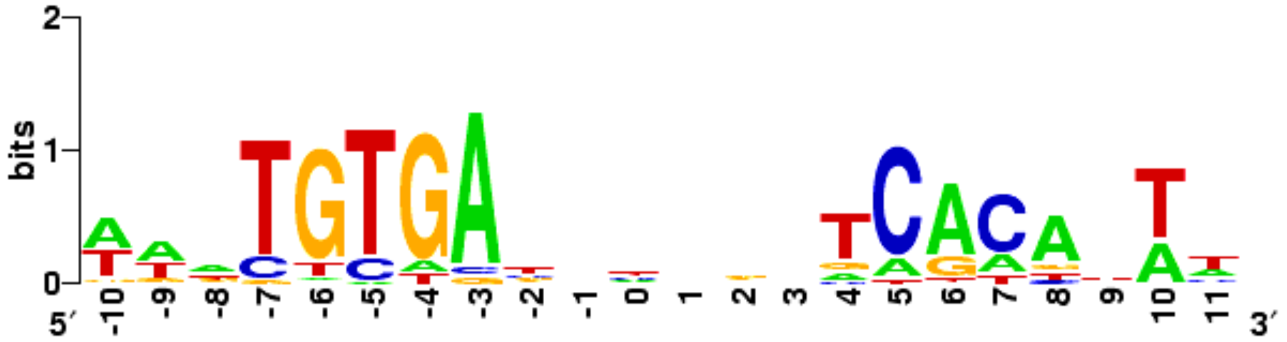
Sequence Logos



- 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

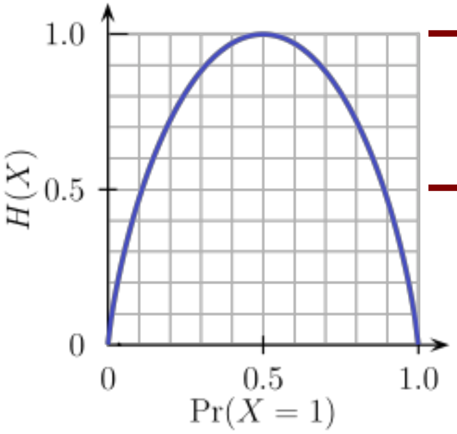
Sequence Logos



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

$$H_{\max} - H(C) = \log_2 N - \left(-\sum_{c} P(c)\log_2 P(c) \right)$$

# 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

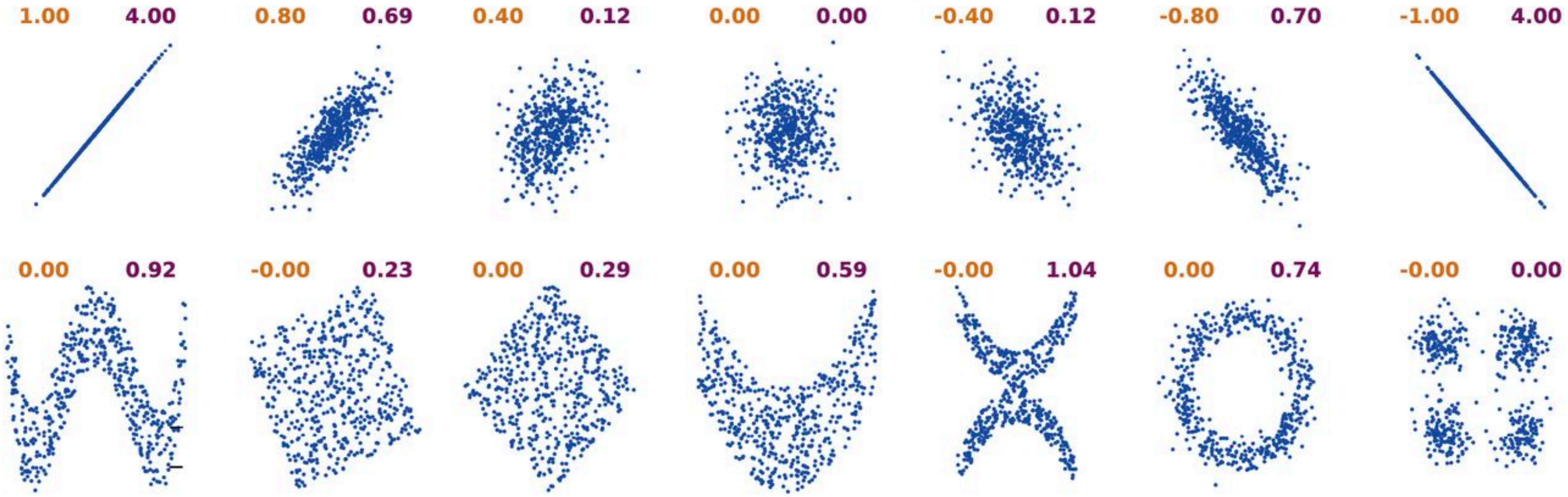
entropy of M

entropy of M conditioned on C

