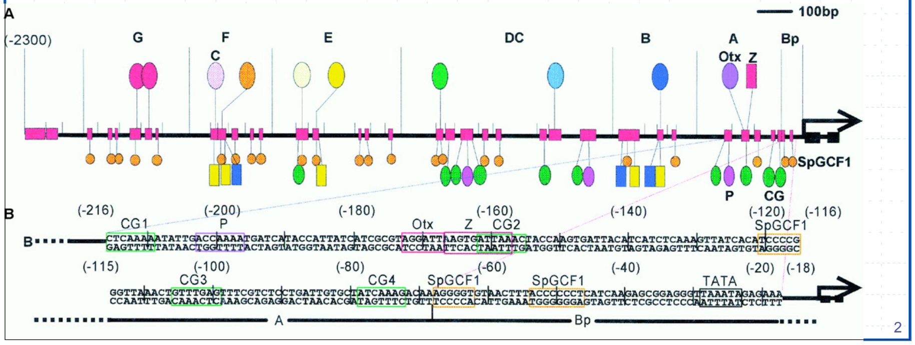
# Chapter 7: Regulatory Networks

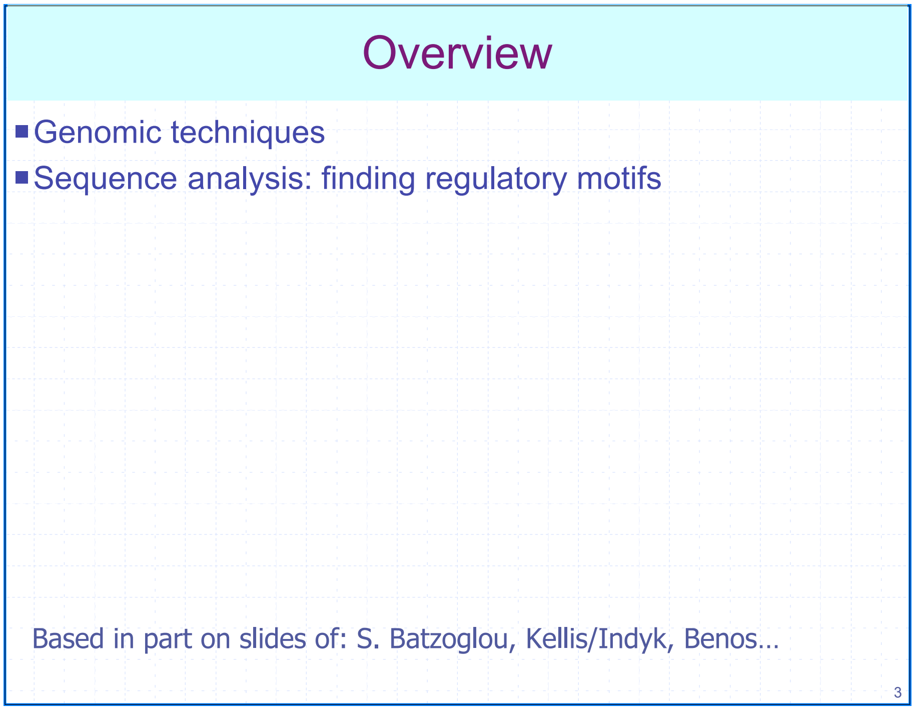
7.2 Analyzing Regulation

Prof. Yechiam Yemini (YY)
Computer Science Department
Columbia University

## The Challenge

- How do we discover regulatory mechanisms?
- Complexity: hundreds of cooperating factors
- Cis-regulation can extend over long stretches
- Trans-regulation involves long-distance interactions
- Many-to-many relationship between TF and genes

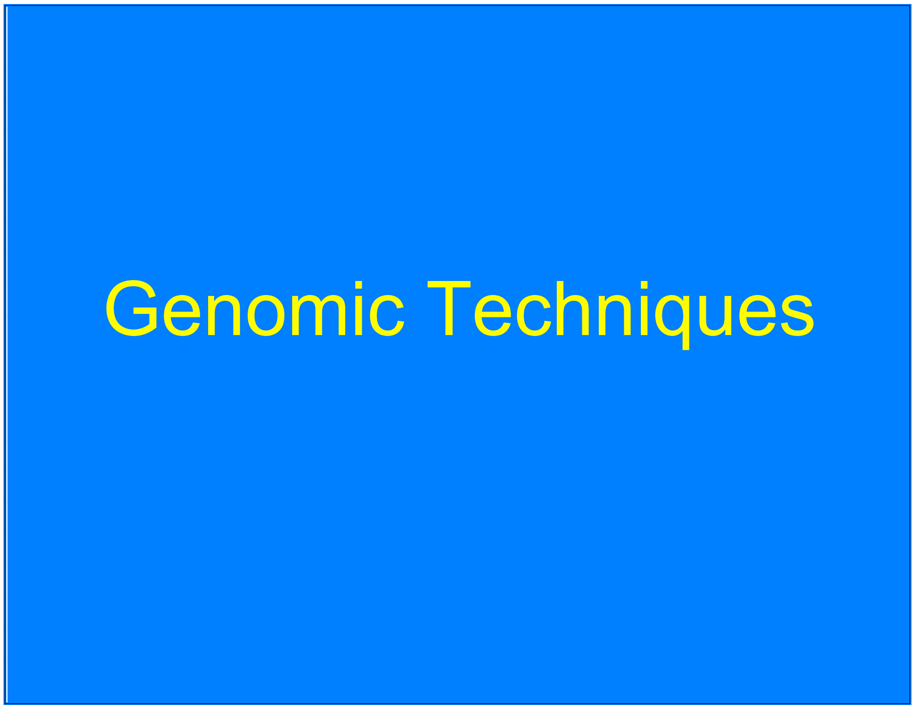

A
100bp
(-2300)
G
F
C
E
DC
B
A
Otx
Z
Bp
SpGCF1
P
CG
B
(-216)
(-200)
(-180)
(-160)
(-140)
(-120)
(-116)
CG1
P
Otx
Z
CG2
SpGCF1
CTCAAAAATATTGACCAAAATGATCATACCATTATCATCGCGTAGGATTAAGTGATTAAACTACCAAGTGATTACATCATCTCAAAGTTATCACATCCCCG
GAGTTTTTATAACTGGTTTTACTAGTATGGTAATAGTAGCGCATCCTAATTCACTAATTTGATGGTTCACTAATGTAGTAGAGTTTCAATAGTGTAGGGGC
(-115)
(-100)
(-80)
(-60)
(-40)
(-20)
(-18)
CG3
CG4
SpGCF1
SpGCF1
TATA
GGTTAAACTGTTTGAGTTTCGTCTCCTGATTGTGCTATCAAAGACAAAGGGGTGTAACTTTACCCCCCTCATCAAGAGCGGAGGGTTAAATAGAGAAA
CCAATTTGACAAACTCAAAGCAGAGGACTAACACGATAGTTTCTGTTTCCCCACATTGAAATGGGGGGAGTAGTTCTCGCCTCCCAATTTATCTCTTT
A
Bp



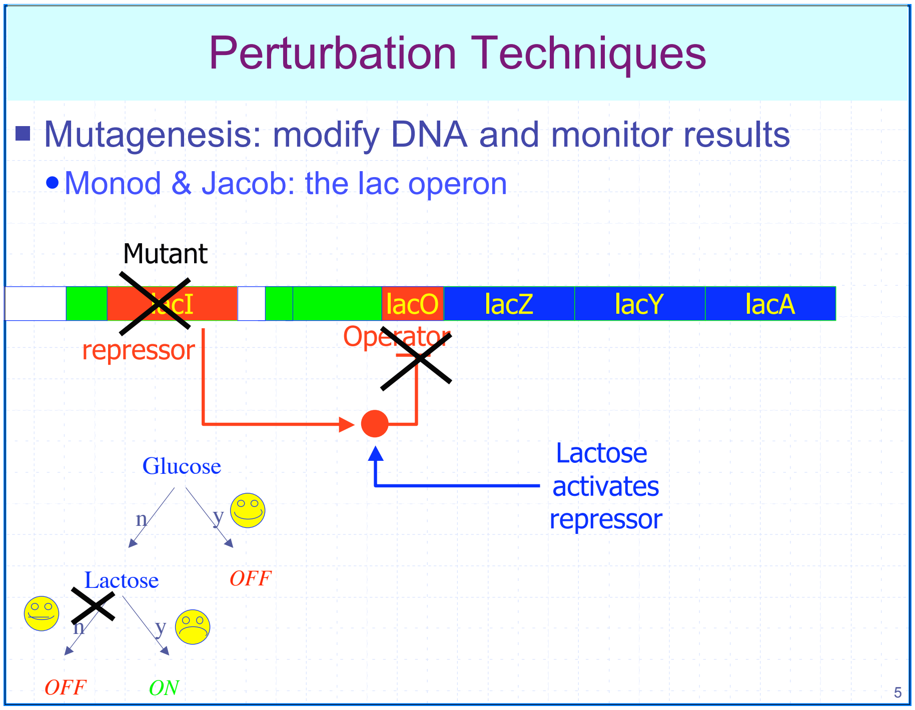
## Overview

- Genomic techniques
- Sequence analysis: finding regulatory motifs

Based in part on slides of: S. Batzoglou, Kellis/Indyk, Benos…



# Genomic Techniques

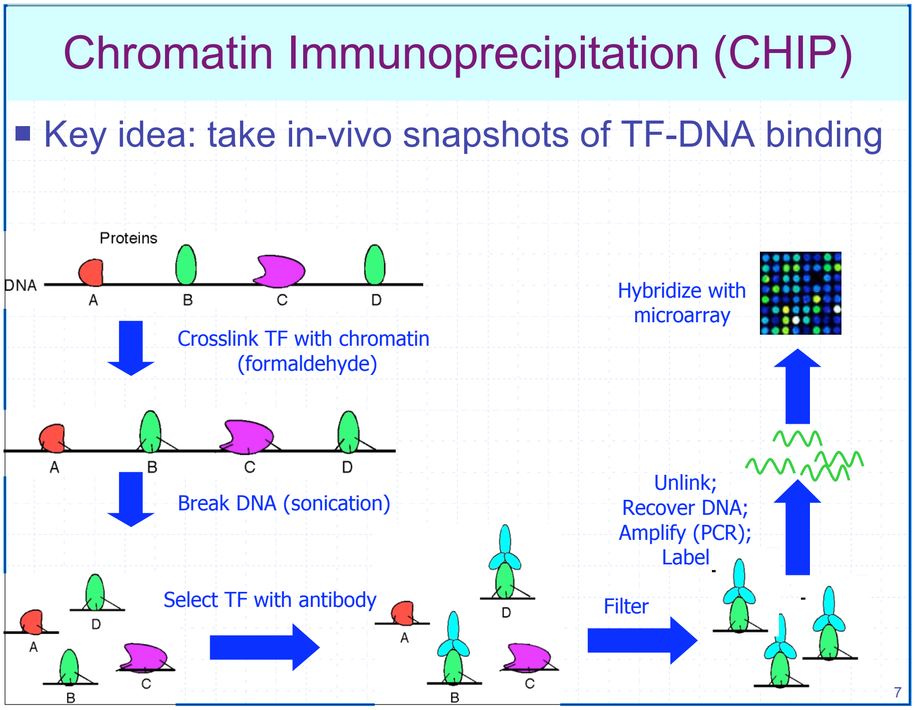


## Perturbation Techniques

- Mutagenesis: modify DNA and monitor results
  - Monod & Jacob: the lac operon

## Perturbations

- Perturb DNA: regulatory gene, promoter region…
- Perturb RNA expression: RNAi…
- Key challenge: high-thruput techniques



