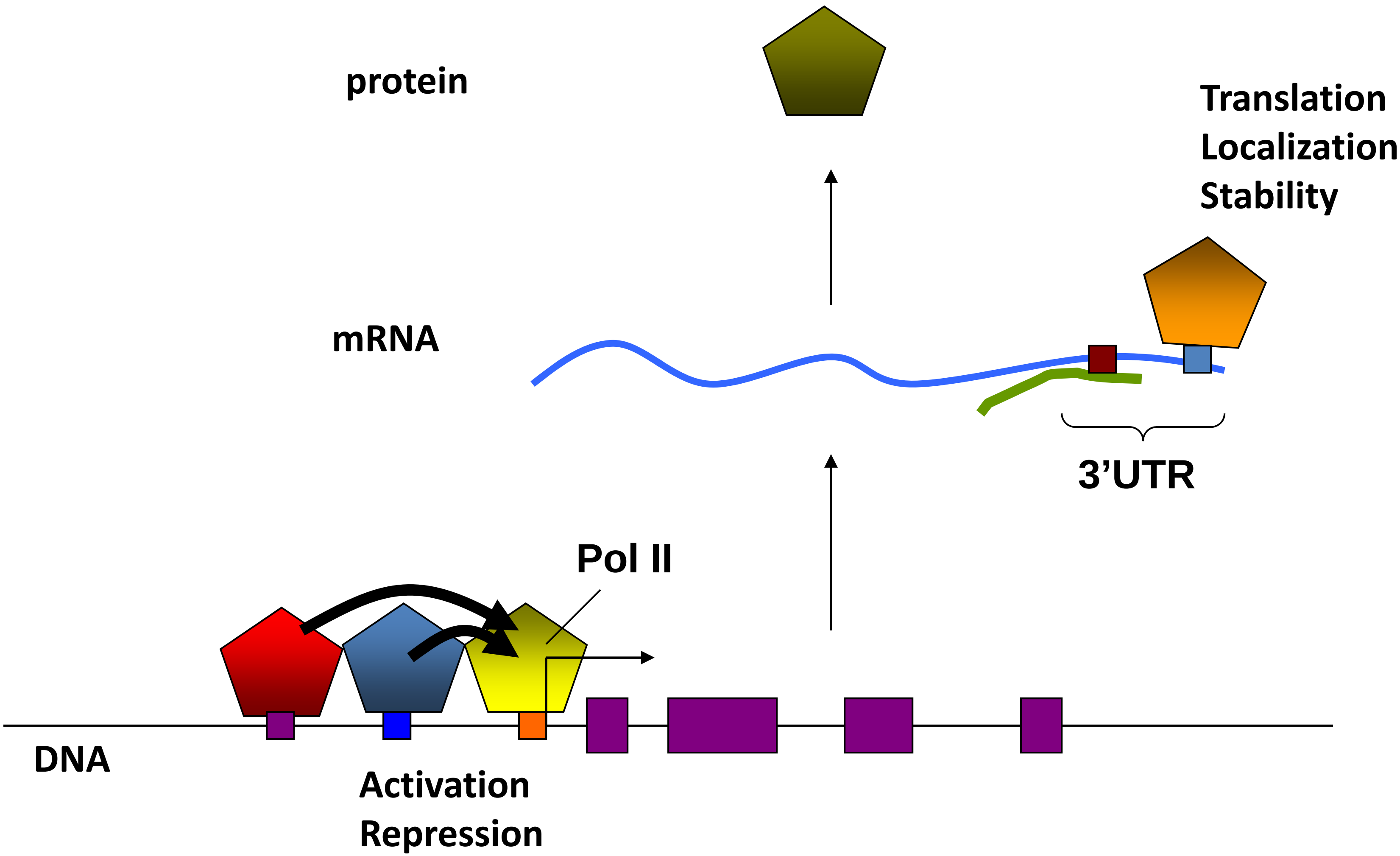
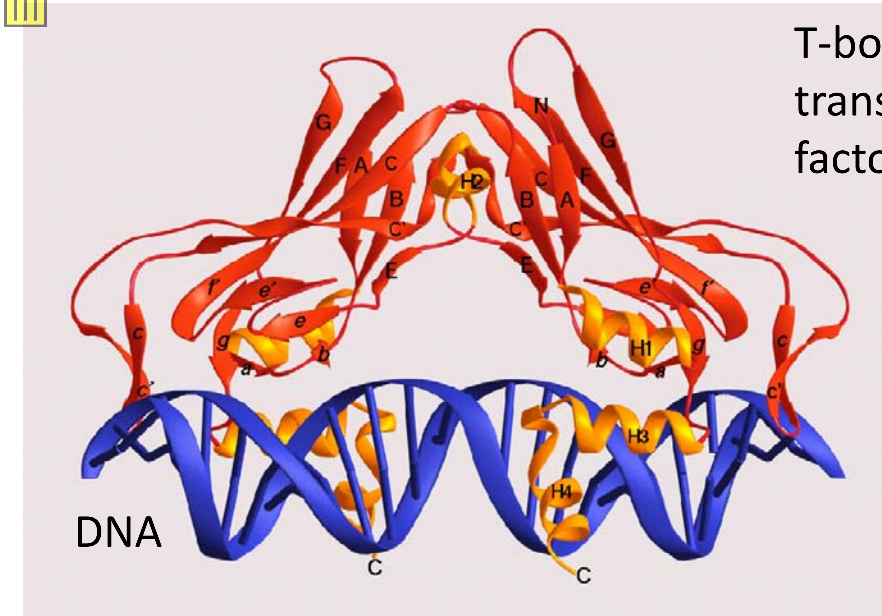
Discovering regulatory sequences from expression data

- Unsupervised clustering
- Information theory
- Optimization
- Non-parametric statistical testing
- Multiple testing
- Overfitting