$$I(M;C) = H(M) - H(M \mid C)$$

$$= \sum_{m} \sum_{c} P(m,c) \log_2 \left(\frac{P(m,c)}{P(m)P(c)} \right)$$

Correlation vs. Mutual information

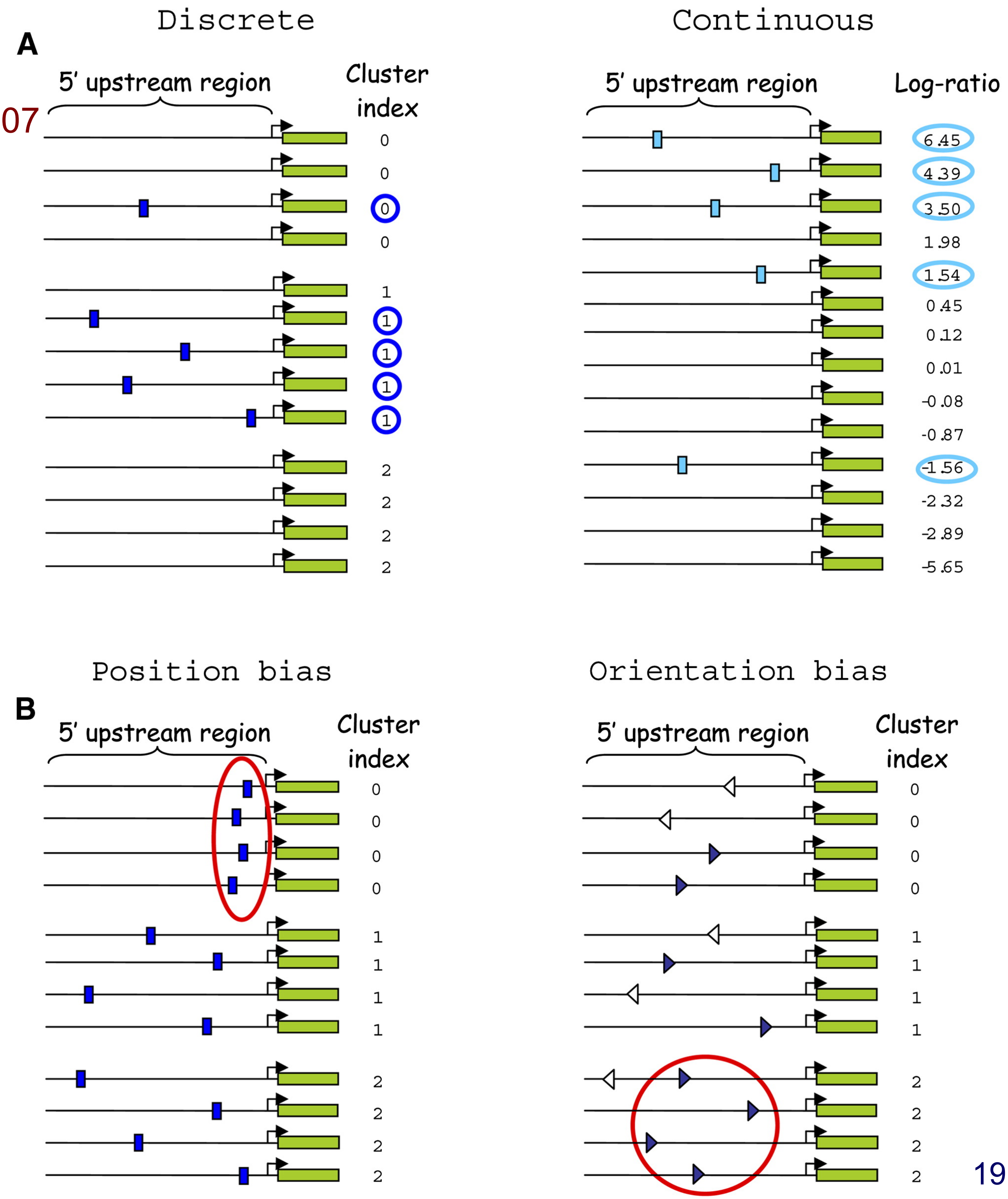


Human Brain Mapping, Volume: 38, Issue: 3, Pages: 1541-1573, First published: 17 November 2016, DOI: (10.1002/hbm.23471)

FIRE

Elemento et al., *Molecular Cell* 2007

- **F**inding **I**nformative **R**egulatory **E**lements (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