# Chromatin Immunoprecipitation (CHIP)

- Key idea: take in-vivo snapshots of TF-DNA binding

Crosslink TF with chromatin (formaldehyde)

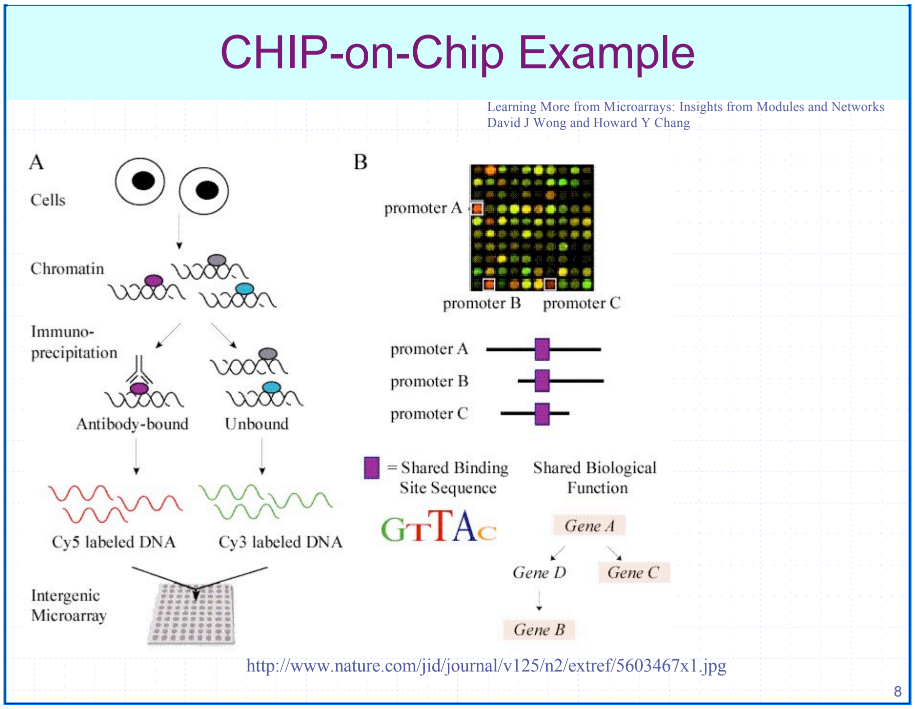
Break DNA (sonication)

Select TF with antibody

Filter

Unlink; Recover DNA; Amplify (PCR); Label

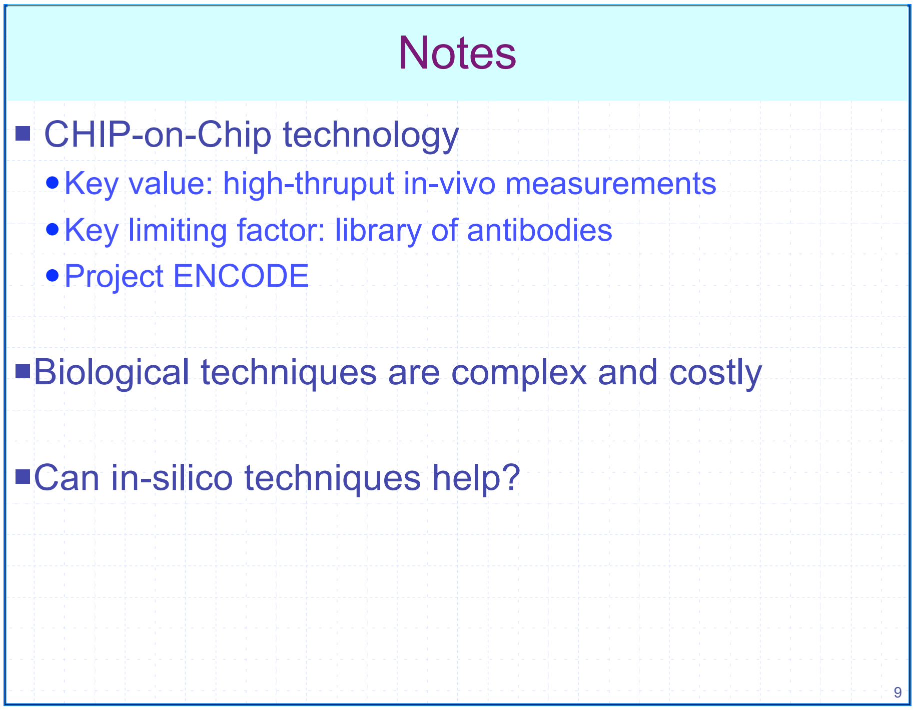
Hybridize with microarray



# CHIP-on-Chip Example

Learning More from Microarrays: Insights from Modules and Networks
David J Wong and Howard Y Chang

http://www.nature.com/jid/journal/v125/n2/extref/5603467x1.jpg



## Notes

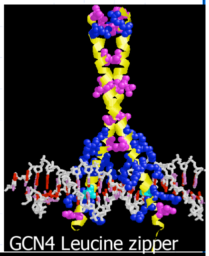
- CHIP-on-Chip technology
  - Key value: high-thruput in-vivo measurements
  - Key limiting factor: library of antibodies
  - Project ENCODE
- Biological techniques are complex and costly
- Can in-silico techniques help?

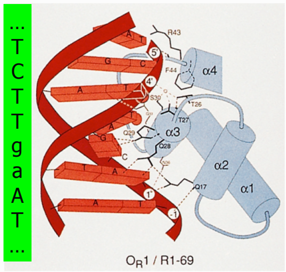
9

# Regulatory motifs

## Transcription Factors (TF) Bind To DNA

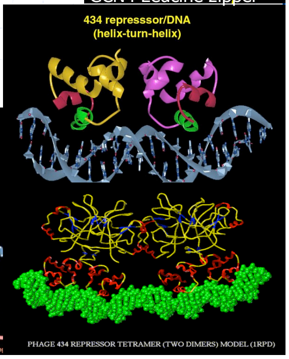
- TF active site binds with DNA
- Several binding mechanisms exist
  - Helix-Turn-Helix
  - Leucine zipper
  - Zinc finger
- How do TFs know where to bind?
  - Find sequence motifs (signals)



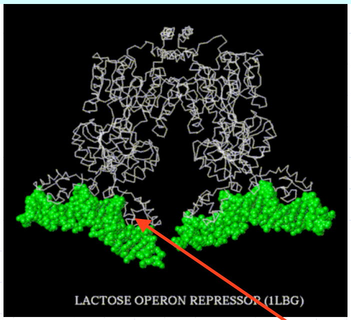


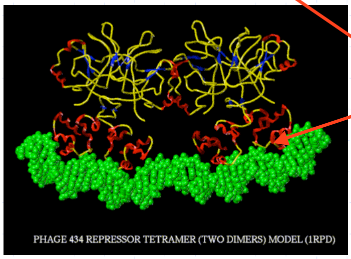
Courtesy of Aneel Aggarwal, John Anderson, and Stephen Harrison, Harvard University.

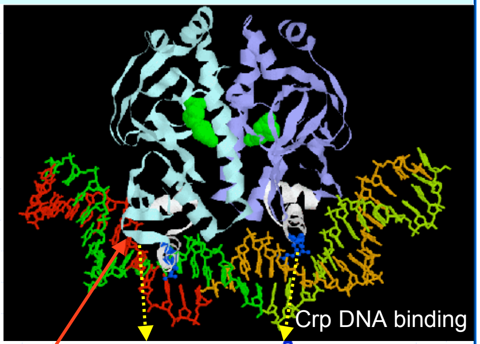
zif268/DNA
(zinc finger)