Transcriptional and post-transcriptional regulation of gene expression



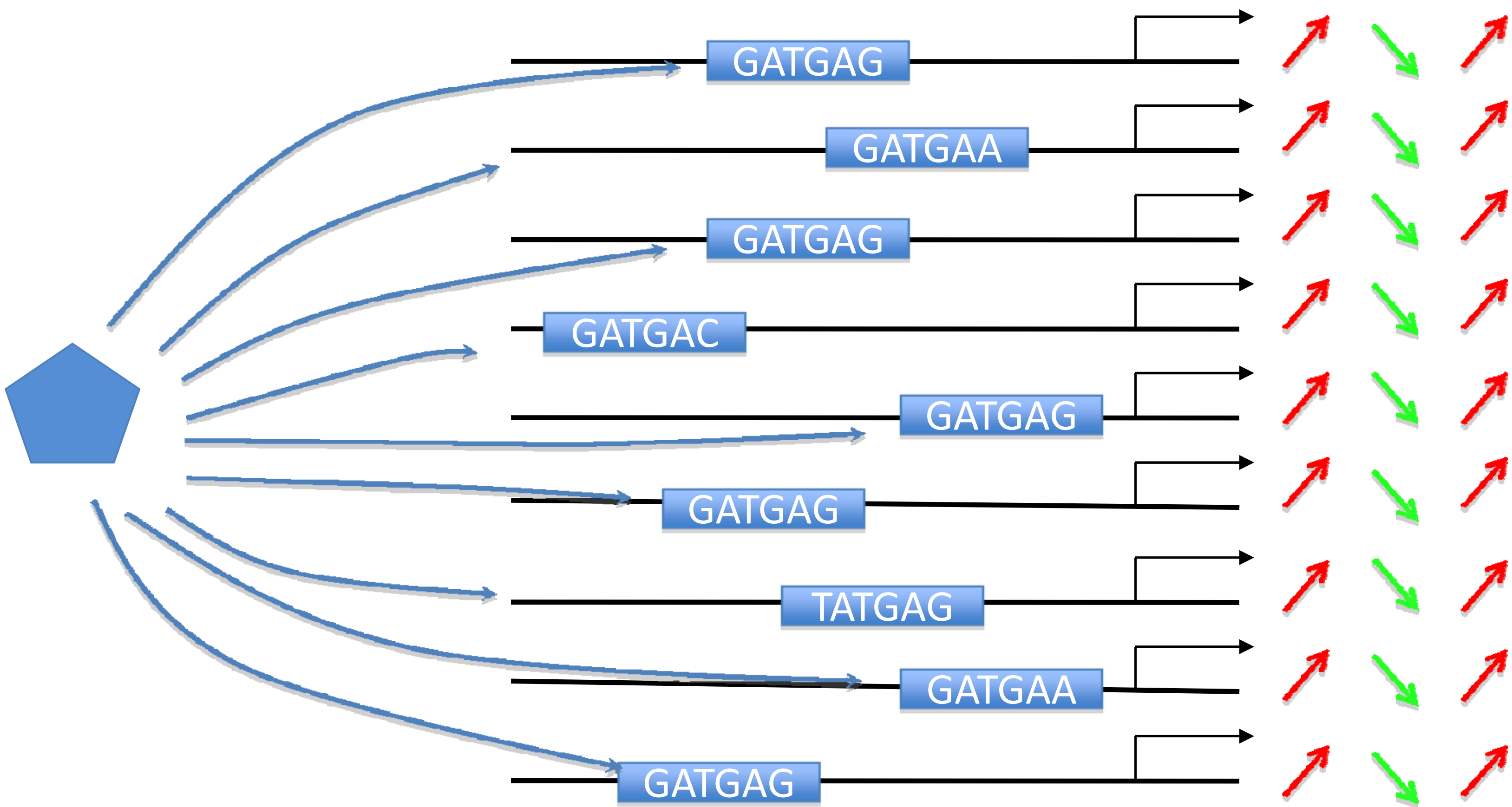
T-box
transcription
factor



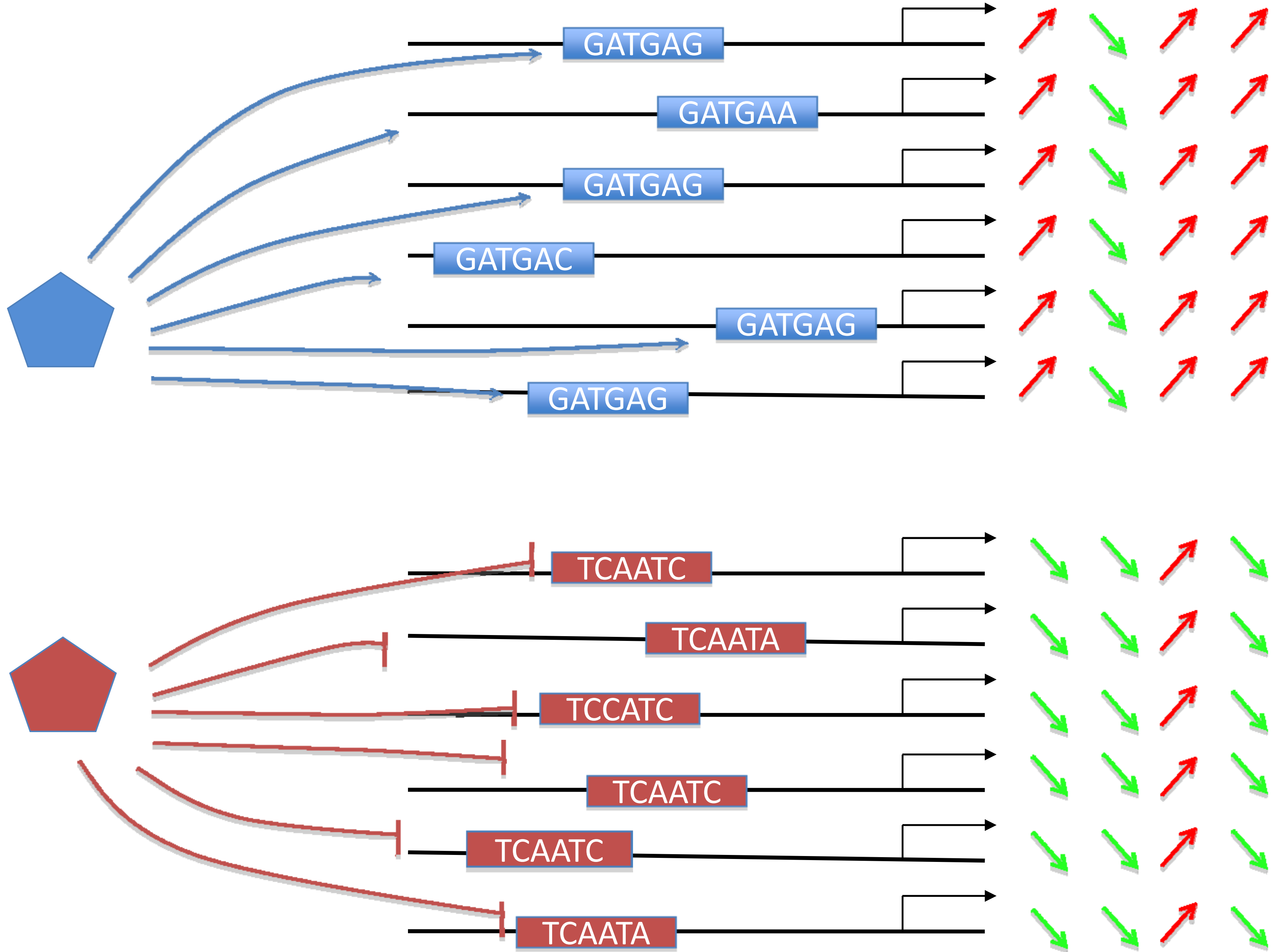
CRP

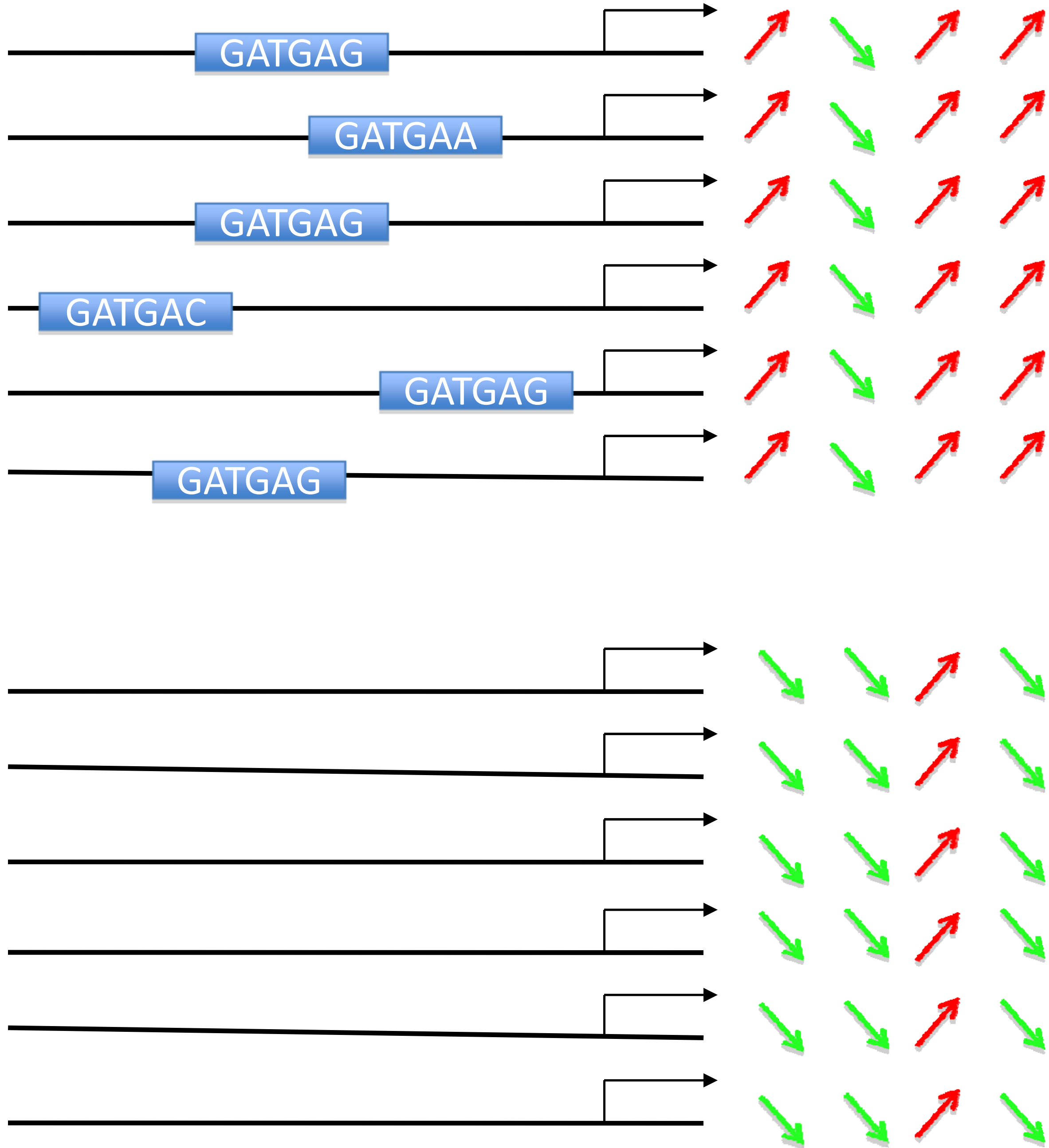
DNA

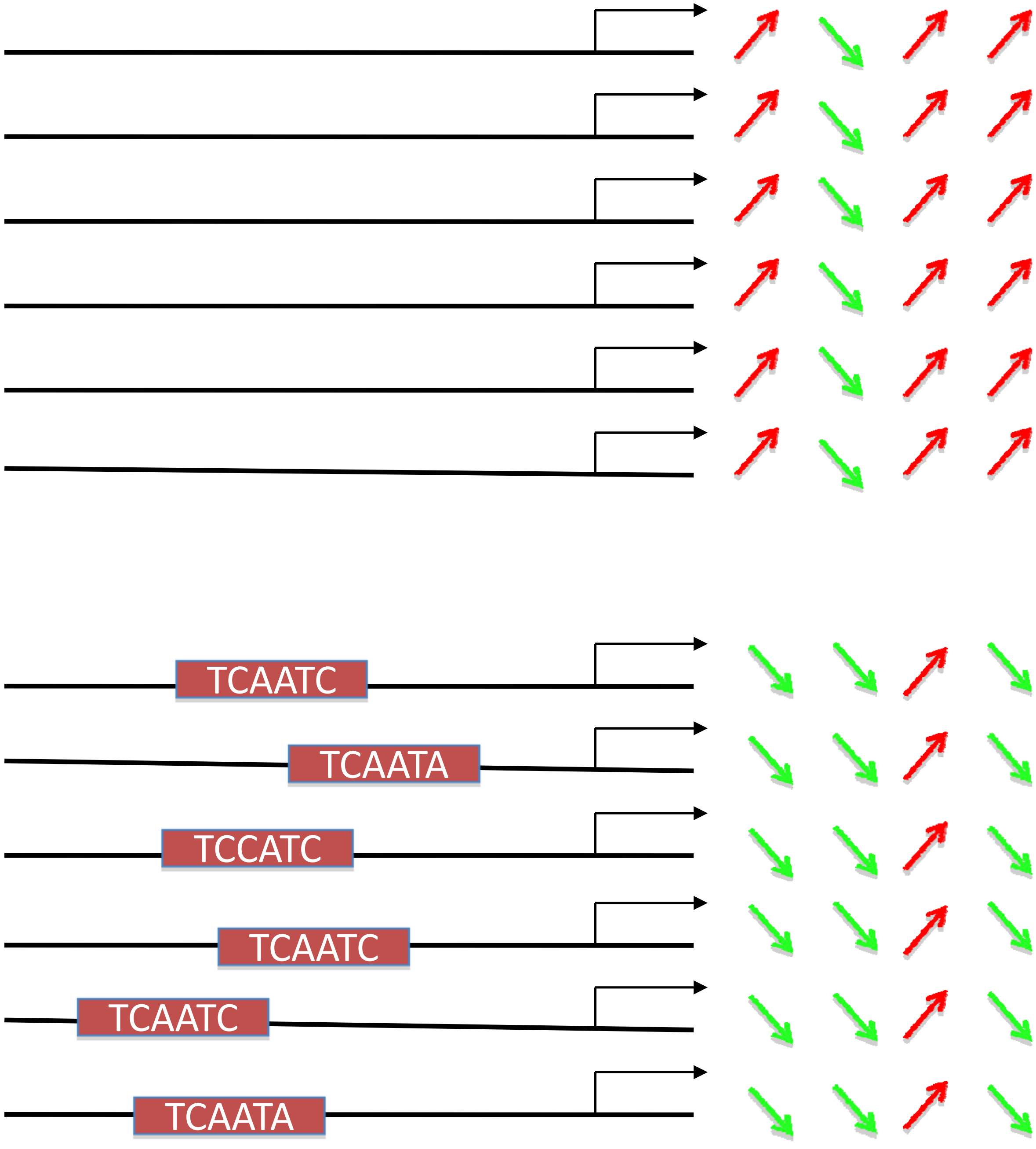
Transcription factor binding sites are ~6-12 bp-long



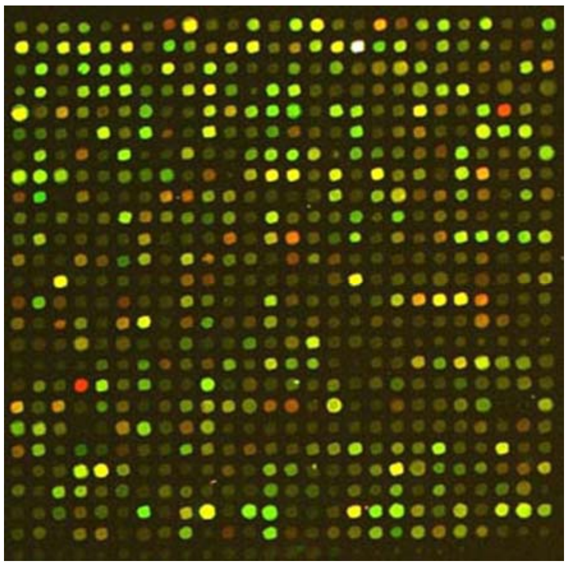
Genes regulated by the same TF will be co-expressed


GATGAG
GATGAA
GATGAG
GATGAC
GATGAG
GATGAG
TCAATC
TCAATA
TCCATC
TCAATC
TCAATC
TCAATA

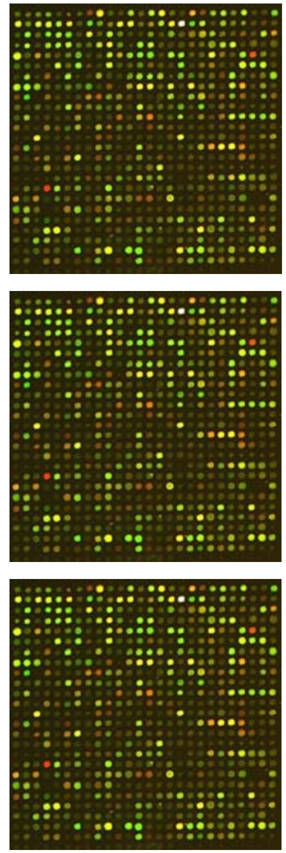

GATGAG
GATGAA
GATGAG
GATGAC
GATGAG
GATGAG


TCAATC
TCAATA
TCCATC
TCAATC
TCAATC
TCAATA

Microarray experiment



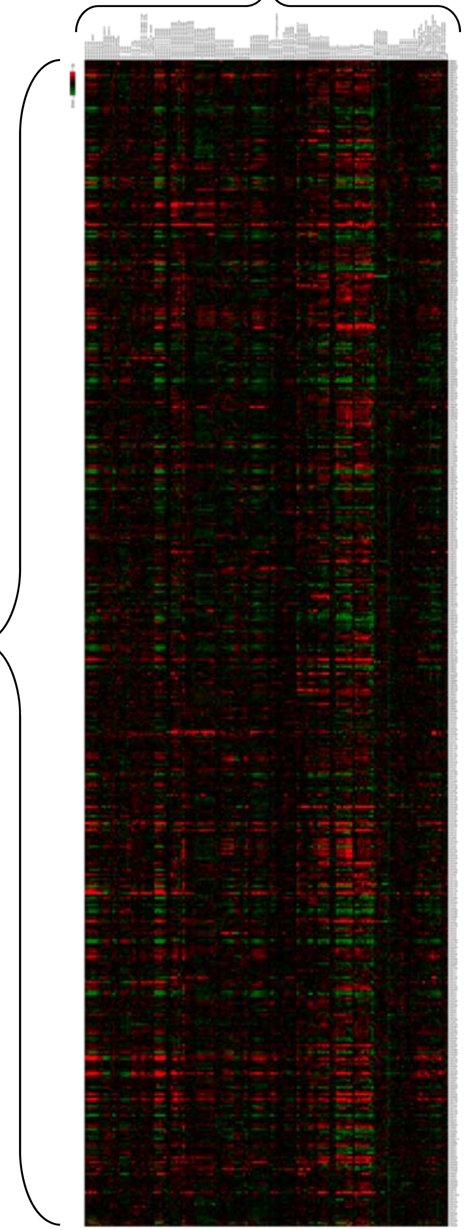
Several microarray experiments (conditions, time points, treatments)



...