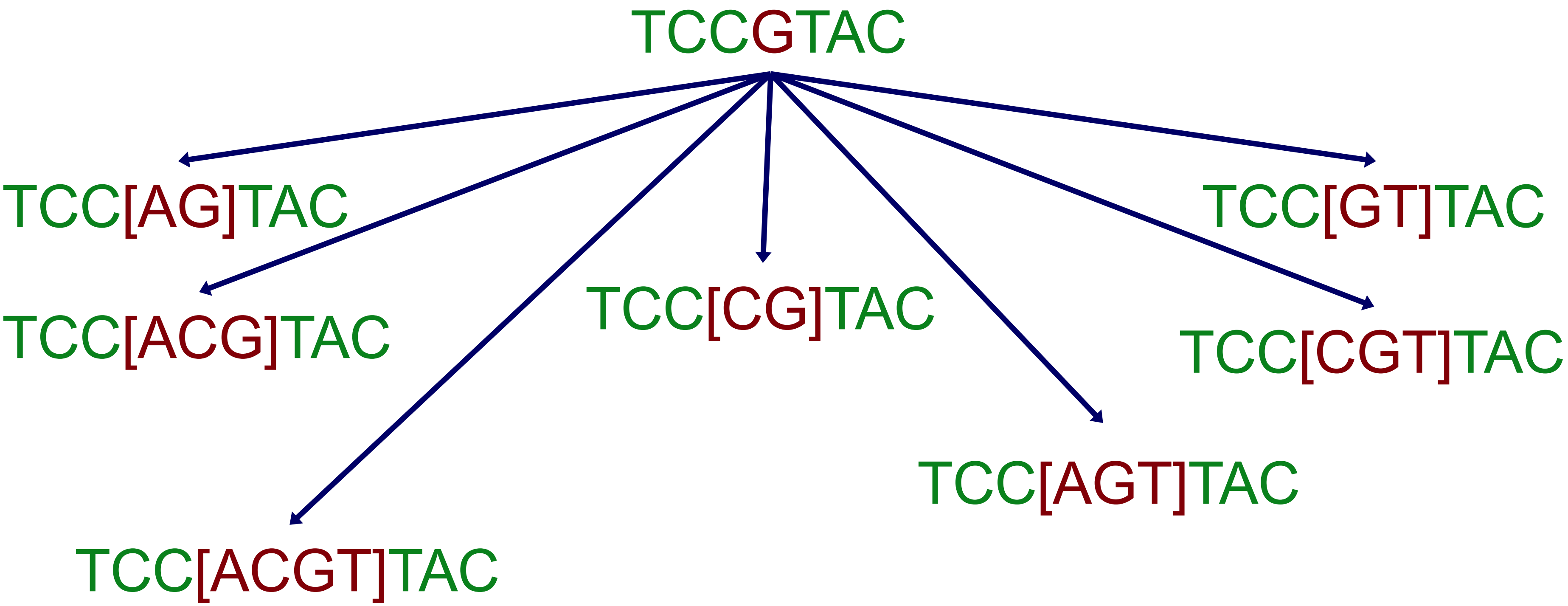
Significance test via randomization

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



TCCGTAC

TCC[AG]TAC

TCC[ACG]TAC

TCC[ACGT]TAC

TCC[CG]TAC