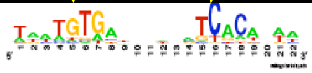


## DNA Motifs Signal Binding Sites

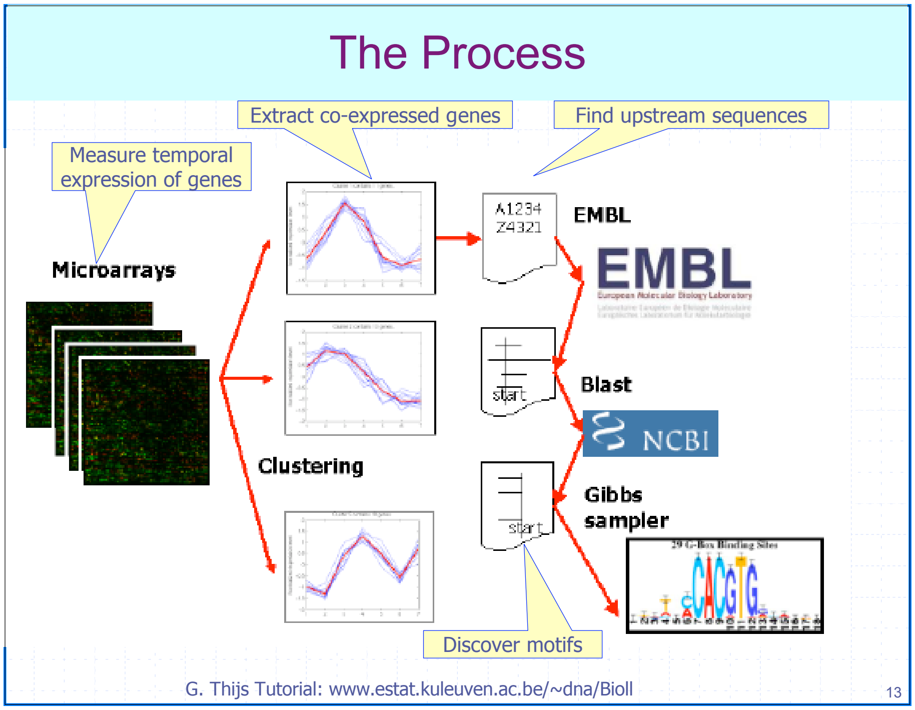








TF recognize motifs



## The Process

Measure temporal expression of genes

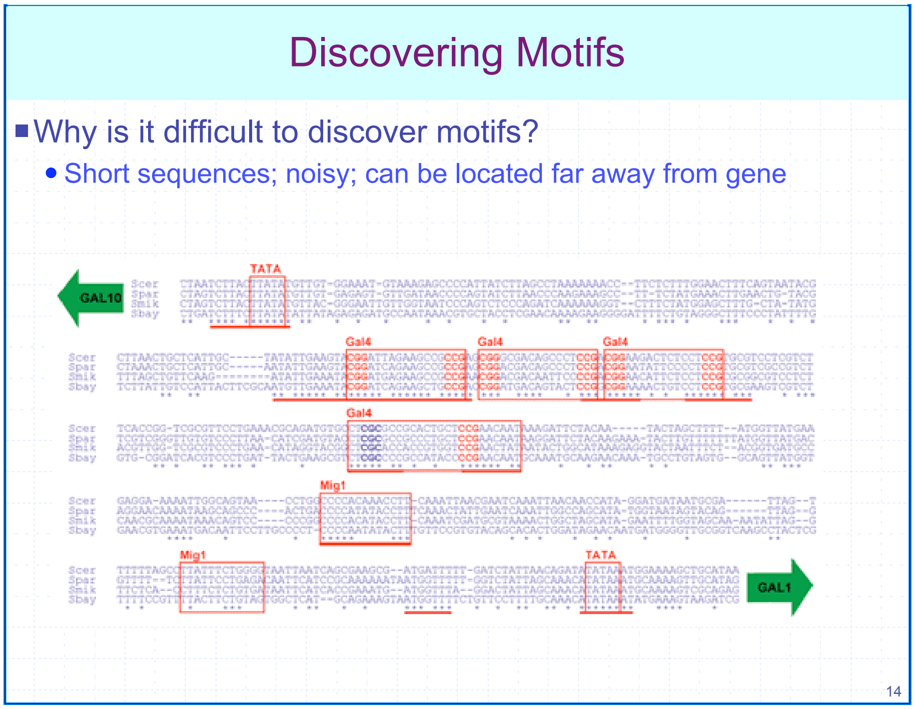
Extract co-expressed genes

Find upstream sequences

Discover motifs

G. Thijs Tutorial: www.estat.kuleuven.ac.be/~dna/Bioll

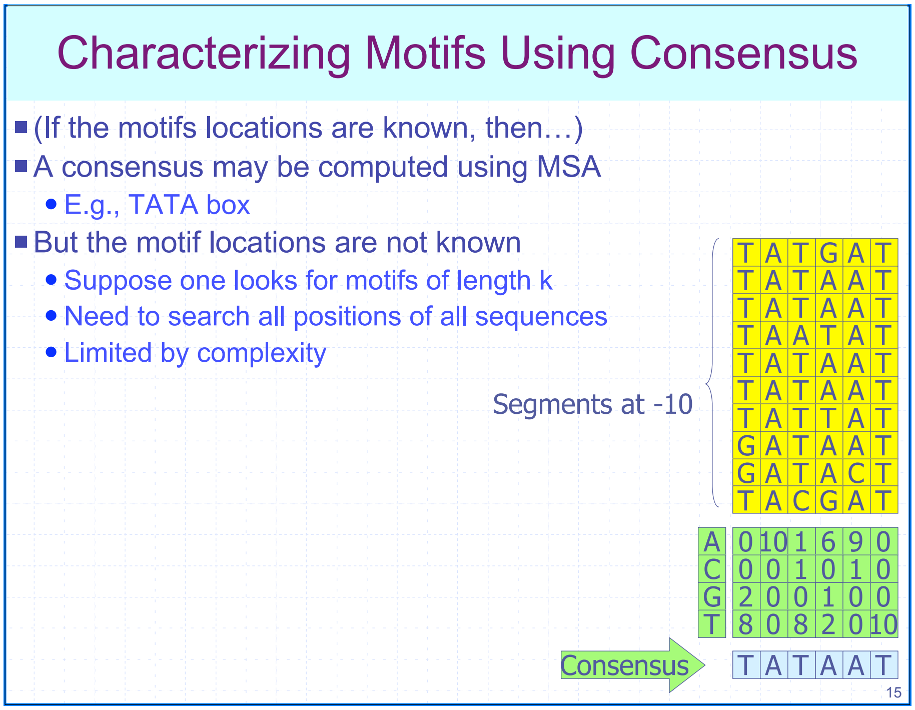
13



## Discovering Motifs

- Why is it difficult to discover motifs?
  - Short sequences; noisy; can be located far away from gene

14



## Characterizing Motifs Using Consensus

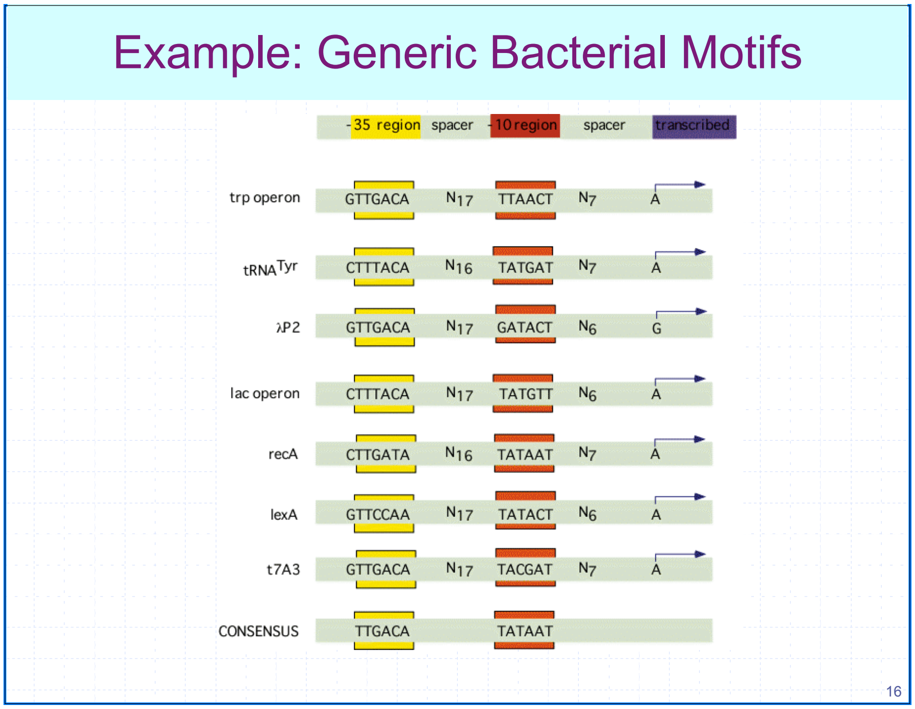
- (If the motifs locations are known, then…)
- A consensus may be computed using MSA
  - E.g., TATA box
- But the motif locations are not known
  - Suppose one looks for motifs of length k
  - Need to search all positions of all sequences
  - Limited by complexity

Segments at -10

| | | | | | |
|---|---|---|---|---|---|
| T | A | T | G | A | T |
| T | A | T | A | A | T |
| T | A | T | A | A | T |
| T | A | A | T | A | T |
| T | A | T | A | A | T |
| T | A | T | A | A | T |
| T | A | T | T | A | T |
| G | A | T | A | A | T |
| G | A | T | A | C | T |
| T | A | C | G | A | T |

| | | | | | | |
|---|---|---|---|---|---|---|
| A | 0 | 10 | 1 | 6 | 9 | 0 |
| C | 0 | 0 | 1 | 0 | 1 | 0 |
| G | 2 | 0 | 0 | 1 | 0 | 0 |
| T | 8 | 0 | 8 | 2 | 0 | 10 |

Consensus → T A T A A T



## Example: Generic Bacterial Motifs

| | -35 region | spacer | -10 region | spacer | transcribed |
|---|---|---|---|---|---|
| trp operon | GTTGACA | $N_{17}$ | TTAACT | $N_7$ | A |
| tRNA$^{Tyr}$ | CTTTACA | $N_{16}$ | TATGAT | $N_7$ | A |
| λP2 | GTTGACA | $N_{17}$ | GATACT | $N_6$ | G |
| lac operon | CTTTACA | $N_{17}$ | TATGTT | $N_6$ | A |
| recA | CTTGATA | $N_{16}$ | TATAAT | $N_7$ | A |
| lexA | GTTCCAA | $N_{17}$ | TATACT | $N_6$ | A |
| t7A3 | GTTGACA | $N_{17}$ | TACGAT | $N_7$ | A |
| CONSENSUS | TTGACA | | TATAAT | | |

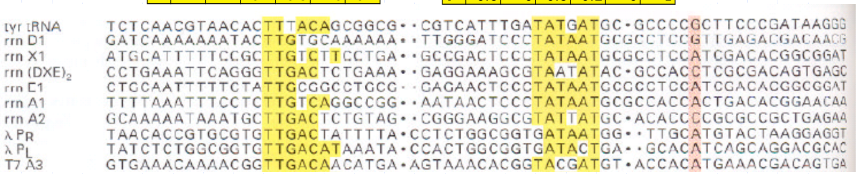
# From Consensus To Probability Profiles

- Consensus sequences do not provide the full story
- Probability of nucleotides is more telling
- Binding motifs may be represented by probability matrices (aka weight/frequency matrix)

Position Weight Matrix (PWM)

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0.6 | 0 | 0.4 |
| C | 0 | 0 | 0 | 0.1 | 0.8 | 0.1 |
| G | 0 | 0 | 0.9 | 0 | 0.2 | 0.1 |
| T | 1 | 1 | 0.1 | 0.3 | 0 | 0.4 |

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 0.1 | 0.6 | 0.9 | 0 |
| C | 0 | 0 | 0.1 | 0 | 0.1 | 0 |
| G | 0.2 | 0 | 0 | 0.1 | 0 | 0 |
| T | 0.8 | 0 | 0.8 | 0.2 | 0 | 1 |



```
tyr tRNA     TCTCAACGTAACACTTTACAGCGGCG••CGTCATTTGATATGATGC•GCCCCGCTTCCCGATAAGGG
rrn D1       GATCAAAAAAATACTTGTGCAAAAAA••TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1       ATGCATTTTTCCGCTTGTCTTCCTGA••GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2   CCTGAAATTCAGGGTTGACTCTGAAA••GAGGAAAGCGTAATATAC•GCCACCTCGCGACAGTGAGC
rrn E1       CTGCAATTTTTCTATTGCGGCCTGCG••GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1       TTTTAAATTTCCTCTTGTCAGGCCGG••AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2       GCAAAAATAAATGCTTGACTCTGTAG••CGGGAAGGCGTATTATGC•ACACCCCGCGCCGCTGAGAA
λ PR         TAACACCGTGCGTGTTGACTATTTTA•CCTCTGGCGGTGATAATGG••TTGCATGTACTAAGGAGGT
λ PL         TATCTCTGGCGGTGTTGACATAAATA•CCACTGGCGGTGATACTGA••GCACATCAGCAGGACGCAC
T7 A3        GTGAAACAAAACGGTTGACAACATGA•AGTAAACACGGTACGATGT•ACCACATGAAACGACAGTGA
```