Microarray Experiments

Genes

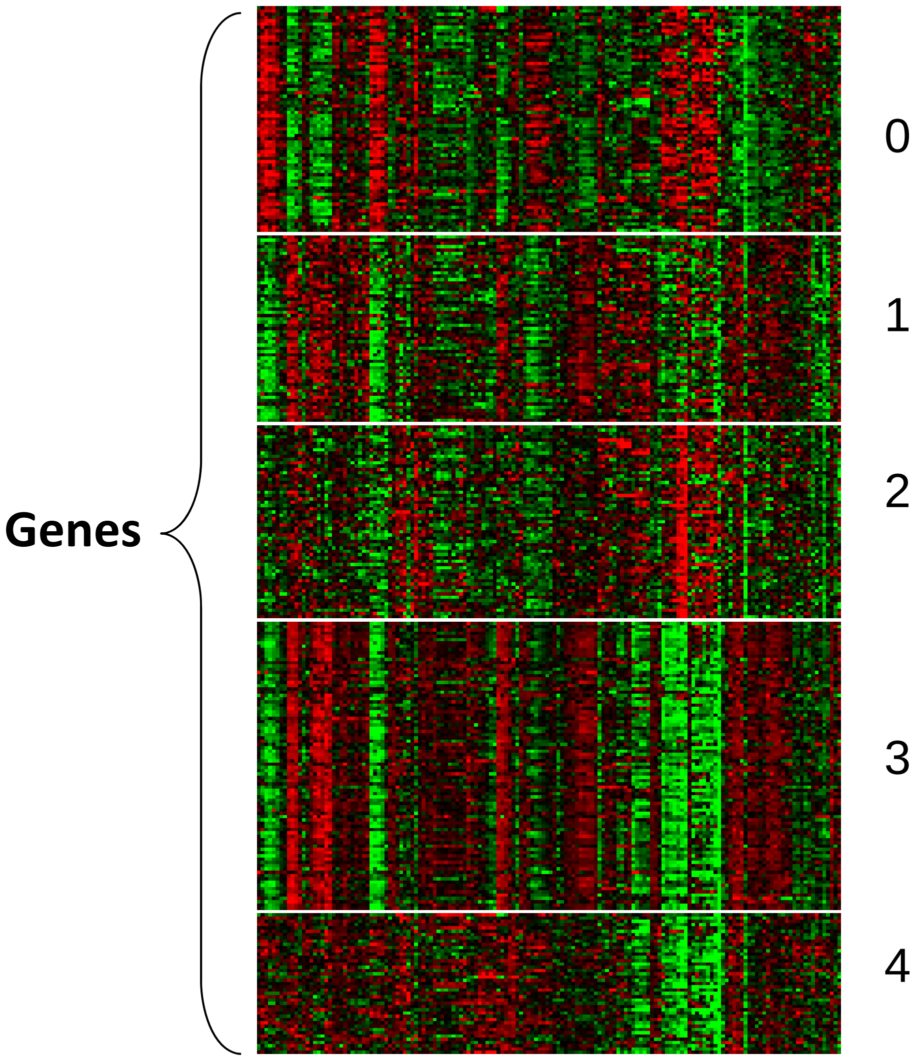


Creating co-expression clusters

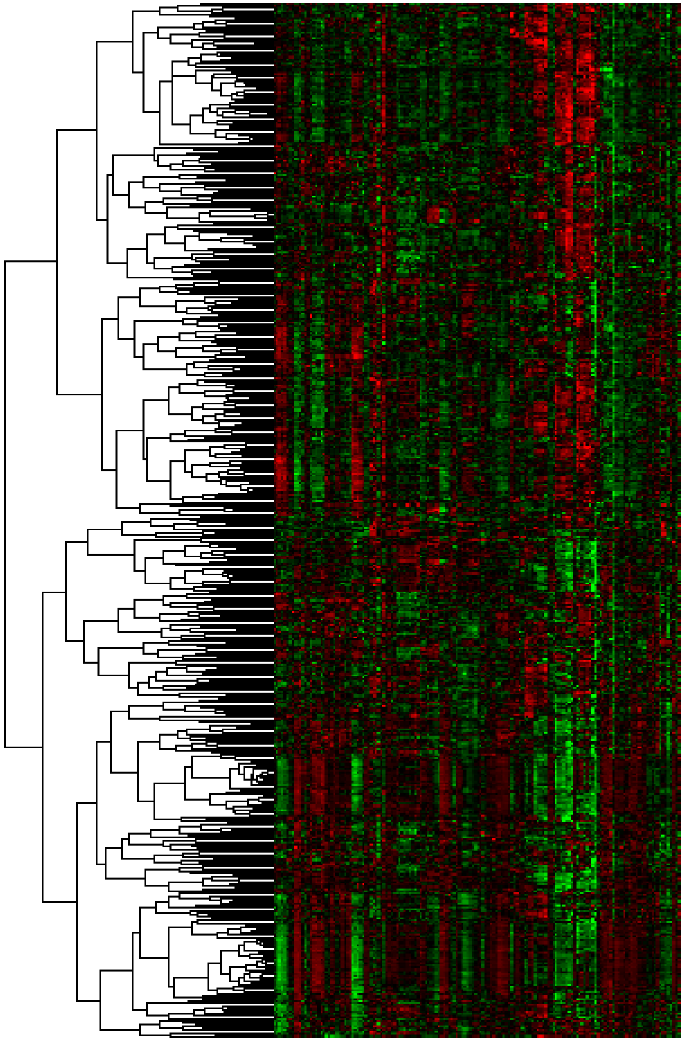
Unsupervised clustering approaches:

- K-means
- Self-organizing maps
- Hierarchical clustering

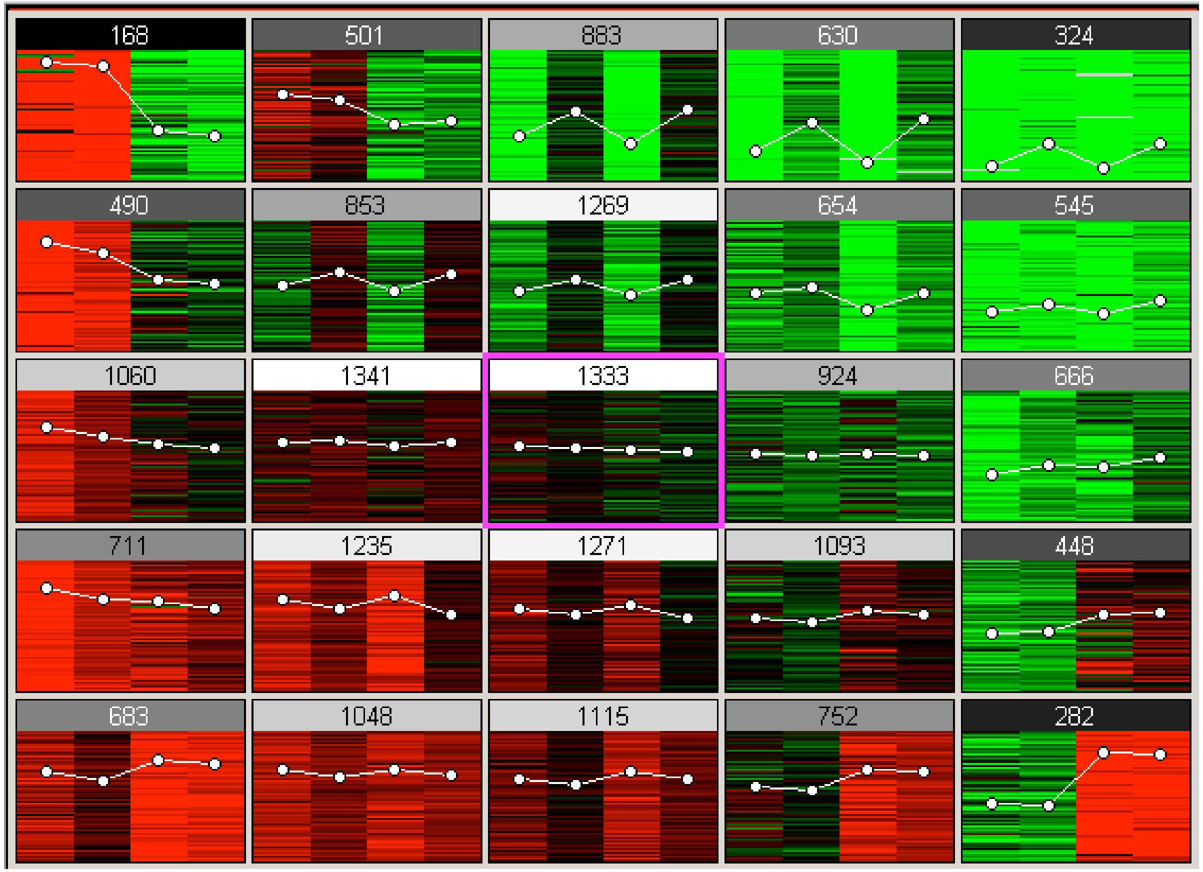
K-means clustering



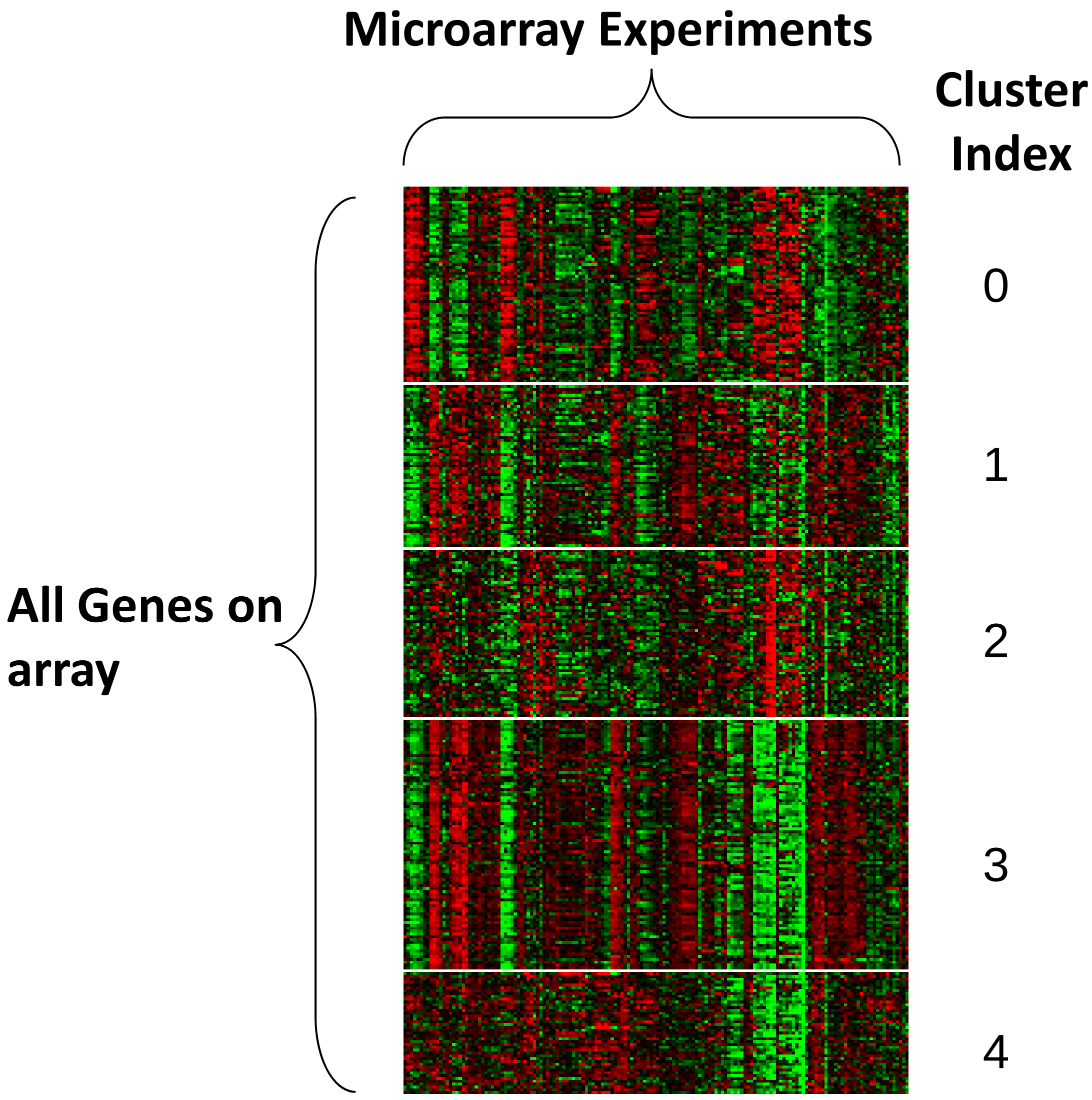
Hierarchical clustering



Self-organizing map



Clusters of co-expressed genes



Molecular Cell
Resource

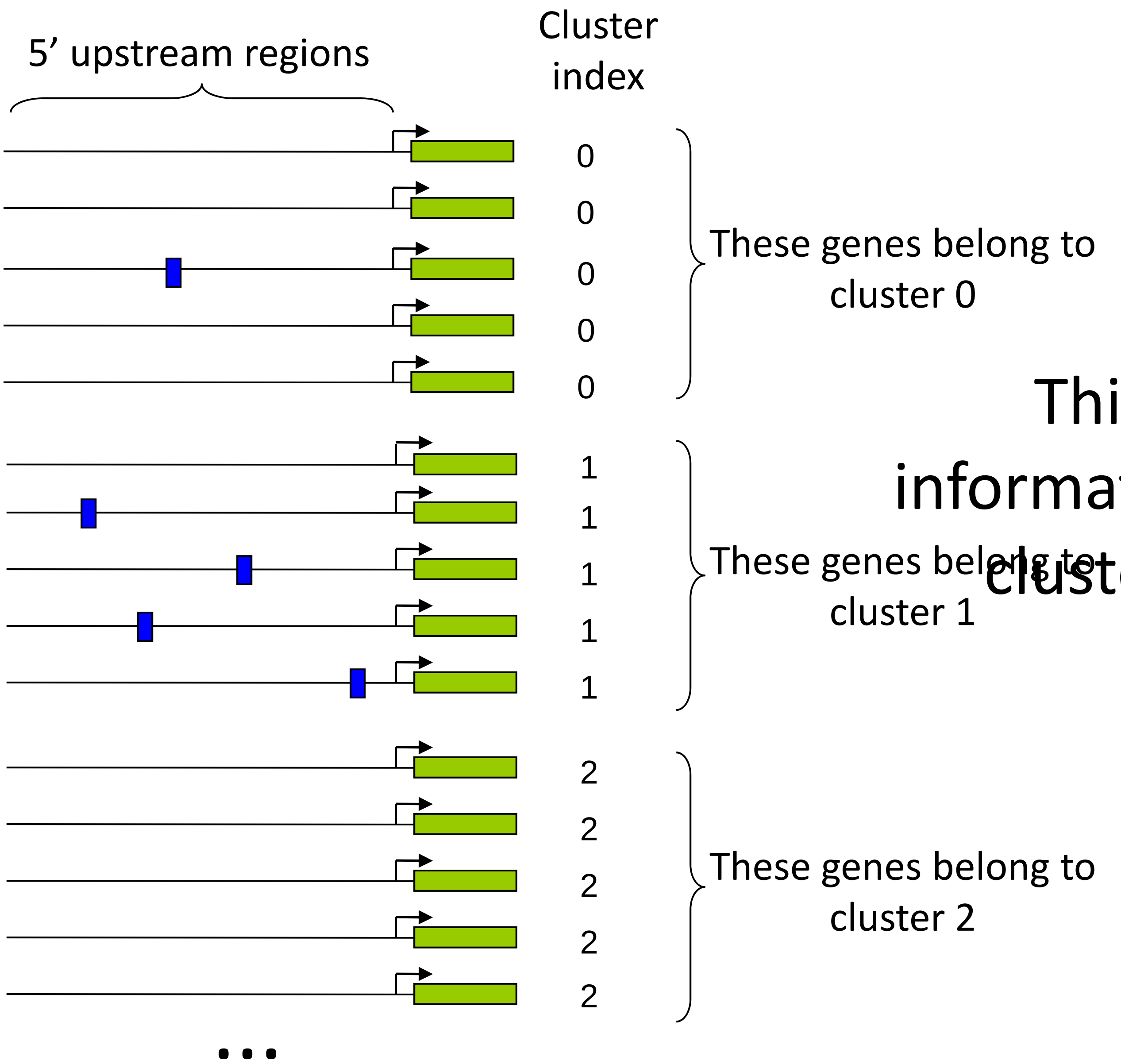
A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types

Olivier Elemento,[1,2,3] Noam Slonim,[1,2,3,4] and Saeed Tavazoie[1,2,*]
[1]Lewis-Sigler Institute for Integrative Genomics
[2]Department of Molecular Biology
Princeton University, Princeton, NJ 08544, USA
[3]These authors contributed equally to this work.
[4]Present 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 mea-

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;



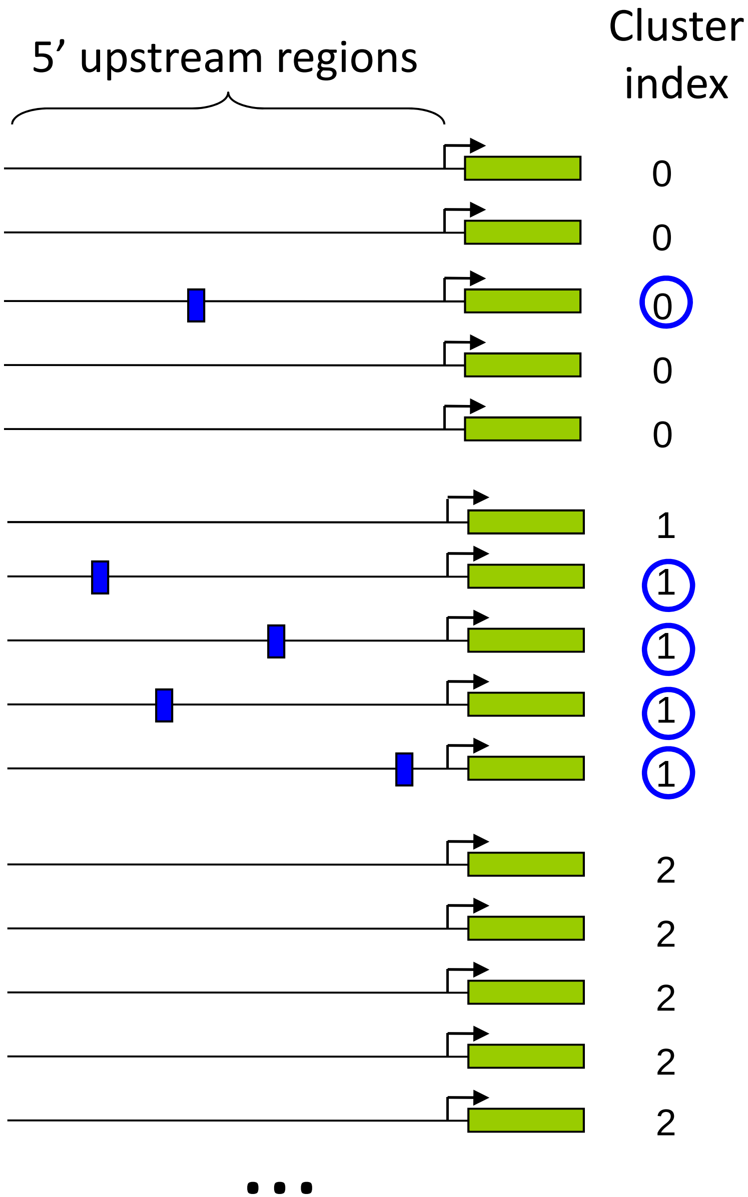
This motif is informative about the cluster indices !

Mutual Information

$$I(\mathrm{X};\mathrm{Y}) = \sum_{x}\sum_{y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

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

Expressed in bits



		Expression (Cluster Indices)		
		0	1	2
Motif	Absent	0.27	0.07	0.33
	Present	0.07	0.27	0.00

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

I(M;E) = 0.34 bits

MI estimator is biased (sample size bias)

N=10

X1={1,0,0,1,0,1,1,0,1,0}

X2={0,0,1,1,0,1,0,1,1,0}

I(X1;X2) =?

100,000 times

N=100

X1={1,0,0,1,.... ,1,0,1,0}

X2={0,0,1,1,.... ,0,1,1,0}

I(X1;X2) =?

100,000 times

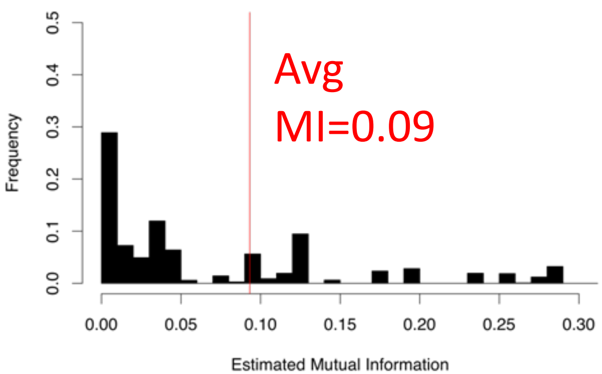
N=1000

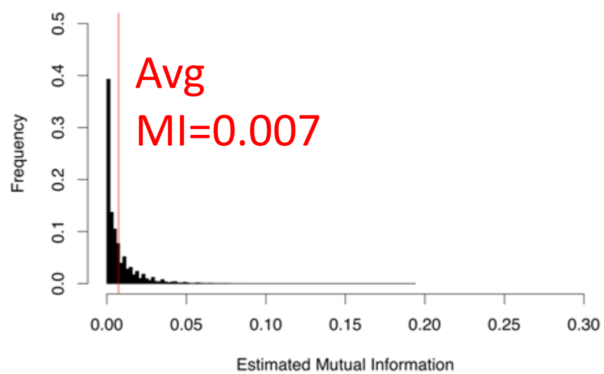
X1={1,0,0,1,.... ,1,0,1,0}

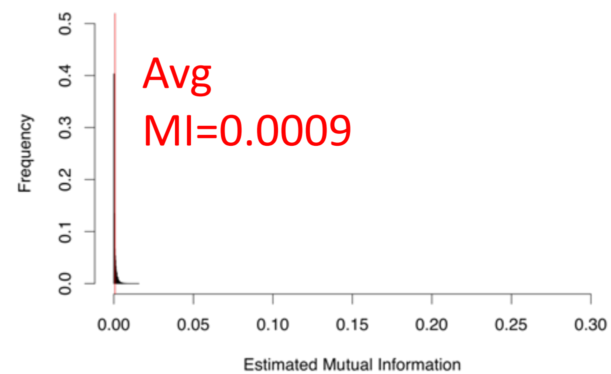
X2={0,0,1,1,.... ,0,1,1,0}

I(X1;X2) =?

100,000 times







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

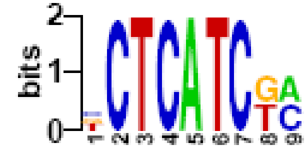
	Accuracy	Search space
Degenerate code [AC]CGATGAG[TC]	good	large
Words (*k*-mers) GCGATGAG	acceptable	small

Motif Search Algorithm

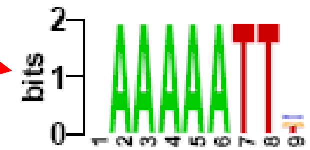
Highly informative

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
...	
...	
ACGCGCG	0.0018
CGACGCG	0.0012
TACGCTA	0.0011
ACCCCCT	0.0010
CCACGGC	0.0009
TTCAAAA	0.0005
AGACGCG	0.0004
CGAGAGC	0.0003
CTTATTA	0.0002