TCC[AGT]TAC

TCC[CGT]TAC

TCC[GT]TAC

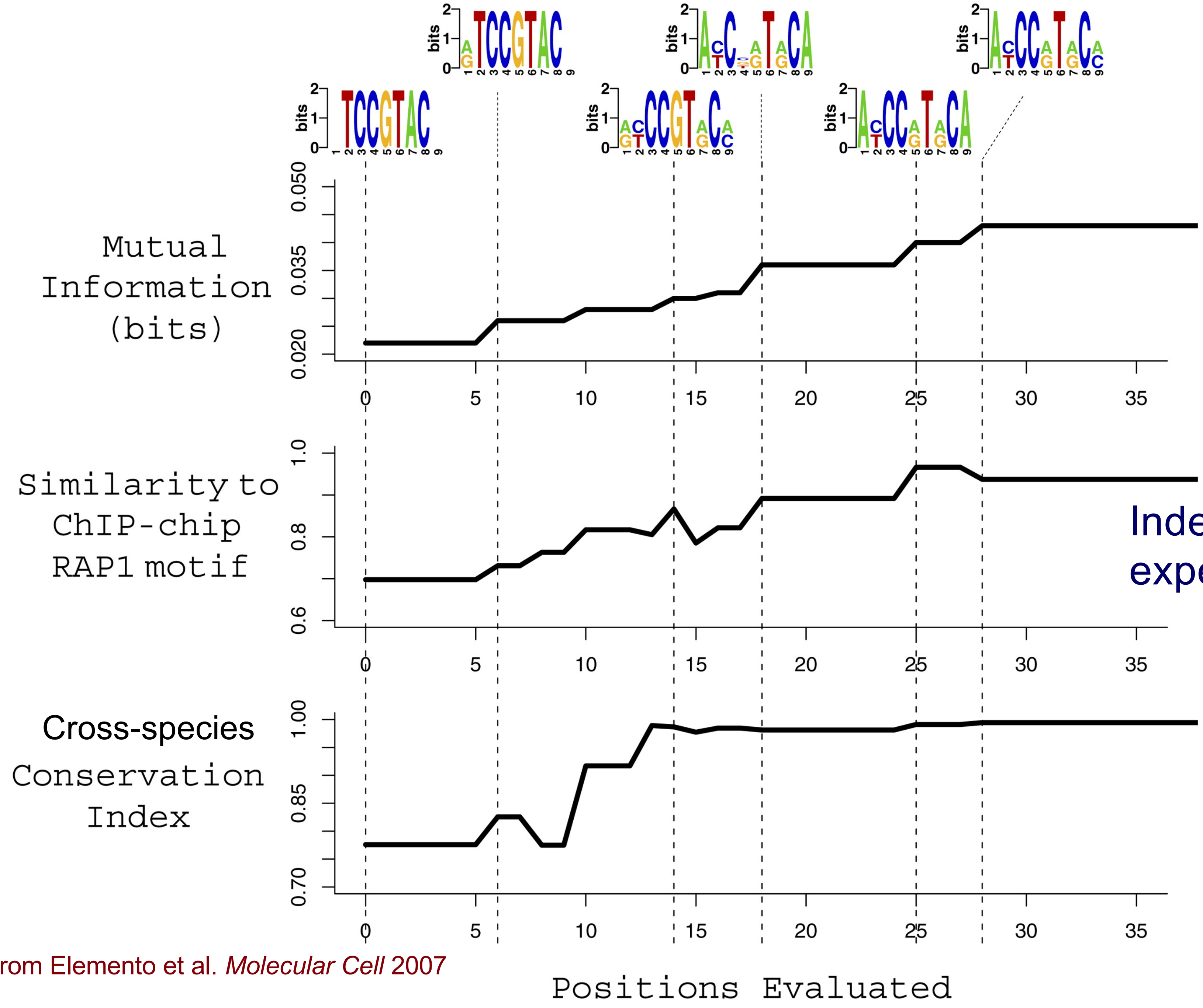
Generalizing a Motif in FIRE

```
given: k-mer, n

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

return: best
```