# From Probability To Log-odds Scoring

```
tyr tRNA     TCTCAACGTAACACTTTACAGCGGCG••CGTCATTTGATATGATGC•GCCCCGCTTCCCGATAAGGG
rrn D1       GATCAAAAAAATACTTGTGCAAAAAA••TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1       ATGCATTTTTCCGCTTGTCTTCCTGA••GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2   CCTGAAATTCAGGGTTGACTCTGAAA••GAGGAAAGCGTAATATAC•GCCACCTCGCGACAGTGAGC
rrn E1       CTGCAATTTTTCTATTGCGGCCTGCG••GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1       TTTTAAATTTCCTCTTGTCAGGCCGG••AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2       GCAAAAATAAATGCTTGACTCTGTAG••CGGGAAGGCGTATTATGC•ACACCCCGCGCCGCTGAGAA
λ PR         TAACACCGTGCGTGTTGACTATTTTA•CCTCTGGCGGTGATAATGG••TTGCATGTACTAAGGAGGT
λ PL         TATCTCTGGCGGTGTTGACATAAATA•CCACTGGCGGTGATACTGA••GCACATCAGCAGGACGCAC
T7 A3        GTGAAACAAAACGGTTGACAACATGA•AGTAAACACGGTACGATGT•ACCACATGAAACGACAGTGA
```

PWM

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 0.1 | 0.6 | 0.9 | 0 |
| C | 0 | 0 | 0.1 | 0 | 0.1 | 0 |
| G | 0.2 | 0 | 0 | 0.1 | 0 | 0 |
| T | 0.8 | 0 | 0.8 | 0.2 | 0 | 1 |

Log-odds

Position Specific Scoring Matrix (PSSM)

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | -3.32 | 2.04 | -1 | 1.32 | 1.89 | -3.32 |
| C | -3.32 | -3.32 | -1 | -3.32 | -1 | -3.32 |
| G | -0.15 | -3.32 | -3.32 | -1 | -3.32 | -3.32 |
| T | 1.72 | -3.32 | 1.72 | -0.15 | -3.32 | 2.04 |

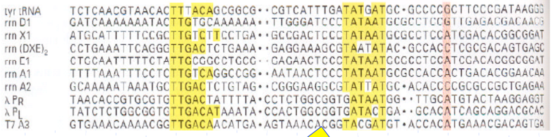
Score=log($p_\alpha$/$b_\alpha$)

$p_\alpha$= probability of nucleotide $\alpha$

$b_\alpha$= background probability of $\alpha$

The background probability may be computed from the sequences. In here we use $b_\alpha$=0.25 for all nucleotides $\alpha$.

## Logo Profiles Represent Motif Signal



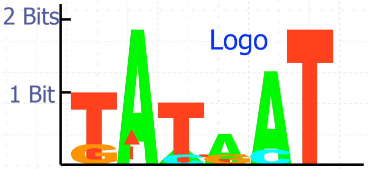
```
tyr tRNA   TCTCAACGTAACACTTTACAGCGGCG··CGTCATTTGATATGATGC·GCCCCGCTTCCCGATAAGGG
rrn D1     GATCAAAAAAATACTTGTGCAAAAAA··TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1     ATGCATTTTTCCGCTTGTCTTCCTGA··GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2 CCTGAAATTCAGGGTTGACTCTGAAA··GAGGAAAGCGTAATATAC·GCCACCTCGCGACAGTGAGC
rrn E1     CTGCAATTTTTCTATTGCGGCCTGCG··GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1     TTTTAAATTTCCTCTTGTCAGGCCGG··AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2     GCAAAAATAAATGCTTGACTCTGTAG··CGGGAAGGCGTATTATGC·ACACCCCGCGCCGCTGAGAA
λ PR       TAACACCGTGCGTGTTGACTATTTTA·CCTCTGGCGGTGATAATGG··TTGCATGTACTAAGGAGGT
λ PL       TATCTCTGGCGGTGTTGACATAAATA·CCACTGGCGGTGATACTGA··GCACATCAGCAGGACGCAC
T7 A3      GTGAAACAAAACGGTTGACAACATGA·AGTAAACACGGTACGATGT·ACCACATGAAACGACAGTGA
```

PWM

|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 0 | 1 | 0.1 | 0.6 | 0.9 | 0 |
| C | 0 | 0 | 0.1 | 0 | 0.1 | 0 |
| G | 0.2 | 0 | 0 | 0.1 | 0 | 0 |
| T | 0.8 | 0 | 0.8 | 0.2 | 0 | 1 |

PSSM

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| A | -3.32 | 2.04 | -1 | 1.32 | 1.89 | -3.32 |
| C | -3.32 | -3.32 | -1 | -3.32 | -1 | -3.32 |
| G | -0.15 | -3.32 | -3.32 | -1 | -3.32 | -3.32 |
| T | 1.72 | -3.32 | 1.72 | -0.15 | -3.32 | 2.04 |



|   | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| A | 0.0 | 1.9 | 0.1 | 0.4 | 1.3 | 0.0 |
| C | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.0 |
| G | 0.3 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 |
| T | 1.0 | 0.0 | 0.9 | 0.1 | 0.0 | 1.9 |

$p_\alpha$*D(p||b)

Contribution of a given nucleotide to the signal

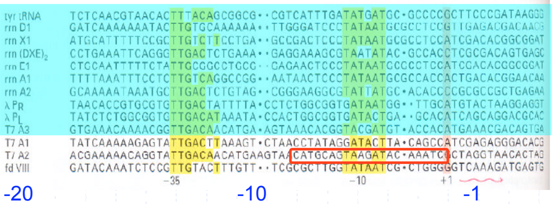
The expected log-odds score of a position (column) corresponds to its relative entropy, aka Kullback-Liebler divergence:

$$D(p||b)=\Sigma_{\alpha\varepsilon\{A,C,T,G\}}p_\alpha\log(p_\alpha/b_\alpha)$$

D(p||b) measures the information signal of a position.

19

## PSSM Profile May Be Aligned With Sequence To Detect Signal (Motif)



```
tyr tRNA   TCTCAACGTAACACTTTACAGCGGCG··CGTCATTTGATATGATGC·GCCCCGCTTCCCGATAAGGG
rrn D1     GATCAAAAAAATACTTGTGCAAAAAA··TTGGGATCCCTATAATGCGCCTCCGTTGAGACGACAACG
rrn X1     ATGCATTTTTCCGCTTGTCTTCCTGA··GCCGACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn (DXE)2 CCTGAAATTCAGGGTTGACTCTGAAA··GAGGAAAGCGTAATATAC·GCCACCTCGCGACAGTGAGC
rrn E1     CTGCAATTTTTCTATTGCGGCCTGCG··GAGAACTCCCTATAATGCGCCTCCATCGACACGGCGGAT
rrn A1     TTTTAAATTTCCTCTTGTCAGGCCGG··AATAACTCCCTATAATGCGCCACCACTGACACGGAACAA
rrn A2     GCAAAAATAAATGCTTGACTCTGTAG··CGGGAAGGCGTATTATGC·ACACCCCGCGCCGCTGAGAA
λ PR       TAACACCGTGCGTGTTGACTATTTTA·CCTCTGGCGGTGATAATGG··TTGCATGTACTAAGGAGGT
λ PL       TATCTCTGGCGGTGTTGACATAAATA·CCACTGGCGGTGATACTGA··GCACATCAGCAGGACGCAC
T7 A3      GTGAAACAAAACGGTTGACAACATGA·AGTAAACACGGTACGATGT·ACCACATGAAACGACAGTGA
T7 A1      TATCAAAAAGAGTATTGACTTAAAGT·CTAACCTATAGGATACTTA·CAGCCATCGAGAGGGACACG
T7 A2      ACGAAAAACAGGTATTGACAACATGAAGTAACATGCAGTAAGATAC·AAATCGCTAGGTAACACTAG
fd VIII    GATACAAATCTCCGTTGTACTTTGTT··TCGCGCTTGGTATAATCG·CTGGGGGTCAAAGATGAGTG
                         -35                       -10       +1
```

-20 -10 -1

| C | A | T | G | C | A | G | T | A | A | G | A | T | A | C | A | A | A | T | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

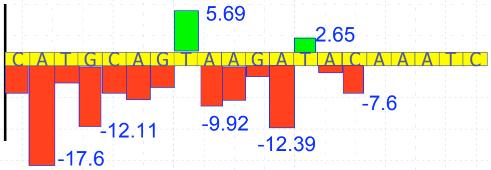
CATGCA Score=-4.88

ATGCAG Score=-17.6

TGCAGT Score=-2.56

........

TAAGAT Score=5.69



Score=-3.32+2.04+1.72-1-1-3.32=-4.88

|   | C | A | T | G | C | A |
|---|---|---|---|---|---|---|
| A | -3.32 | 2.04 | -1 | 1.32 | 1.89 | -3.32 |
| C | -3.32 | -3.32 | -1 | -3.32 | -1 | -3.32 |
| G | -0.15 | -3.32 | -3.32 | -1 | -3.32 | -3.32 |
| T | 1.72 | -3.32 | 1.72 | -0.15 | -3.32 | 2.04 |

|   | T | A | A | G | A | T |
|---|---|---|---|---|---|---|
| A | -3.32 | 2.04 | -1 | 1.32 | 1.89 | -3.32 |
| C | -3.32 | -3.32 | -1 | -3.32 | -1 | -3.32 |
| G | -0.15 | -3.32 | -3.32 | -1 | -3.32 | -3.32 |
| T | 1.72 | -3.32 | 1.72 | -0.15 | -3.32 | 2.04 |

Score=1.72+2.04 -1-1+1.89+2.04=5.69

20

# The Motif Finding Problem

- Motif finding algorithm:
  - Input: a set of sequences $X^1,\ldots X^n$ ; a motif length k
  - Output: PSSM motif M of length k, and its positions in the sequences