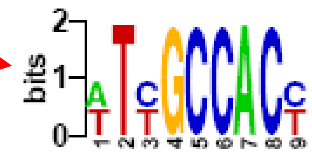
Not informative



MI=0.081



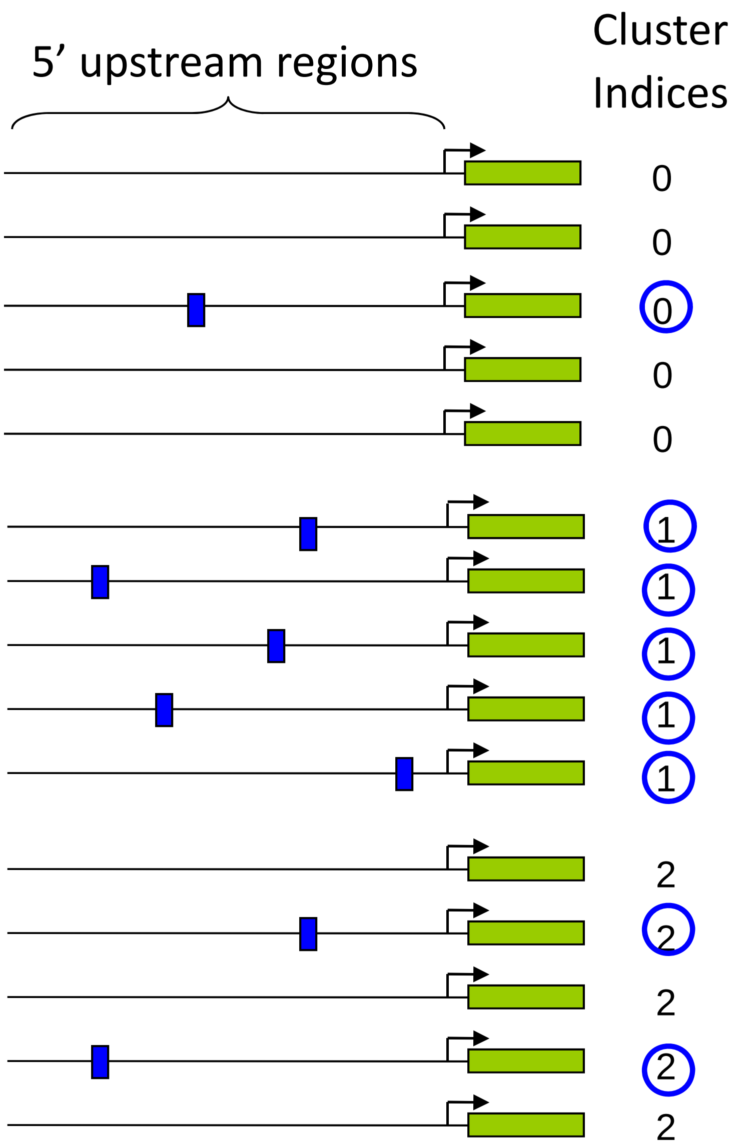
MI=0.045

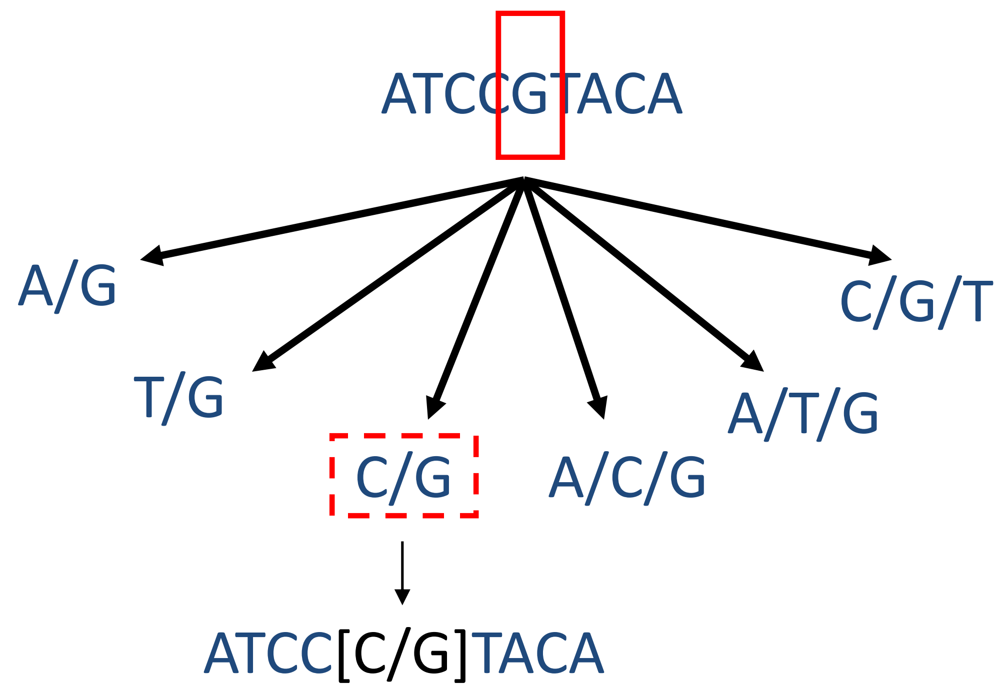


MI=0.040

...

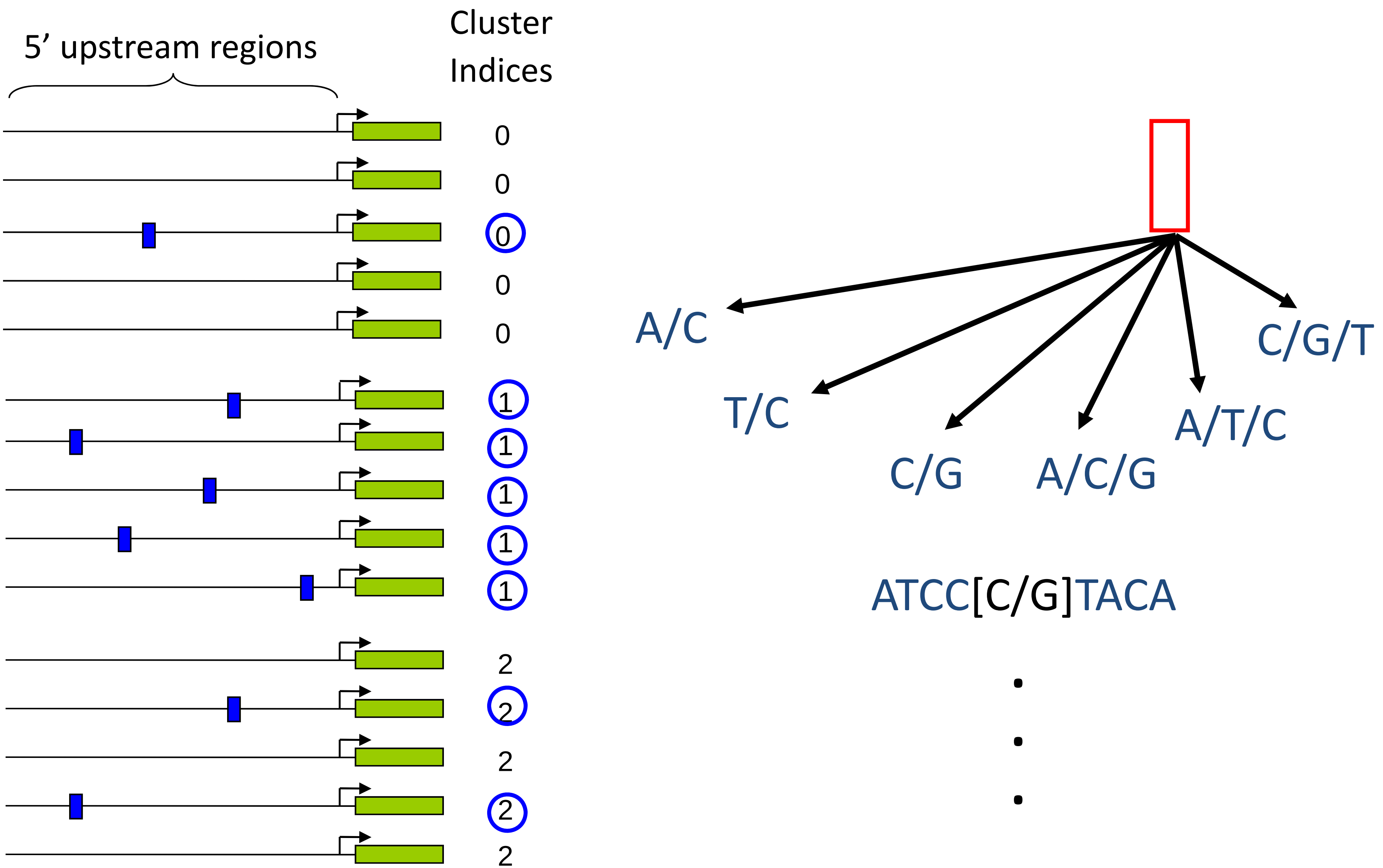
Optimizing *k*-mers into more informative degenerate motifs



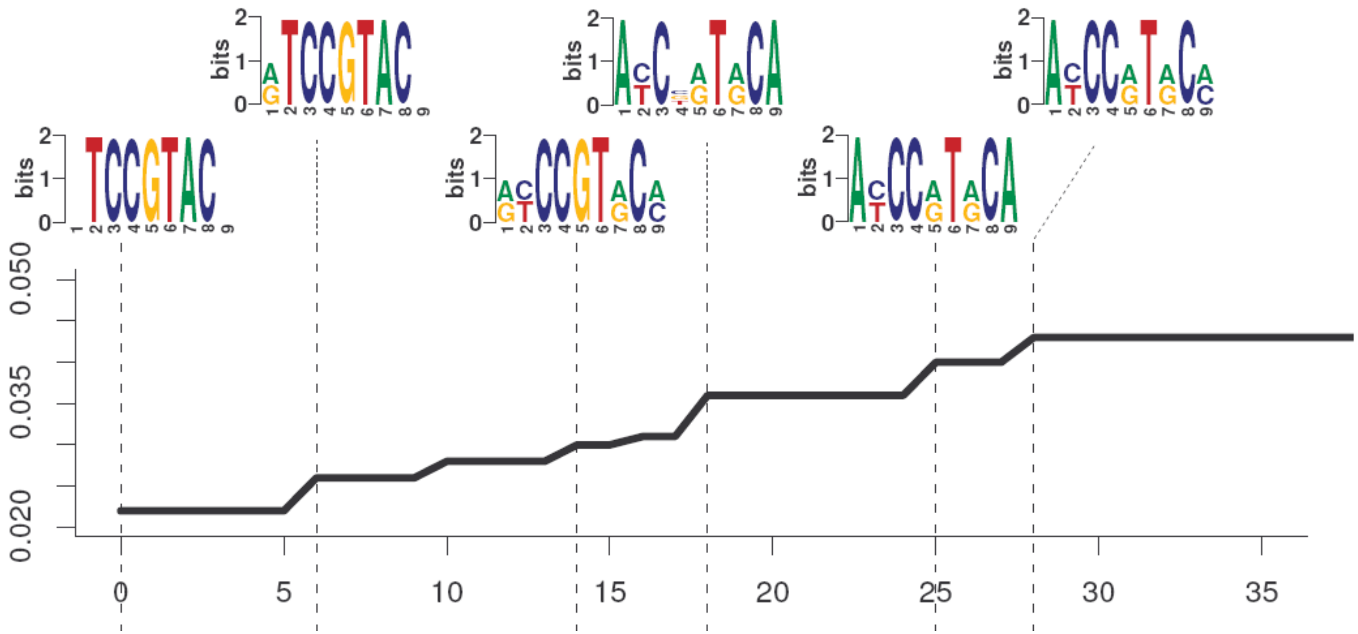


which character increases the mutual information by the largest amount ?