Generalizing a Motif in FIRE: Example



Independent experiment

Figure from Elemento et al. *Molecular Cell* 2007

Avoiding Redundant Motifs

- Different seeds could converge to similar motifs

TCCGTAC → TCC[CG]TAC

TCCCTAC → TCC[CG]TAC

- 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 M new candidate motif C expression clusters

Characterizing Predicted Motifs in FIRE

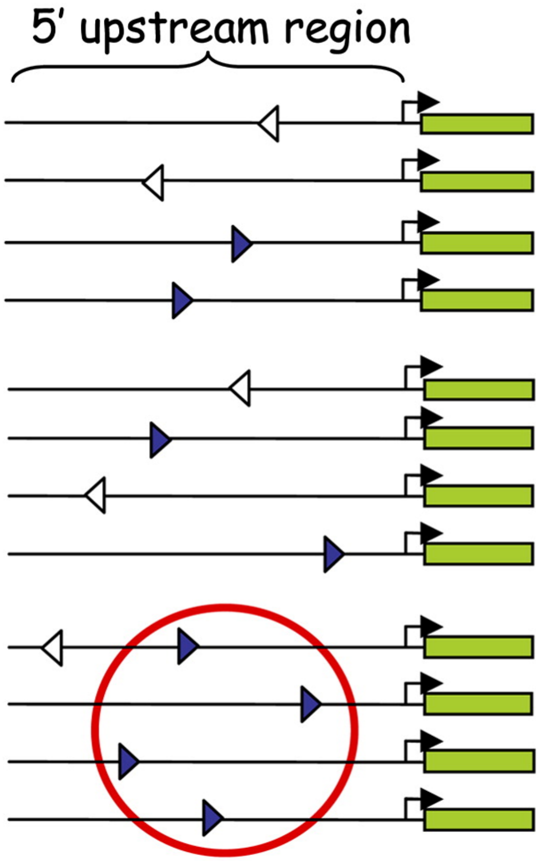
- Mutual information is also used to assess various properties of found motifs
  - orientation bias
  - position bias
  - interaction with another motif

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)



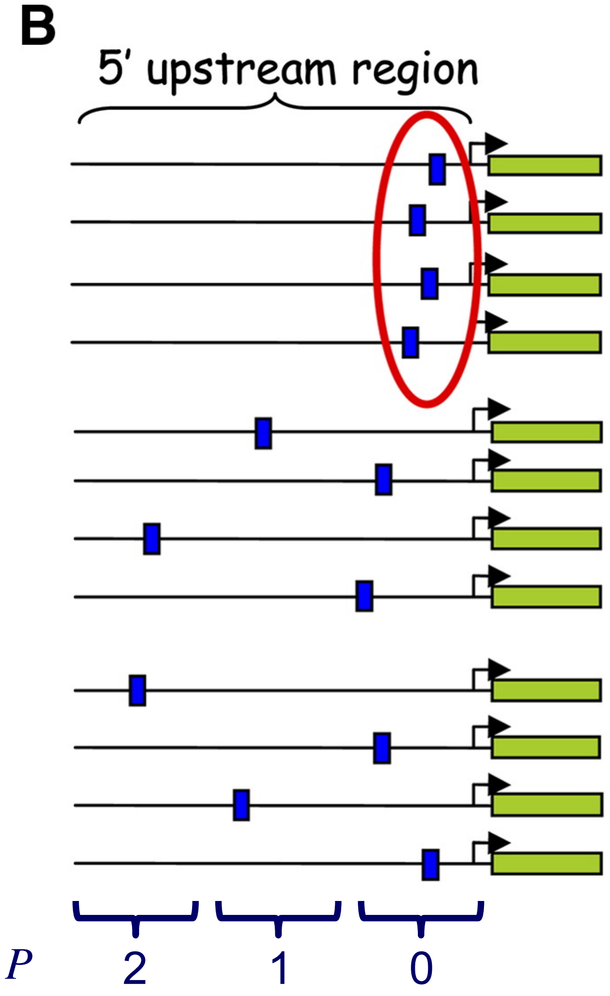
C	S
0	0
0	0
0	1
0	1
1	0
1	1
1	0
1	1
2	1
2	1
2	1
2	1