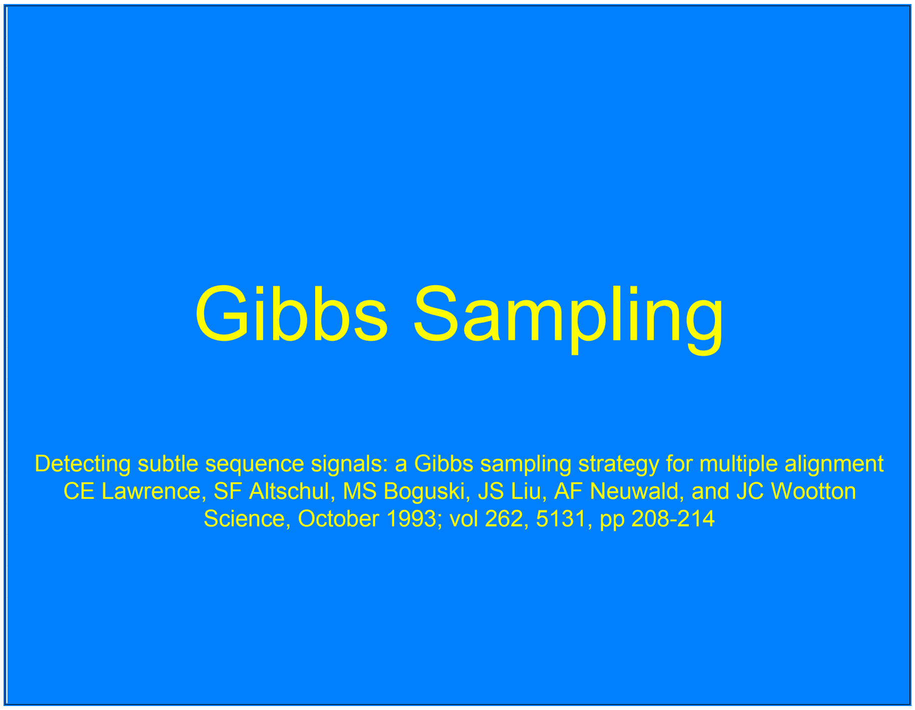
```
1: actcgtcggggcgtacgtacgtaacgtacgtaCGGACAACTGTTGACCG
2: cggagcactgttgagcgacaagtaCGGAGCACTGTTGAGCGgtacgtac
3: ccccgtaggCGGCGCACTCTCGCCCGggcgtacgtacgtaacgtacgta
4: agggcgcgtacgctaccgtcgacgtcgCGCGCCGCACTGCTCCGacgct
```

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | $\frac{3}{4}$ | 0 | $\frac{1}{4}$ | $\frac{1}{2}$ | 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ | 0 | 0 | 0 |
| C | $\frac{4}{4}$ | 0 | 0 | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ | 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ | 0 | $\frac{1}{4}$ | $\frac{3}{4}$ | $\frac{4}{4}$ | 0 |
| G | 0 | $\frac{4}{4}$ | $\frac{4}{4}$ | 0 | $\frac{3}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | 0 | 0 | $\frac{3}{4}$ | 0 | 0 | $\frac{3}{4}$ | 0 | $\frac{1}{4}$ | 0 | $\frac{4}{4}$ |
| T | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\frac{1}{2}$ | $\frac{3}{4}$ | 0 | $\frac{3}{4}$ | $\frac{1}{2}$ | 0 | $\frac{1}{4}$ | 0 | 0 | 0 |

- If we knew the positions the problem would be easy
  - Align the sequences and extract PSSM
- If we knew the PSSM, the problem would be easy
  - Align the PSSM with the sequences to find positions
- The problem is that both the motif PSSM and positions are unknown

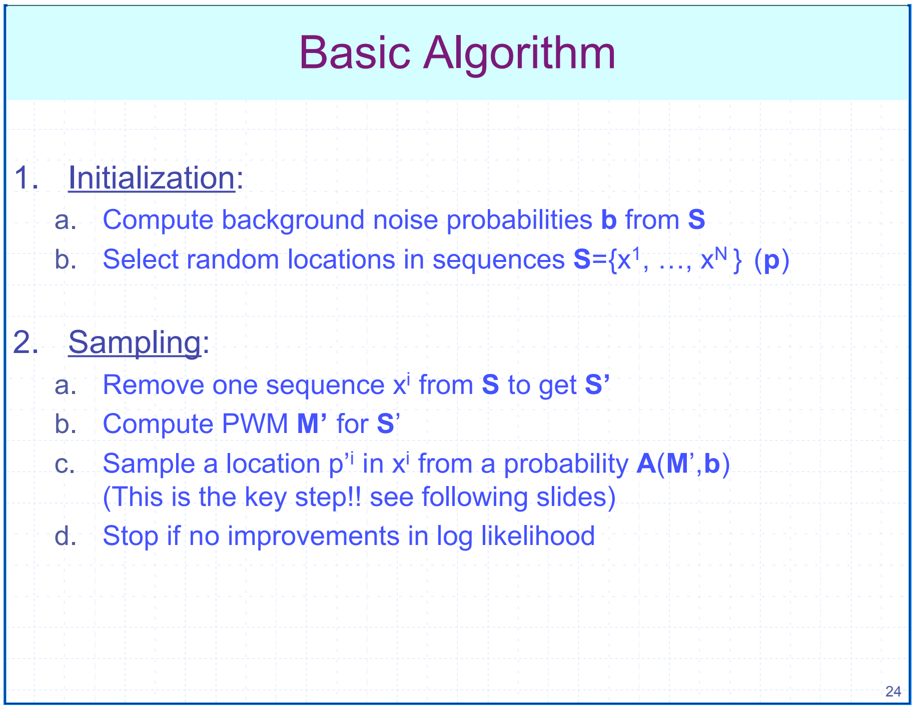
# Finding PSSM Motifs

- Brute force search:
  - Consider all k-mers; compute PSSM M for each subset of k-mers
  - Find M which maximizes the expected relative entropy $D(M||b)=\Sigma_{position-i}D(m^i||b)$
- Impractical complexity
- Recast the problem:
  - Estimate the parameters **M** (PSSM), **p** (positions) and **b** (background probability) that best explain the sample sequence **S**.
  - Maximize the log likelihood L(**M,p,b**|**S**)
- Strategy: iterate the following two steps
  - Estimate **p** from (**M,b**) and **S** (**M,b**) ➔**p**
  - Estimate (**M,b**) from **p** and **S** **p** ➔(**M,b**)
- This yields a family of solutions, depending on the estimation technique
  - Gibbs sampling, EM….



# Gibbs Sampling

Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment
CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton
Science, October 1993; vol 262, 5131, pp 208-214



## Basic Algorithm

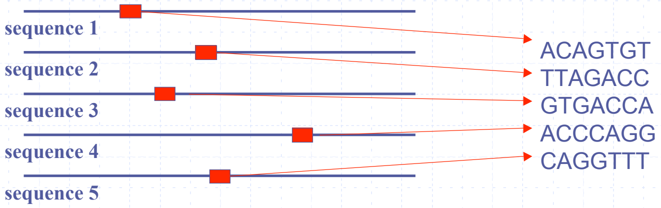
1. Initialization:
   a. Compute background noise probabilities **b** from **S**
   b. Select random locations in sequences **S**=$\{x^1, \ldots, x^N\}$ (**p**)

2. Sampling:
   a. Remove one sequence $x^i$ from **S** to get **S’**
   b. Compute PWM **M’** for **S**’
   c. Sample a location $p'^i$ in $x^i$ from a probability **A**(**M**’,**b**)
   (This is the key step!! see following slides)
   d. Stop if no improvements in log likelihood

24

# Initialization

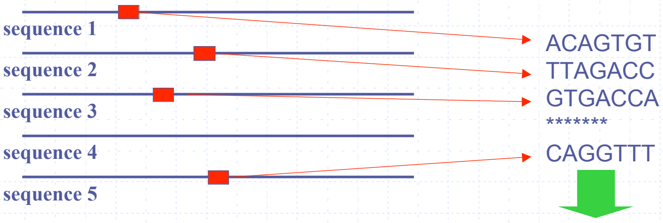
- Select random locations $a_1, \dots, a_N$ in $x^1, \dots, x^N$



ACAGTGT
TTAGACC
GTGACCA
ACCCAGG
CAGGTTT

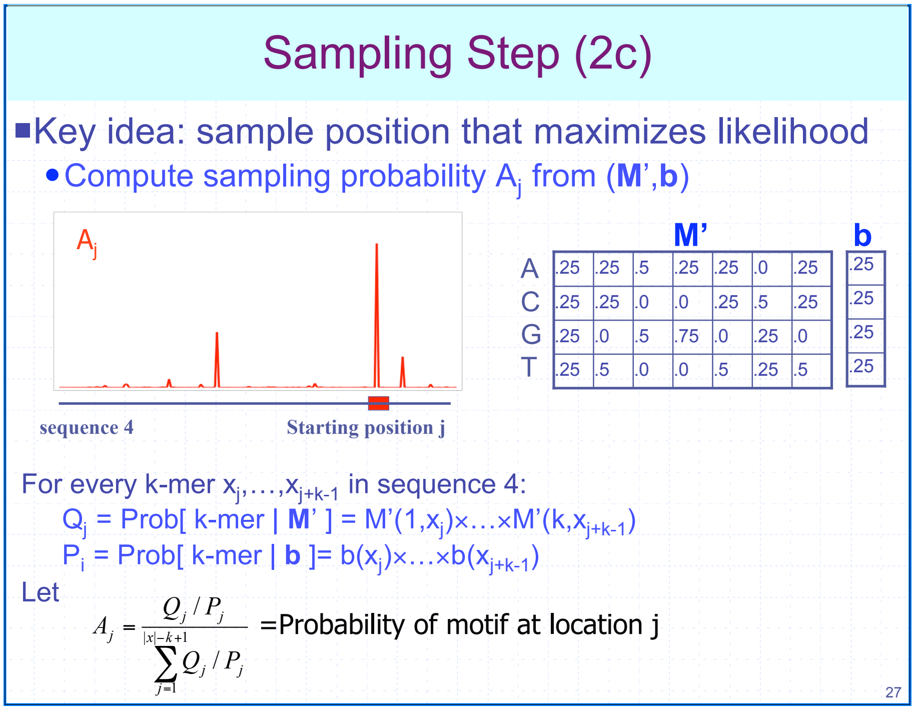
# Iteration: Remove A Sequence (2a,b)

- Select a sequence $x = x^i$
- Remove $x^i$ and compute PWM



ACAGTGT
TTAGACC
GTGACCA
*******
CAGGTTT

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | .25 | .25 | .5 | .25 | .25 | .0 | .25 |
| C | .25 | .25 | .0 | .0 | .25 | .5 | .25 |
| G | .25 | .0 | .5 | .75 | .0 | .25 | .0 |
| T | .25 | .5 | .0 | .0 | .5 | .25 | .5 |



## Sampling Step (2c)

- Key idea: sample position that maximizes likelihood
  - Compute sampling probability $A_j$ from (**M’**,**b**)

|   | **M’** |  |  |  |  |  |  | **b** |
|---|---|---|---|---|---|---|---|---|
| A | .25 | .25 | .5 | .25 | .25 | .0 | .25 | .25 |
| C | .25 | .25 | .0 | .0 | .25 | .5 | .25 | .25 |
| G | .25 | .0 | .5 | .75 | .0 | .25 | .0 | .25 |
| T | .25 | .5 | .0 | .0 | .5 | .25 | .5 | .25 |