Optimizing *k*-mers into more informative degenerate motifs

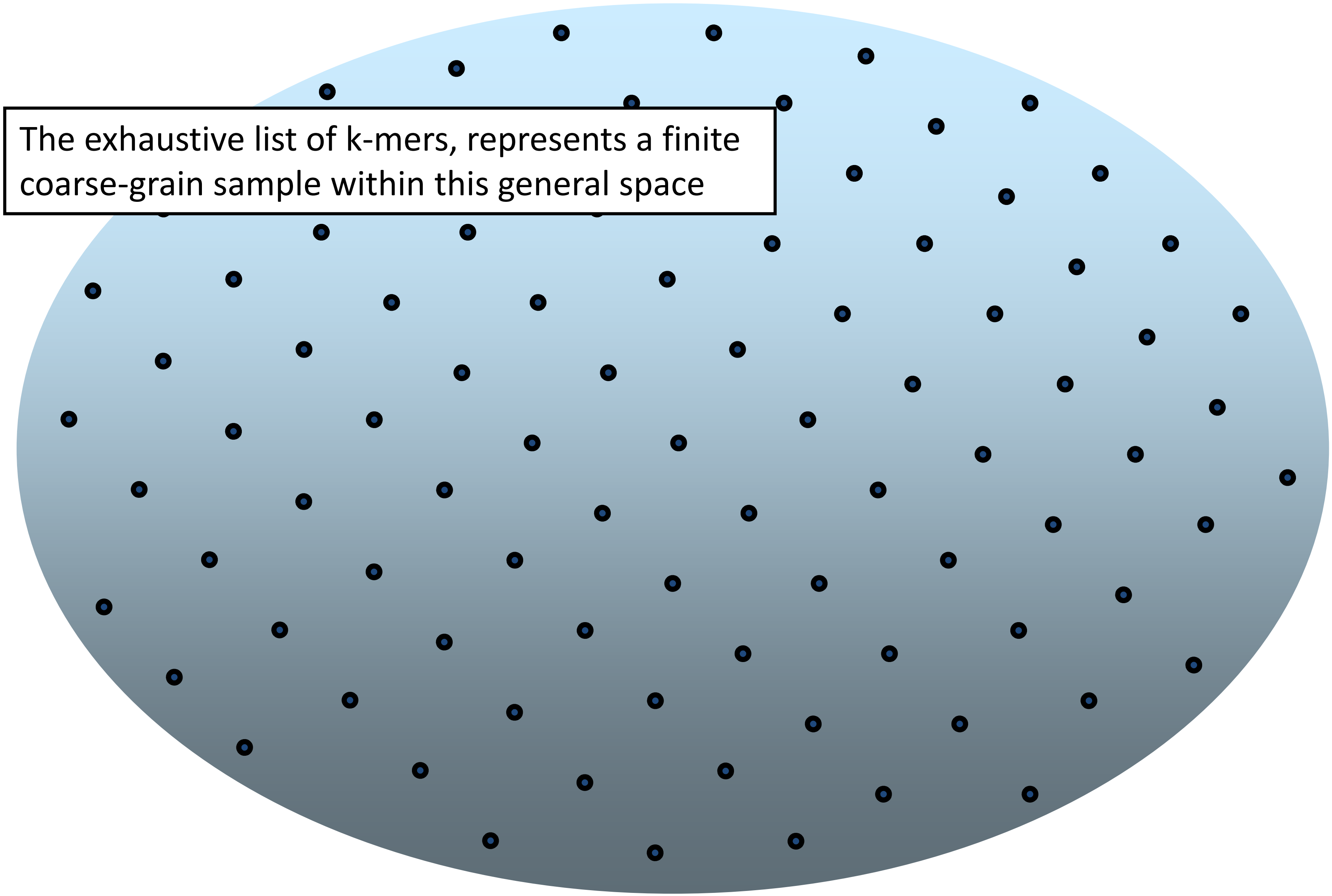


Mutual information

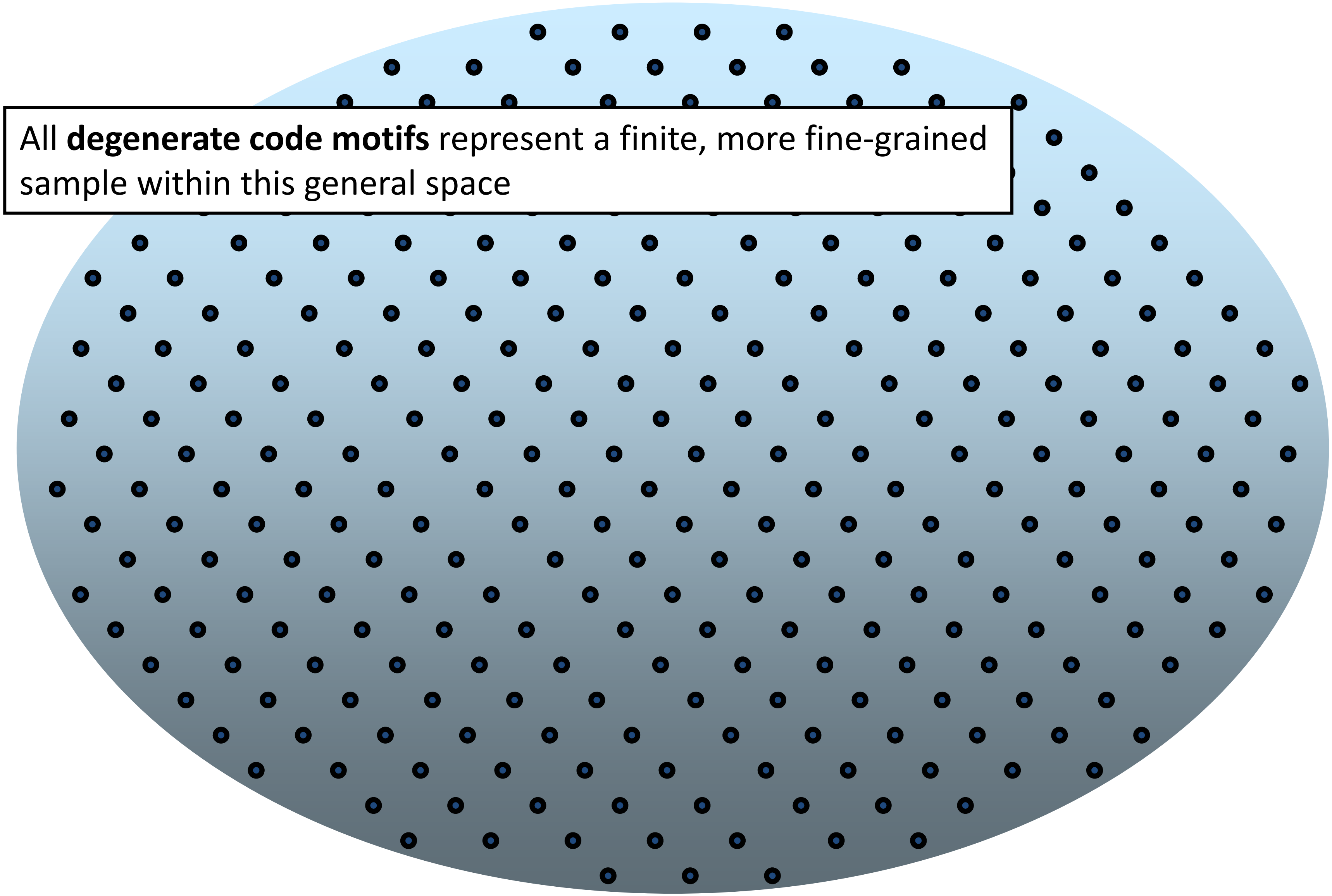


Optimization iterations (visited motif positions)

A schematic view of the optimization process

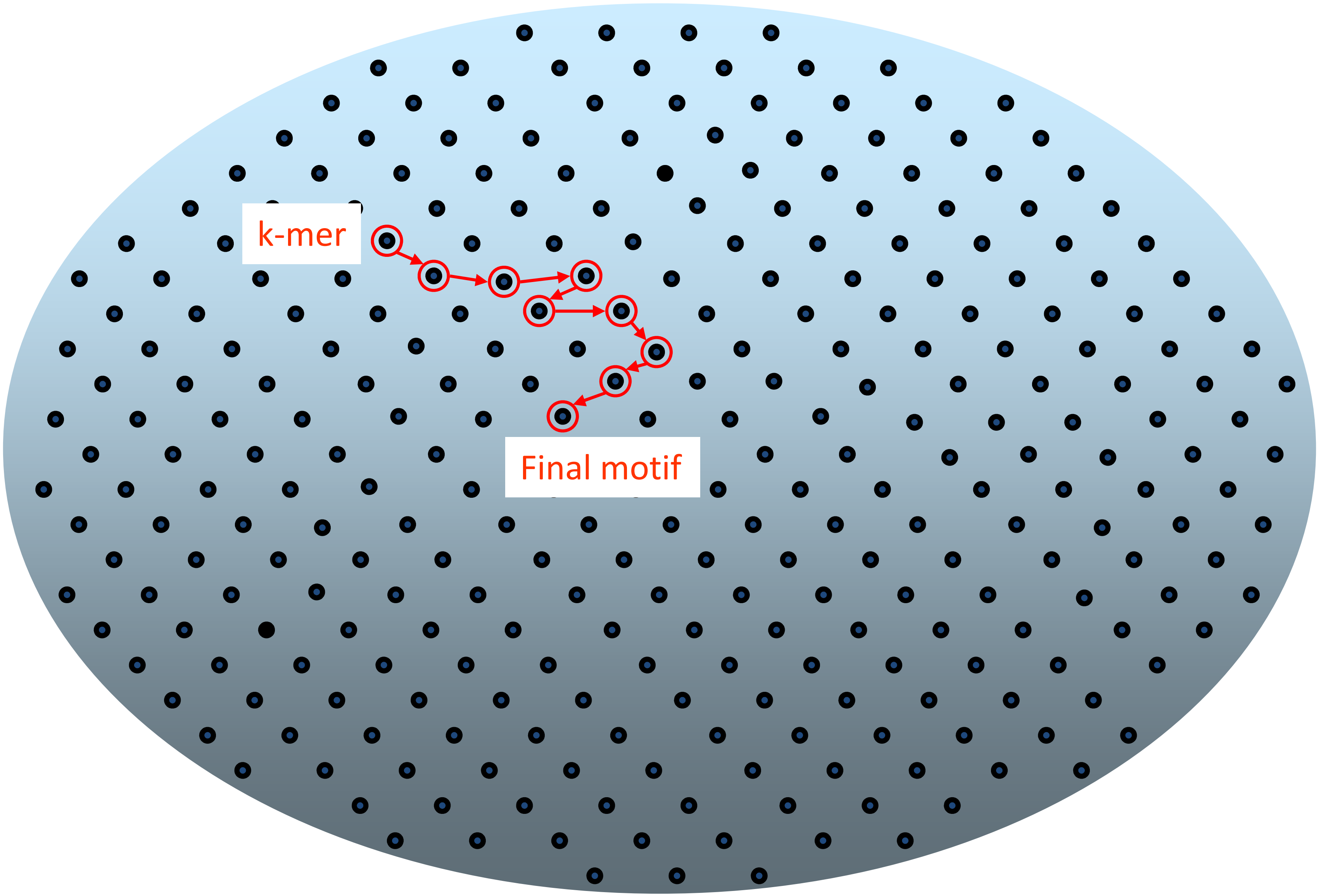


A schematic view of the optimization process

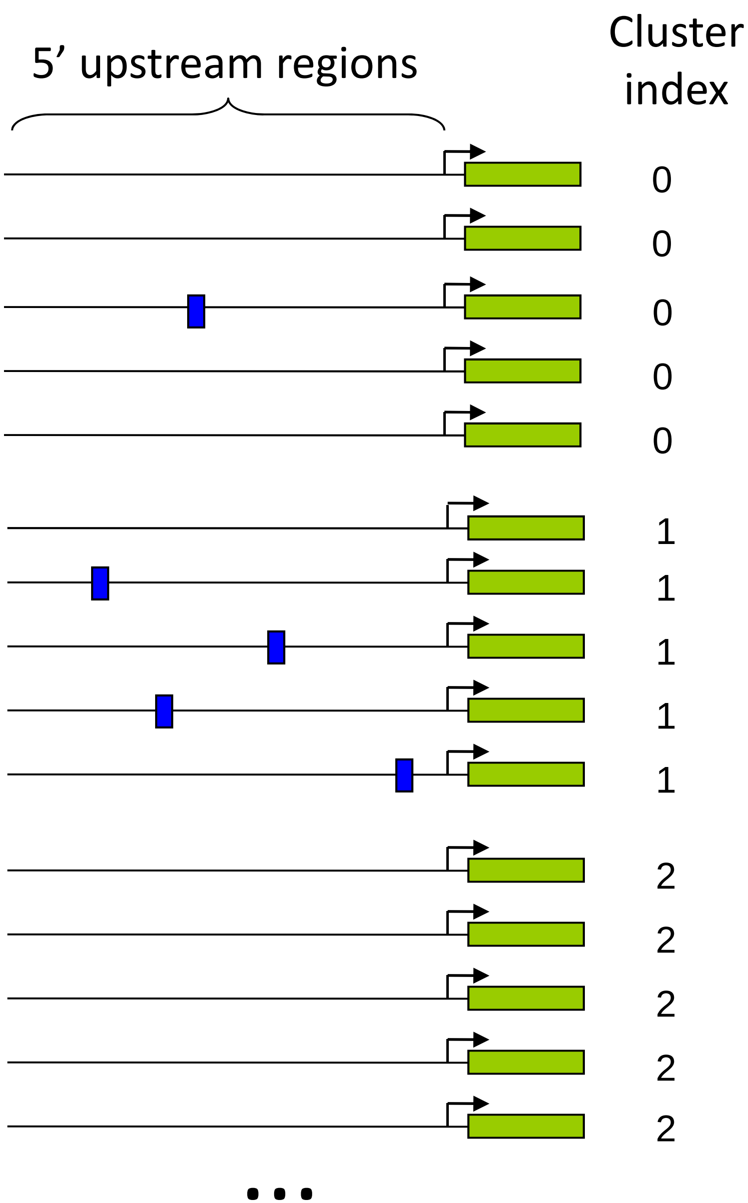


All **degenerate code motifs** represent a finite, more fine-grained sample within this general space

A schematic view of the optimization process

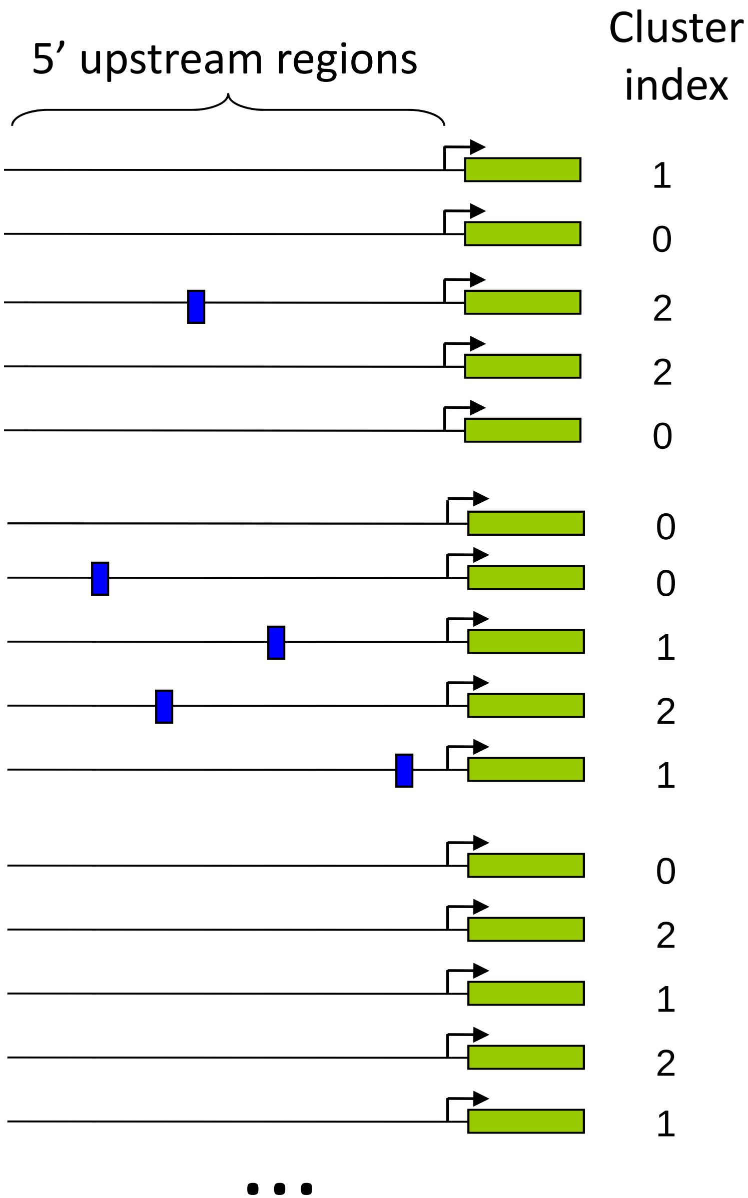


Is a given motif more informative than expected by chance ?