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

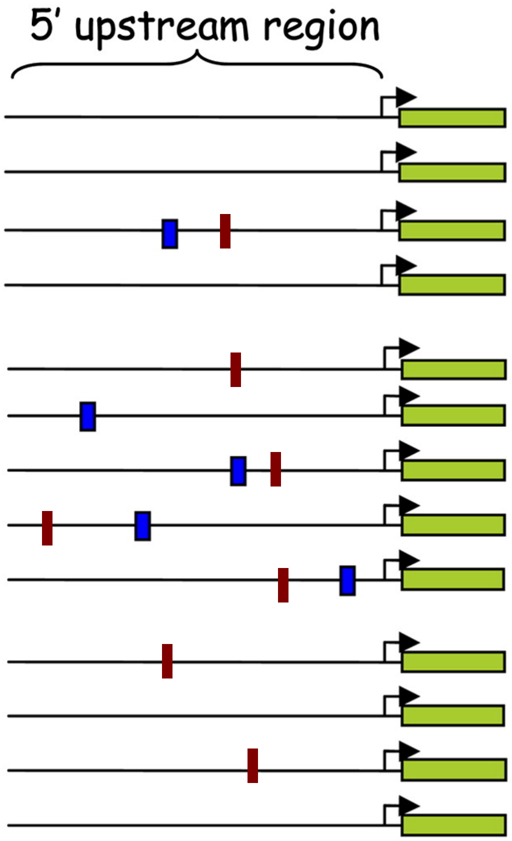


P	O
0	1
0	1
0	1
0	1
1	0
0	0
2	0
0	0
2	0
0	0
1	0
0	0

Only sequences containing the motif are considered for this calculation

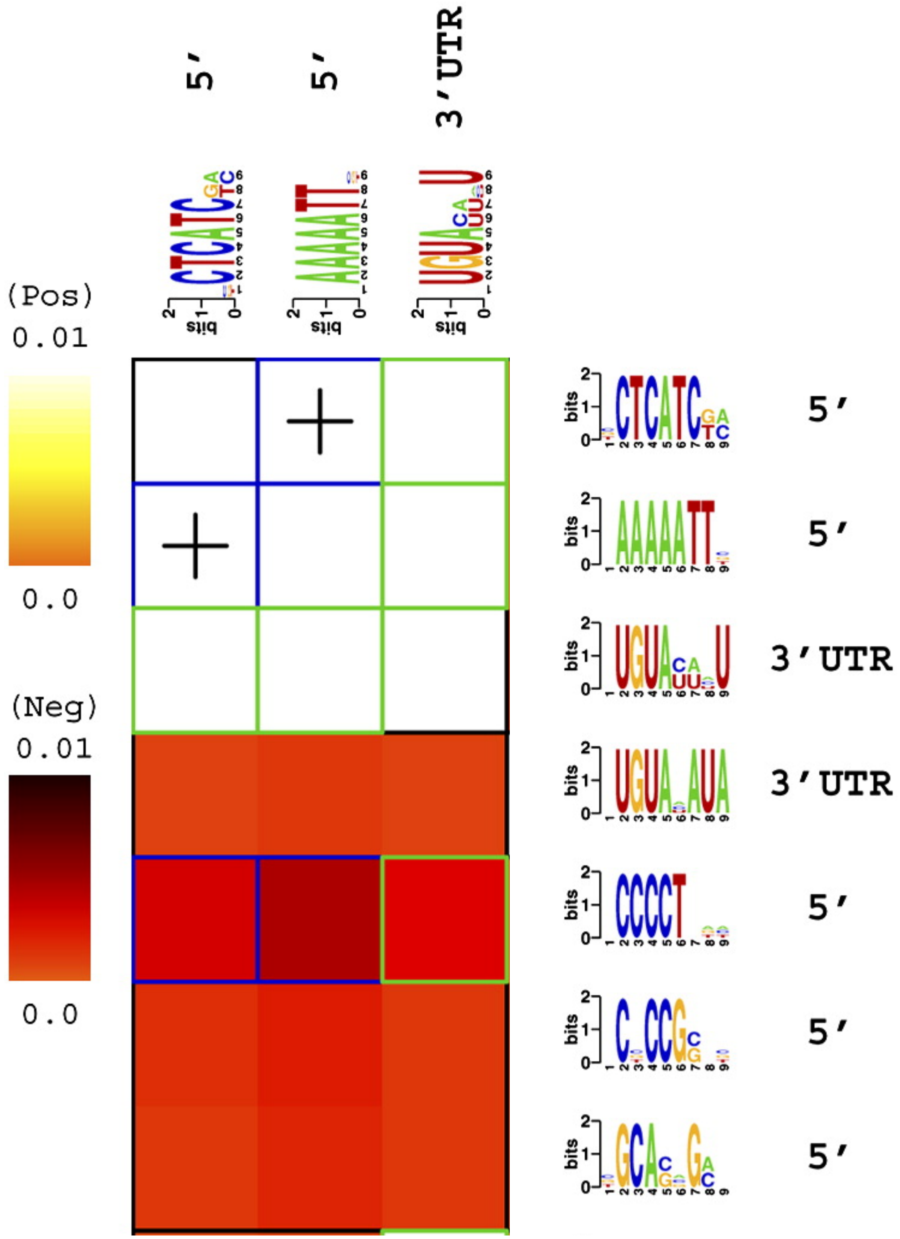
Using MI to Determine Motif Interactions