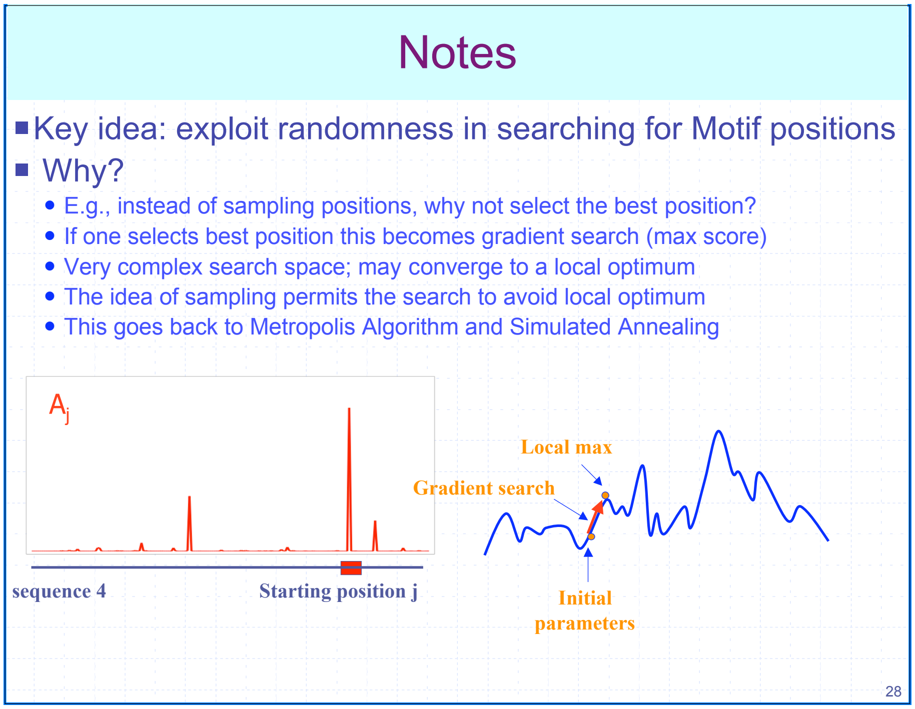
For every k-mer $x_j,\ldots,x_{j+k-1}$ in sequence 4:

$Q_j$ = Prob[ k-mer | **M’** ] = $M'(1,x_j)\times\ldots\times M'(k,x_{j+k-1})$

$P_i$ = Prob[ k-mer | **b** ]= $b(x_j)\times\ldots\times b(x_{j+k-1})$

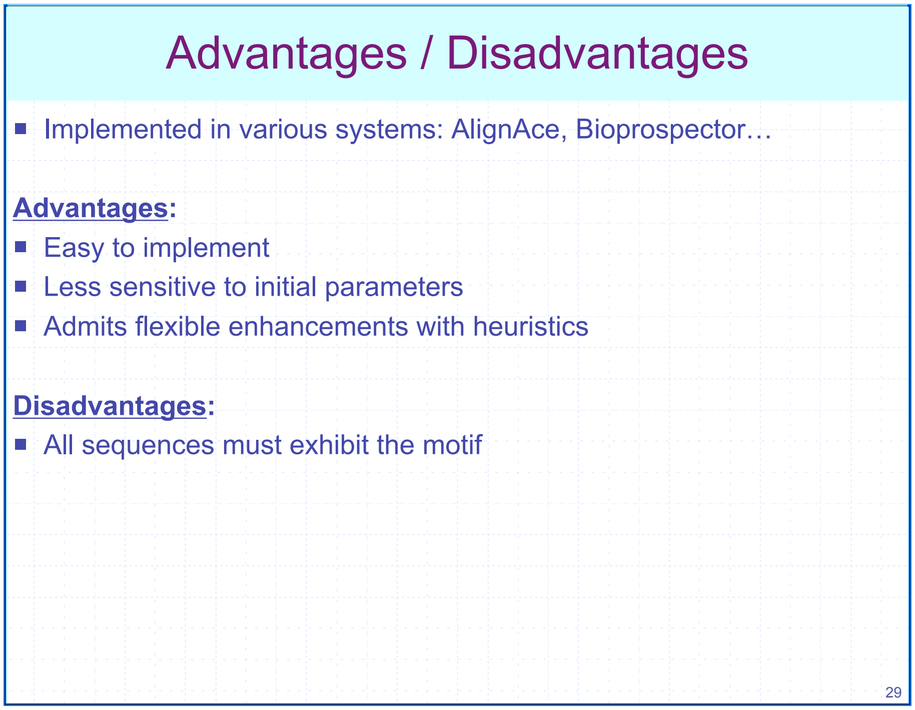
Let

$$A_j = \frac{Q_j / P_j}{\sum_{j=1}^{|x|-k+1} Q_j / P_j}$$ =Probability of motif at location j



## Notes

- Key idea: exploit randomness in searching for Motif positions
- Why?
  - E.g., instead of sampling positions, why not select the best position?
  - If one selects best position this becomes gradient search (max score)
  - Very complex search space; may converge to a local optimum
  - The idea of sampling permits the search to avoid local optimum
  - This goes back to Metropolis Algorithm and Simulated Annealing



## Advantages / Disadvantages

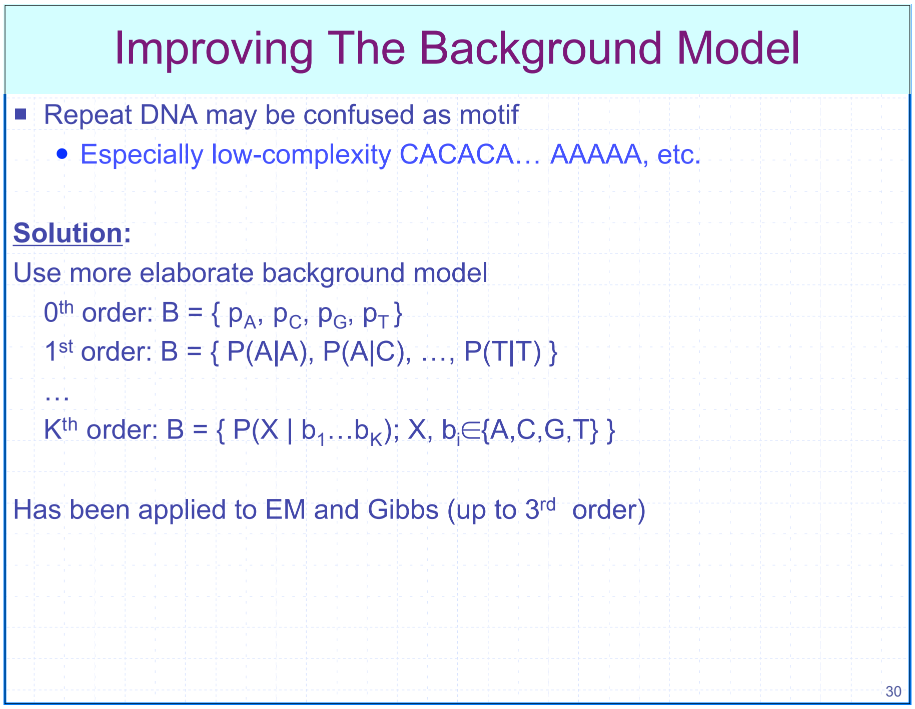
- Implemented in various systems: AlignAce, Bioprospector…

**Advantages:**

- Easy to implement
- Less sensitive to initial parameters
- Admits flexible enhancements with heuristics

**Disadvantages:**

- All sequences must exhibit the motif



## Improving The Background Model

- Repeat DNA may be confused as motif
  - Especially low-complexity CACACA… AAAAA, etc.

**Solution:**

Use more elaborate background model

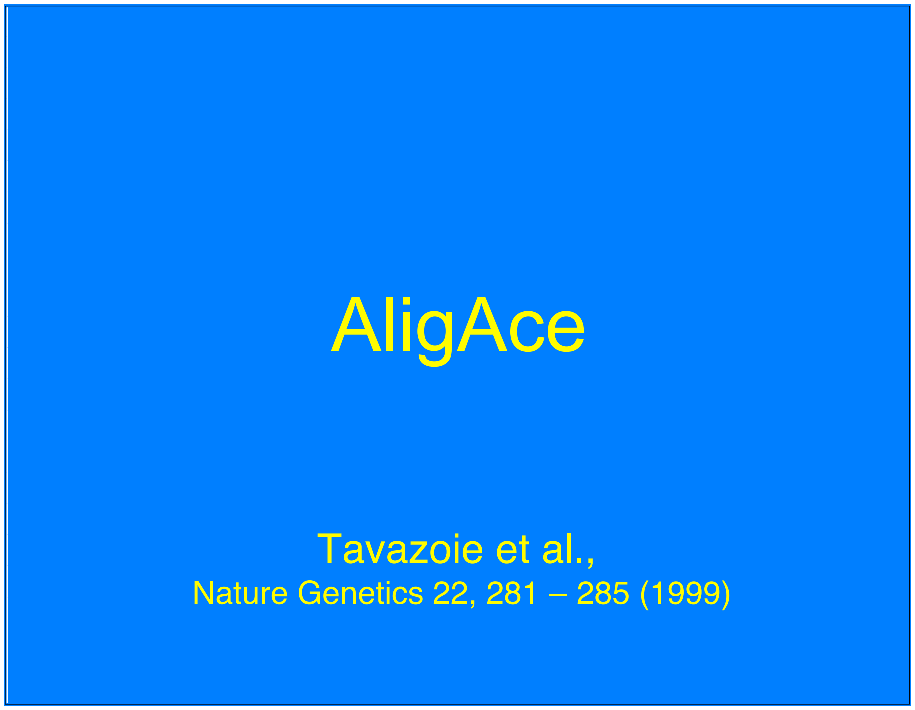
0th order: $B = \{ p_A, p_C, p_G, p_T \}$

1st order: $B = \{ P(A|A), P(A|C), \ldots, P(T|T) \}$

…

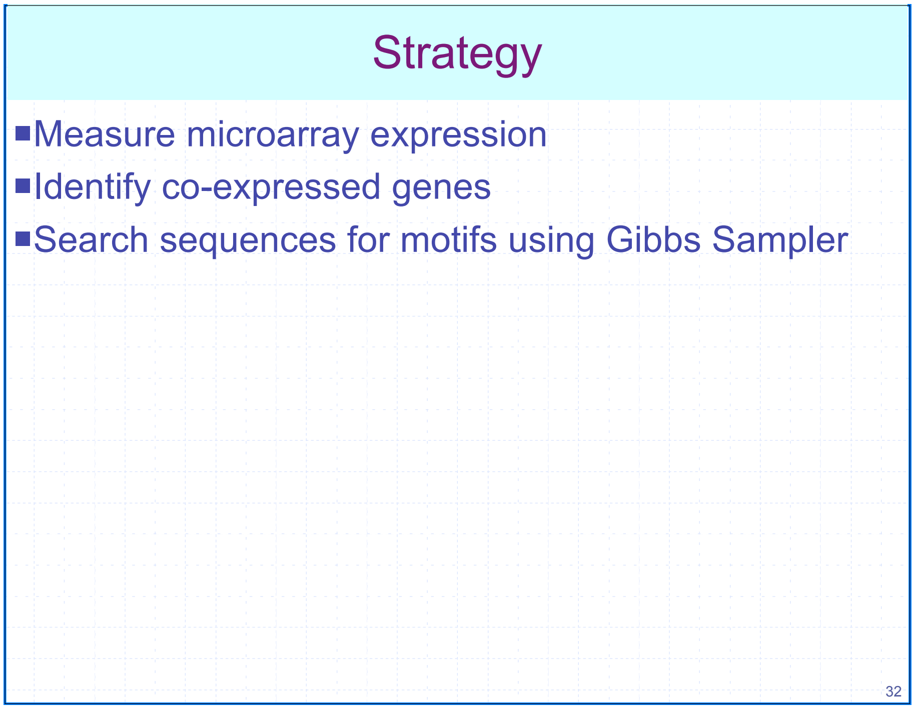
Kth order: $B = \{ P(X \mid b_1 \ldots b_K);\ X, b_i \in \{A,C,G,T\} \}$