		Expression (Cluster Indices) 0	1	2
Motif	Absent	0.27	0.07	0.33
	Present	0.07	0.27	0.00

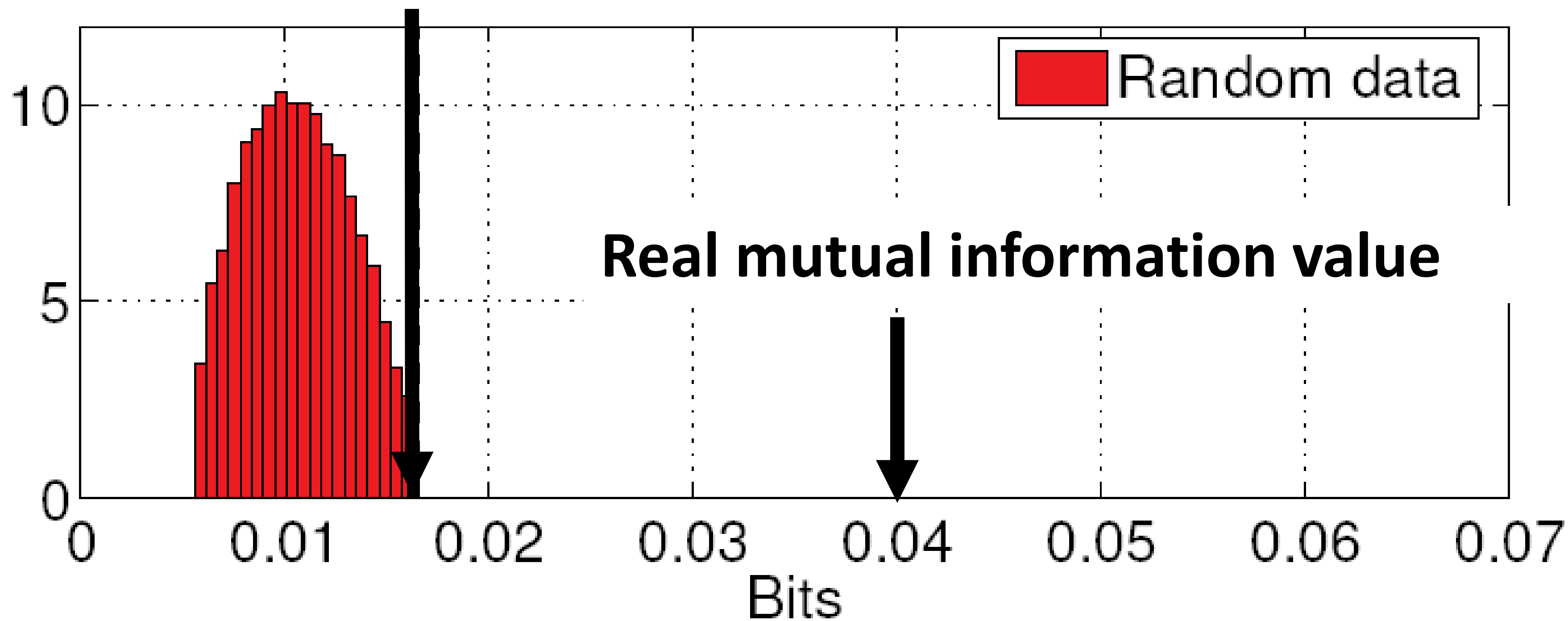
$$I(\text{motif ;expression}) = \sum_{i=1}^{2}\sum_{j=1}^{3} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$$



		Expression (Cluster Indices)		
		0	1	2
Motif	Absent	0.16	0.19	0.18
	Present	0.19	0.13	0.15

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

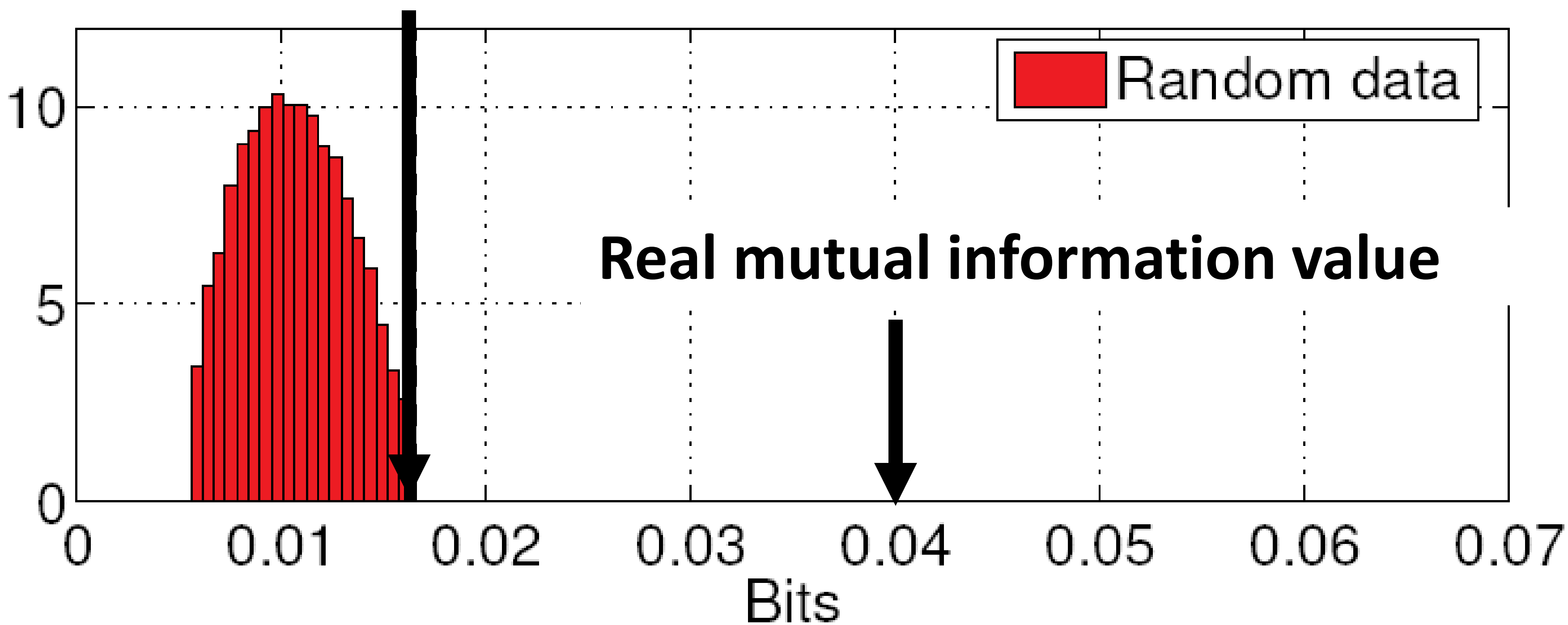
Maximum of 10,000 expression-shuffled mutual information values



P-value : probability of obtaining by chance a result at least as extreme as observed result
P(X>=x)

Maximum of 10,000 expression-shuffled mutual information values

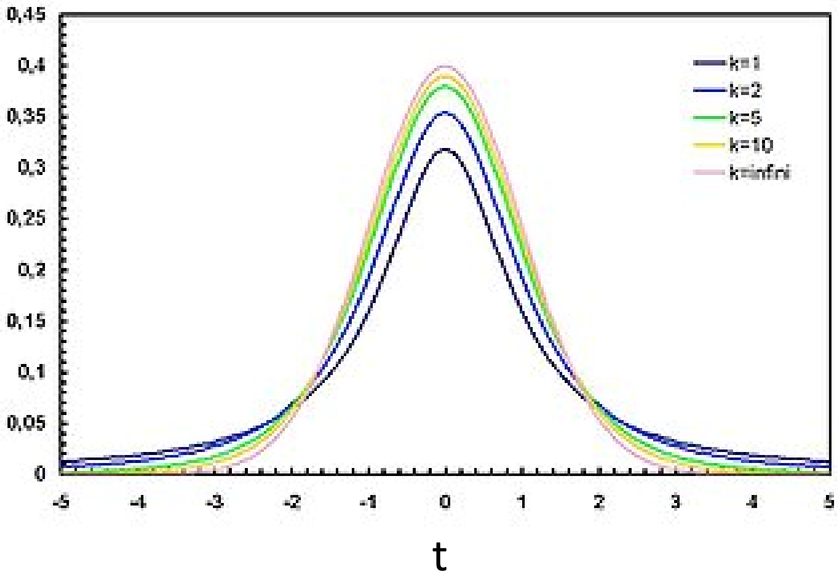
P<10^{-4}



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



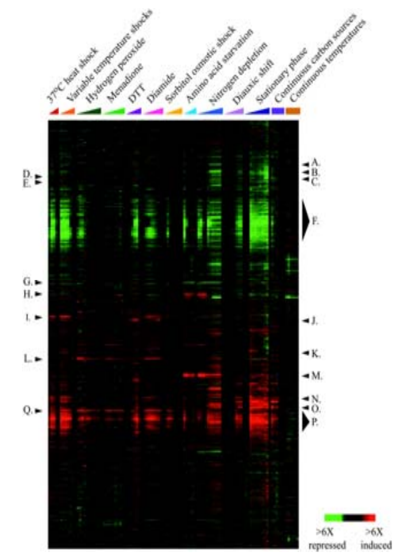
Null Distribution of information values

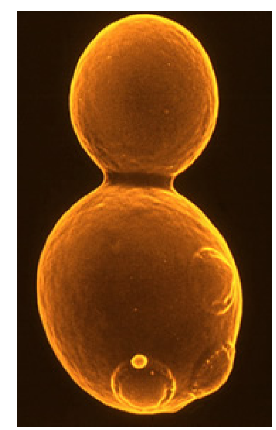
?

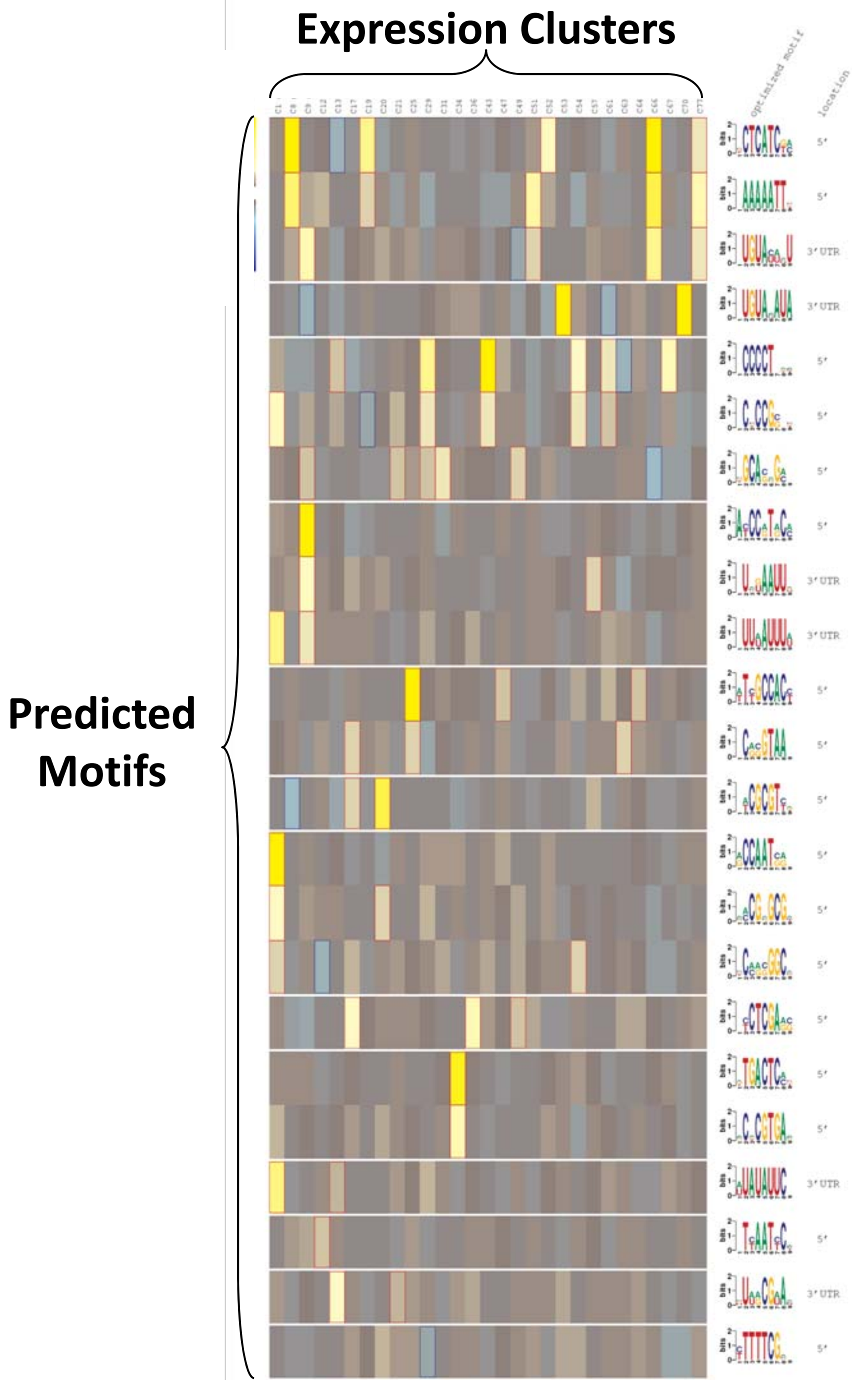
So we estimate it using simulation.