$I(M_1; M_2)$ $M_1=0, 1$ 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



M_1	M_2
0	0
0	0
0	0
0	0
0	1
1	0
1	1
1	1
1	1
0	1
0	0
0	1
0	0

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

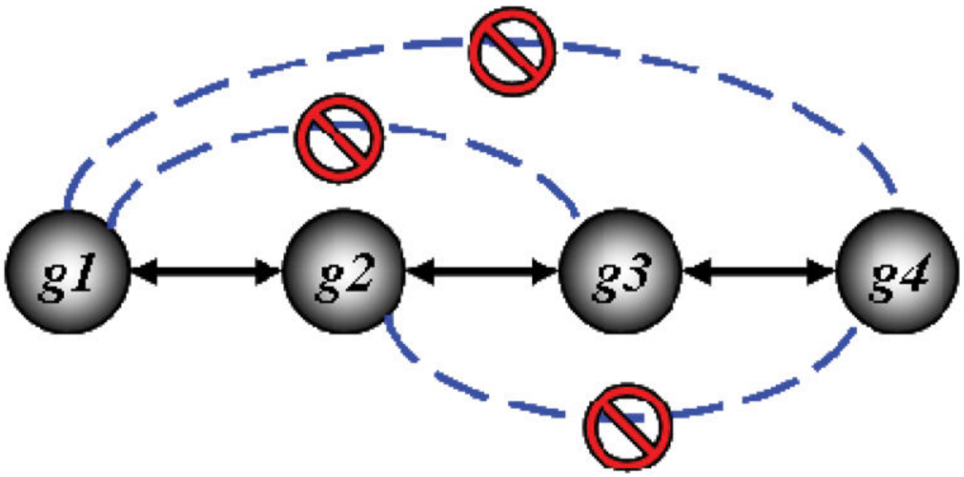
- Build gene-gene network where edges indicate high MI in genes' expression levels

- 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