Has been applied to EM and Gibbs (up to 3rd order)



# AligAce

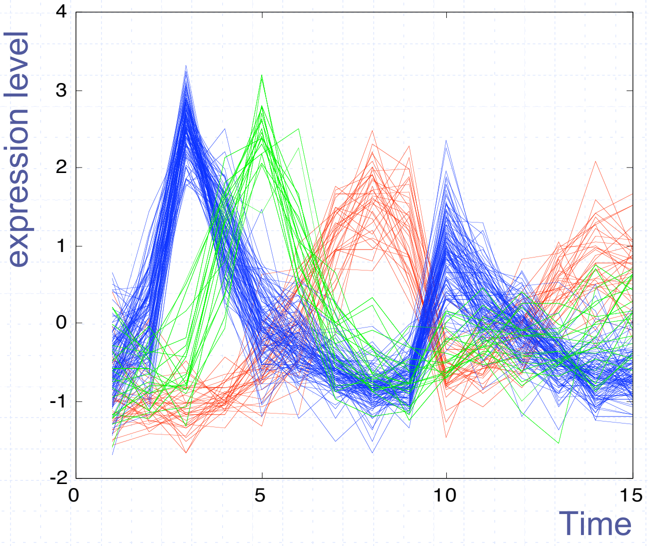
Tavazoie et al.,
Nature Genetics 22, 281 – 285 (1999)



## Strategy

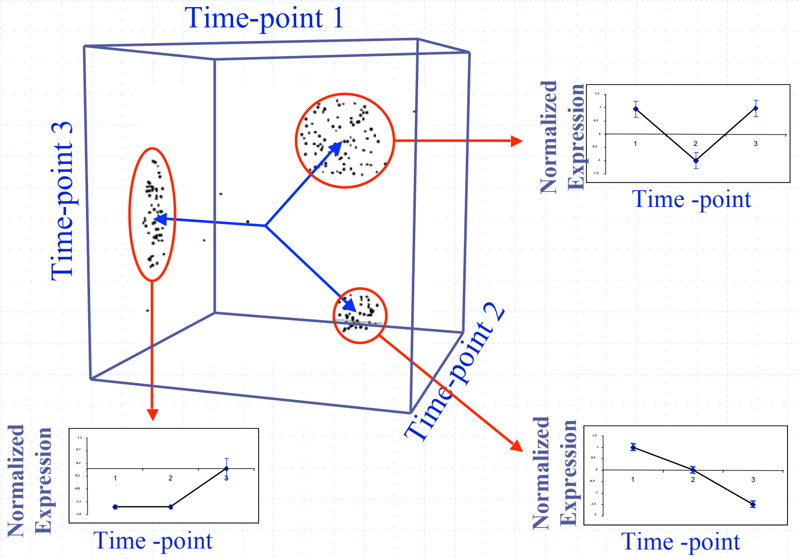
- Measure microarray expression
- Identify co-expressed genes
- Search sequences for motifs using Gibbs Sampler

# Correlation of Expression Profiles



# AlignAce

Tavazoie et al., Nature Genetics 22, 281 – 285 (1999)



# Search for Motifs in Promoter Regions

- The problem: finding motifs
- Use Gibbs sampling

5′ - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...*HIS7*

5′ - ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...*ARO4*

5′ - CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*

5′ - TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*

5′ - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA ...*ARO1*

5′ - ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...*HOM2*

5′ - GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

300-600 bp of upstream sequence per gene are searched in *Saccharomyces cerevisiae*.

Source: G.M. Church

# Motif Found by AlignACE

5′ - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAG**AAAAGAGTCA**GACATCGAAACATACAT ...*HIS7*

5′ - ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCG**AAATGACTCA**ACG ...*ARO4*

5′ - CACATCCAACGAATCACCTCACCGTTATCG**TGACTCACTT**TCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*

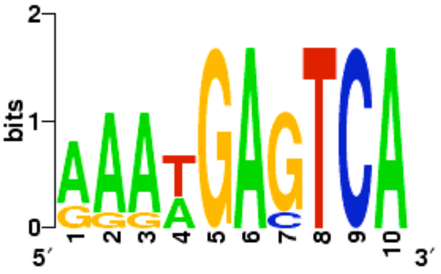
5′ - TGCGAAC**AAAAGAGTCA**TTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*

5′ - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATA**TGACTCATCC**CGAACATGAAA ...*ARO1*

5′ - ATTGAT**TGACTCATTT**TCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA ...*HOM2*

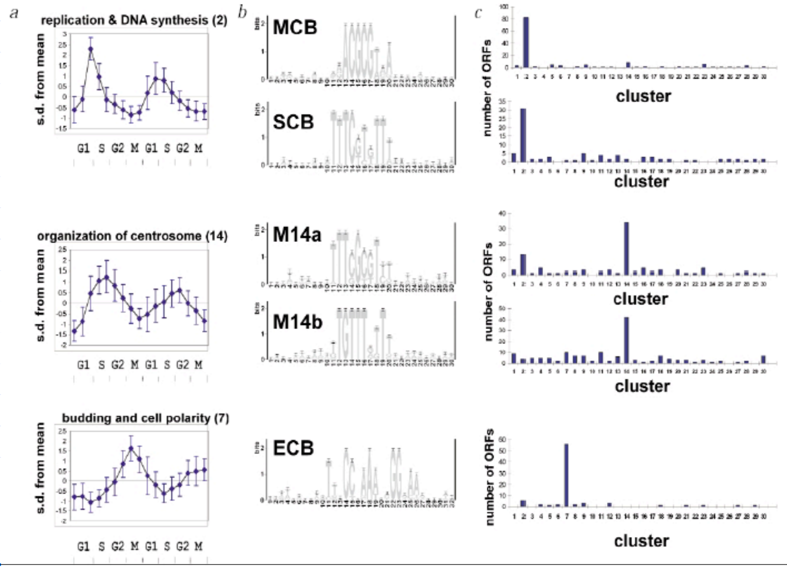
5′ - GGCGCCACAGTCCGCGTTTGGTTATCCGGC**TGACTCATTCTGACTCTTTT**TTGGAAAGTGTGGCATGTGCTTCACACA ...*PRO3*

```
AAAAGAGTCA
AAATGACTCA
AAGTGAGTCA
AAAAGAGTCA
GGATGAGTCA
AAATGAGTCA
GAATGAGTCA
AAAAGAGTCA
**********
```



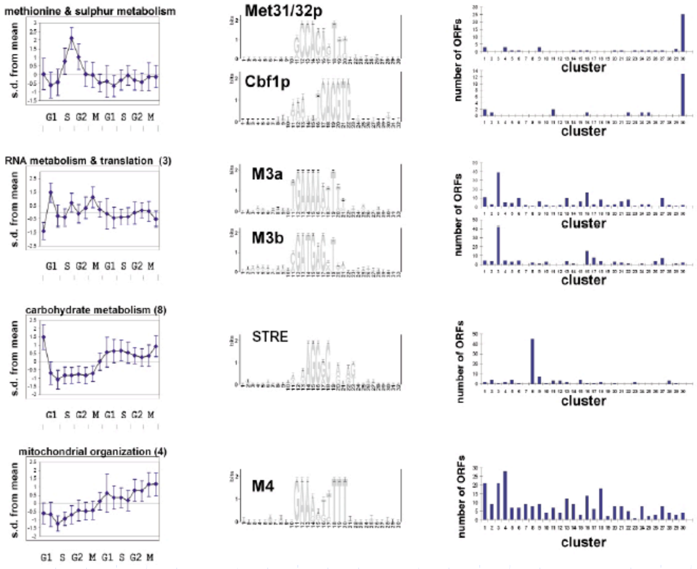
# Yeast Motifs: Periodic Clusters

Tavazoie et al., Nature Genetics 22, 281 – 285 (1999)