Yeast stress gene expression program (Gasch *et al*, 2000)

- 173 microarray conditions
- ~ 5,500 genes
- 80 co-expression clusters
- Runtime ~ 1h (standard PC)







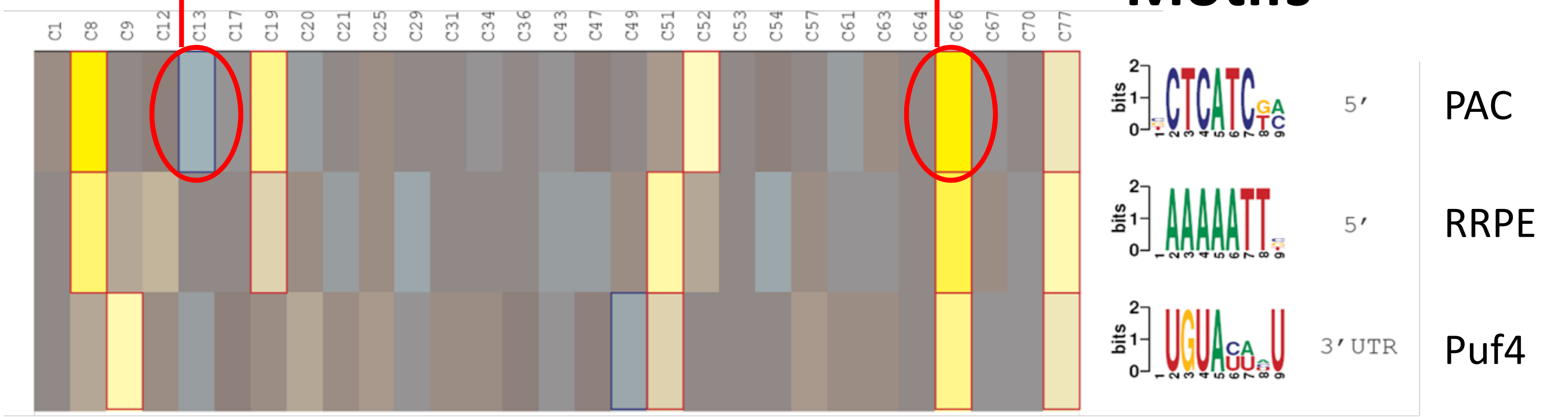
17 motifs in 5’ upstream regions
6 motifs in 3’UTRs

AC is under-represented in cluster 13 (p<1e-5)

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

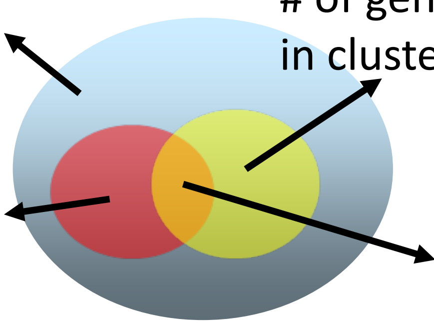
Expression Clusters

Motifs



Total # of genes

# of genes in cluster



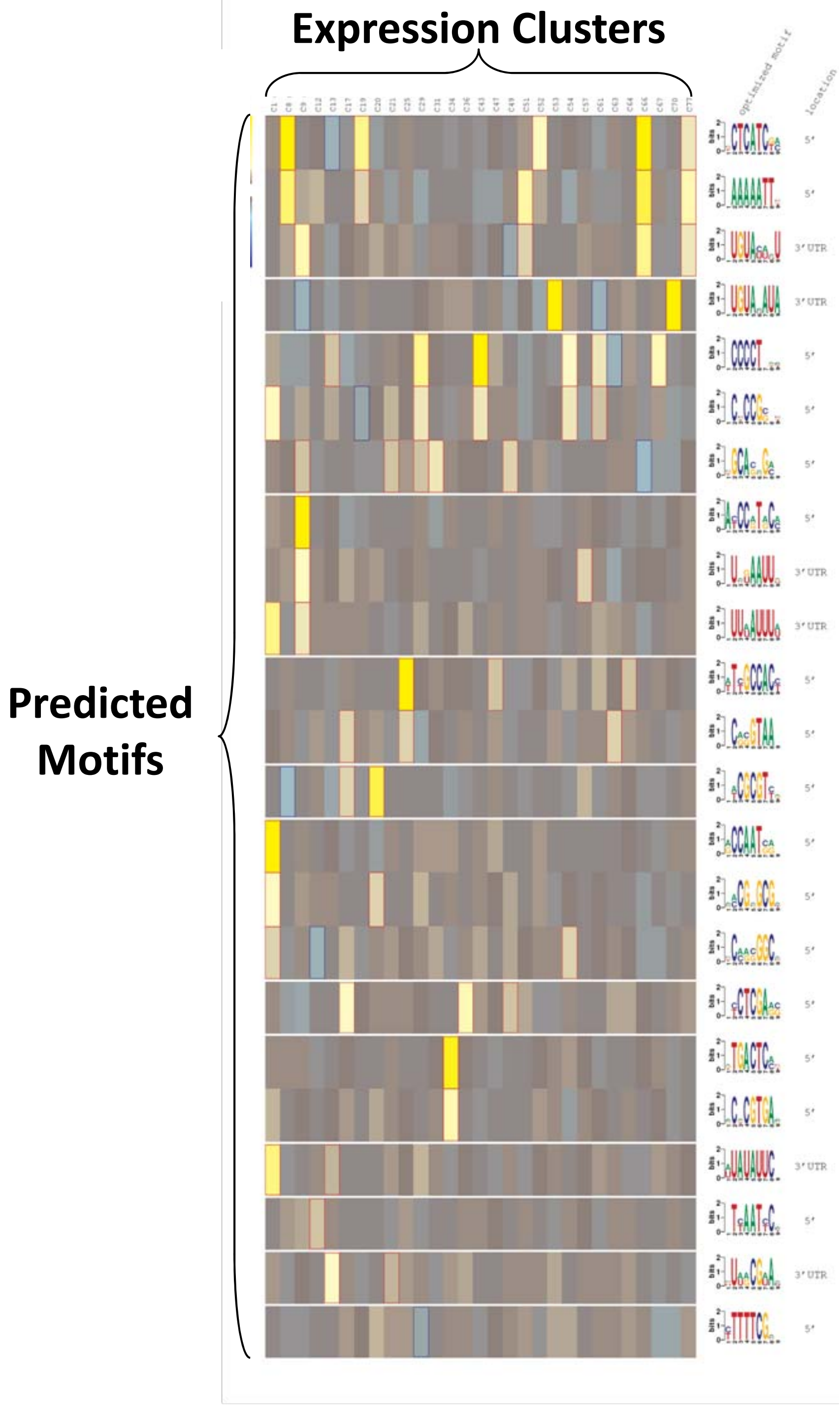
# of genes with motif

High overlap?

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

$$P(X \geq i) = \sum_{x=i}^{min(s_1,s_2)} \frac{\binom{s_1}{x}\binom{N-s_1}{s_2-x}}{\binom{N}{s_2}}$$

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 ?

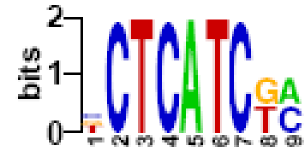
- Multiple hypothesis testing (k-mers)
- Overfit by the optimization procedure

Motif Search Algorithm

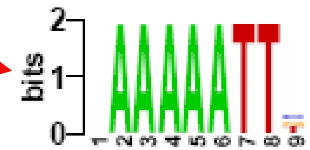
Highly informative

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
...	
...	
ACGCGCG	0.0018
CGACGCG	0.0012
TACGCTA	0.0011
ACCCCCT	0.0010
CCACGGC	0.0009
TTCAAAA	0.0005
AGACGCG	0.0004
CGAGAGC	0.0003
CTTATTA	0.0002

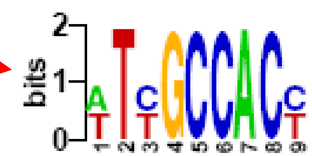
Not informative



MI=0.081



MI=0.045



MI=0.040

...

Does the algorithm overfit motifs to the expression ?

Cluster 0 : 112 genes
Cluster 1 : 132 genes ← enriched in RAP1 motif

Randomly split cluster 1 into cluster 1A and 1B

Cluster 0 : 112 genes
Cluster 1A : 66 genes

Cluster 0 : 112 genes
Cluster 1B : 66 genes

Best motif, MI=0.38 bits
[CT]CC[AG][ACT]AC[AG][CT]

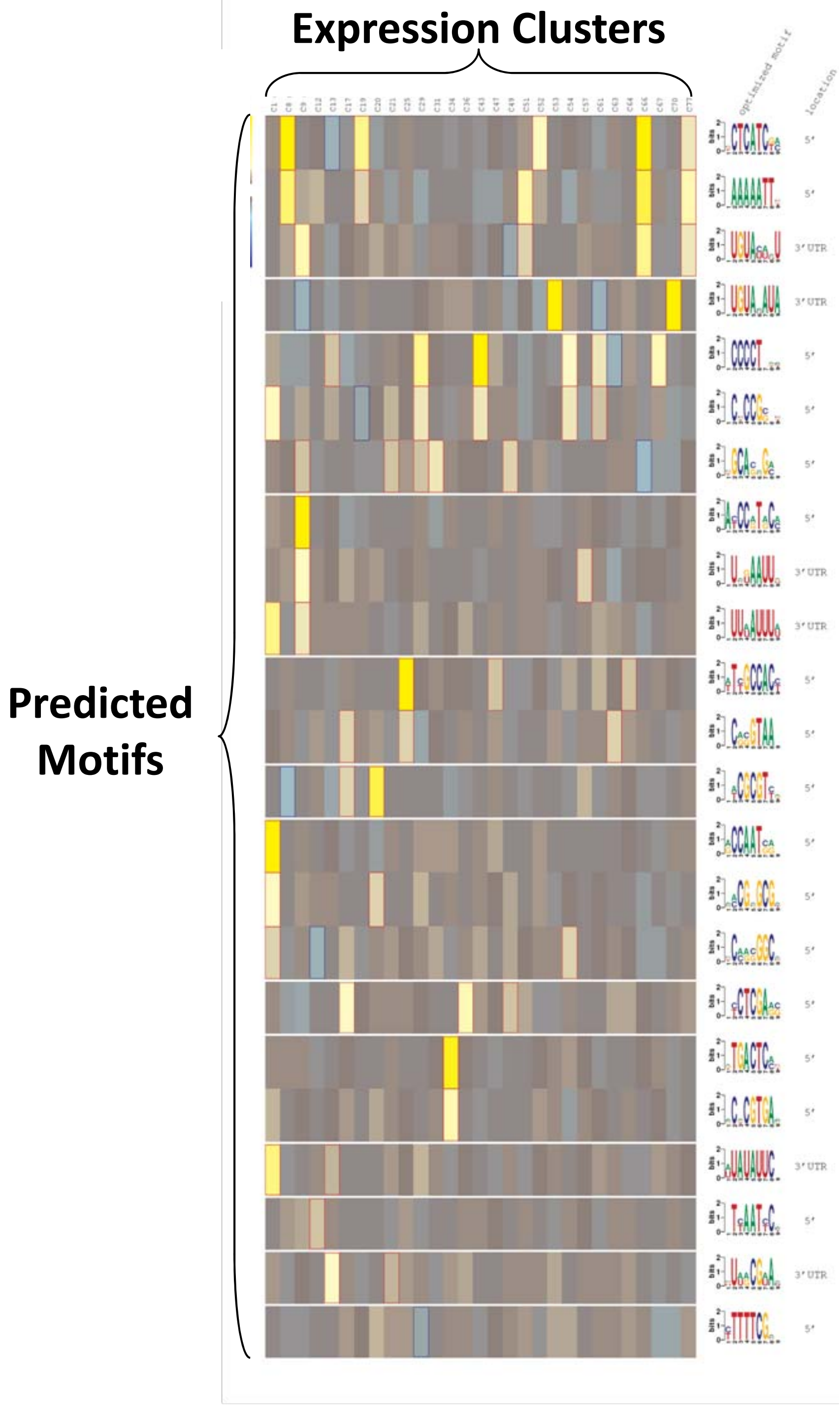
MI=0.301 bits when evaluated on this dataset

Best motif, MI=0.33 bits
[ACG][CT]CC[AG][CT][AG]C[AC]

MI=0.29 bits when evaluated on this dataset

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

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

X	P(X)	
0	0.5	H(X)=1 bit
1	0.5	

X	P(X)	
0	0.8	H(X)=0.72 bits
1	0.2	

X	P(X)	
0	1.0	H(X)=0 bits
1	0.0	

Mutual Information

$$I(\mathrm{X}\,;\mathrm{Y}) = H(Y) - H(Y \mid X)$$

Uncertainty about Y

the amount of uncertainty remaining about Y after X is known