37

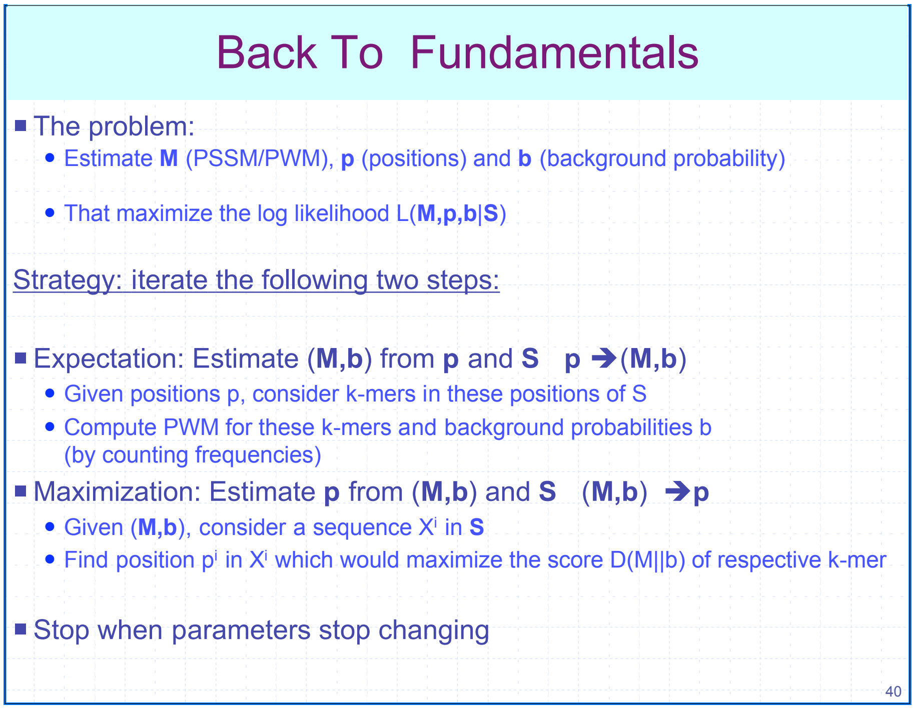
# Motifs in Non-periodic Clusters



38

# Expectation Maximization

MEME: Baily & Elkan 1995

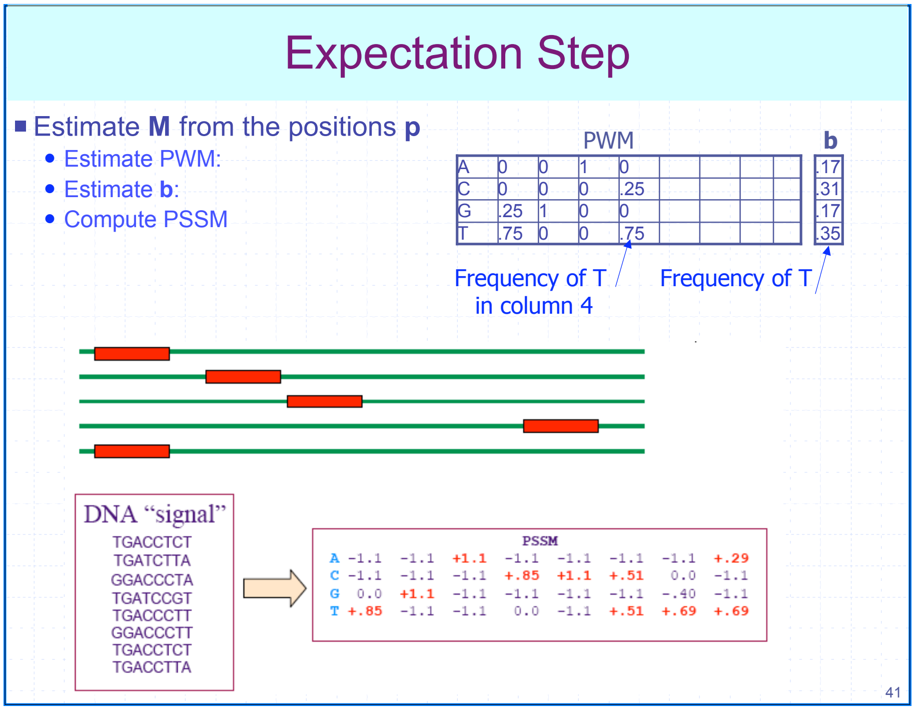


## Back To  Fundamentals

- The problem:
  - Estimate **M** (PSSM/PWM), **p** (positions) and **b** (background probability)
  - That maximize the log likelihood L(**M**,**p**,**b**|**S**)

Strategy: iterate the following two steps:

- Expectation: Estimate (**M**,**b**) from **p** and **S**   **p** ➔(**M**,**b**)
  - Given positions p, consider k-mers in these positions of S
  - Compute PWM for these k-mers and background probabilities b (by counting frequencies)
- Maximization: Estimate **p** from (**M**,**b**) and **S**   (**M**,**b**) ➔**p**
  - Given (**M**,**b**), consider a sequence $X^i$ in **S**
  - Find position $p^i$ in $X^i$ which would maximize the score D(M||b) of respective k-mer
- Stop when parameters stop changing



## Expectation Step

- Estimate **M** from the positions **p**
  - Estimate PWM:
  - Estimate **b**:
  - Compute PSSM

PWM

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 1 | 0 | | | | |
| C | 0 | 0 | 0 | .25 | | | | |
| G | .25 | 1 | 0 | 0 | | | | |
| T | .75 | 0 | 0 | .75 | | | | |

**b**

| |
|---|
| .17 |
| .31 |
| .17 |
| .35 |

Frequency of T in column 4

Frequency of T

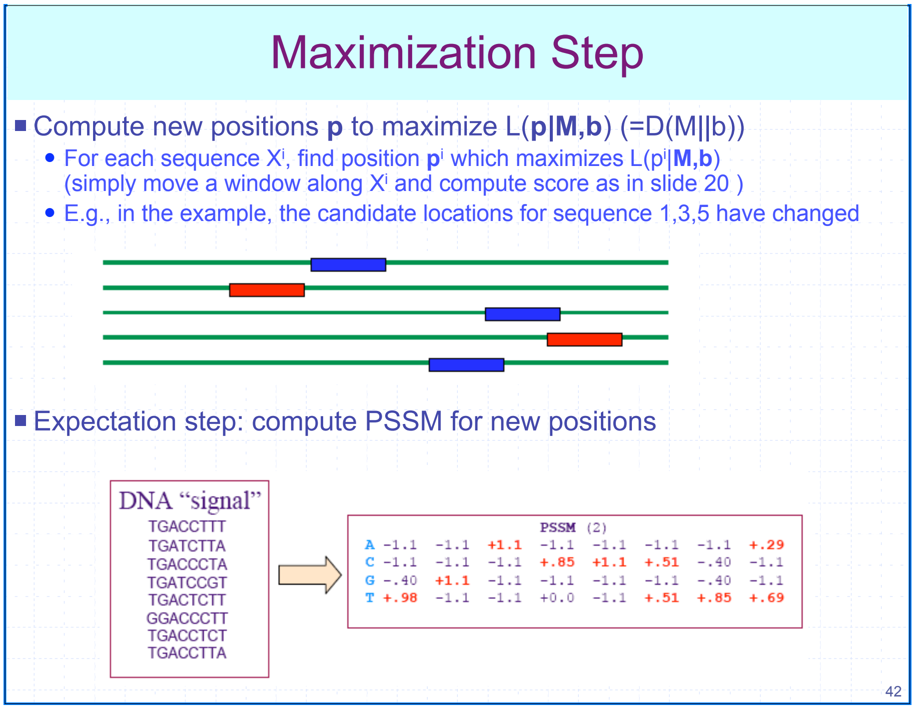
DNA "signal"

TGACCTCT
TGATCTTA
GGACCCTA
TGATCCGT
TGACCCTT
GGACCCTT
TGACCTCT
TGACCTTA

PSSM

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | -1.1 | -1.1 | +1.1 | -1.1 | -1.1 | -1.1 | -1.1 | +.29 |
| C | -1.1 | -1.1 | -1.1 | +.85 | +1.1 | +.51 | 0.0 | -1.1 |
| G | 0.0 | +1.1 | -1.1 | -1.1 | -1.1 | -1.1 | -.40 | -1.1 |
| T | +.85 | -1.1 | -1.1 | 0.0 | -1.1 | +.51 | +.69 | +.69 |

41



## Maximization Step

- Compute new positions **p** to maximize L(**p**|**M**,**b**) (=D(M||b))
  - For each sequence $X^i$, find position $\mathbf{p}^i$ which maximizes L($p^i$|**M**,**b**) (simply move a window along $X^i$ and compute score as in slide 20 )
  - E.g., in the example, the candidate locations for sequence 1,3,5 have changed
- Expectation step: compute PSSM for new positions

DNA "signal"

TGACCTTT
TGATCTTA
TGACCCTA
TGATCCGT
TGACTCTT
GGACCCTT
TGACCTCT
TGACCTTA

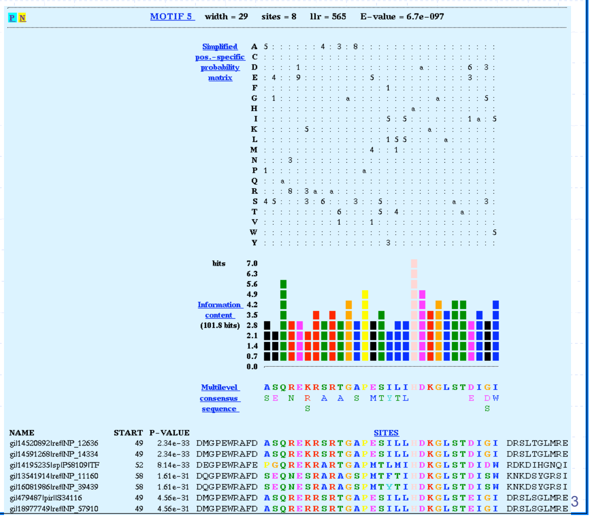
PSSM (2)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A | -1.1 | -1.1 | +1.1 | -1.1 | -1.1 | -1.1 | -1.1 | +.29 |
| C | -1.1 | -1.1 | -1.1 | +.85 | +1.1 | +.51 | -.40 | -1.1 |
| G | -.40 | +1.1 | -1.1 | -1.1 | -1.1 | -1.1 | -.40 | -1.1 |
| T | +.98 | -1.1 | -1.1 | +0.0 | -1.1 | +.51 | +.85 | +.69 |

42

# Example: Discovering Motif With MEME

- MEME an EM based motif finder
- Below is a sample output



```
P N              MOTIF 5   width = 29   sites = 8   llr = 565   E-value = 6.7e-097

Simplified       A  5 : : : : : : 4 : 3 : 8 : : : : : : : : : : : : : : : : :
pos.-specific    C  : : : : : : : : : : : : : : : : : : : : : : : : : : : : :
probability      D  : : : : 1 : : : : : : : : : : : : : : a : : : : : 6 : 3 :
matrix           E  : 4 : : 9 : : : : : : : : 5 : : : : : : : : : : : 3 : : :
                 F  : : : : : : : : : : : : : : : 1 : : : : : : : : : : : : :
                 G  : 1 : : : : : : : : a : : : : : : : : : : a : : : : : 5 :
                 H  : : : : : : : : : : : : : : : : : : a : : : : : : : : : :
                 I  : : : : : : : : : : : : : : : 5 : 5 : : : : : : : 1 a : 5
                 K  : : : : : 5 : : : : : : : : : : : : : : a : : : : : : : :
                 L  : : : : : : : : : : : : : : : 1 5 5 : : : : a : : : : : :
                 M  : : : : : : : : : : : : : 4 : : 1 : : : : : : : : : : : :
                 N  : : : 3 : : : : : : : : : : : : : : : : : : : : : : : : :
                 P  1 : : : : : : : : : : : a : : : : : : : : : : : : : : : :
                 Q  : : a : : : : : : : : : : : : : : : : : : : : : : : : : :
                 R  : : : 8 : 3 a : a : : : : : : : : : : : : : : : : : : : :
                 S  4 5 : : : 3 : 6 : : 3 : : 5 : : : : : : : : : a : : : 3 :
                 T  : : : : : : : : : 6 : : : : 5 : 4 : : : : : : : a : : : :
                 V  : : : : : : : : : 1 : : : 1 : : : : : : : : : : : : : : :
                 W  : : : : : : : : : : : : : : : : : : : : : : : : : : : : 5
                 Y  : : : : : : : : : : : : : : : 3 : : : : : : : : : : : : :

                 bits  7.0
                       6.3
                       5.6
                       4.9
Information            4.2
content                3.5
(101.8 bits)           2.8
                       2.1
                       1.4
                       0.7
                       0.0

Multilevel          A S Q R E K R S R T G A P E S I L I H D K G L S T D I G I
consensus           S E   N   R   A   A   S   M T Y T L               E   D W
sequence                      S                                           S

NAME                      START  P-VALUE                          SITES
gi|14520892|ref|NP_12636     49  2.34e-33  DMGPEWRAFD  A S Q R E K R S R T G A P E S I L L H D K G L S T D I G I  DRSLTGLMRE
gi|14591268|ref|NP_14334     49  2.34e-33  DMGPEWRAFD  A S Q R E K R S R T G A P E S I L L H D K G L S T D I G I  DRSLTGLMRE
gi|14195235|sp|P58109|TF     52  8.14e-33  DEGPEWRAFE  P G Q R E K R A R T G A P M T L M I H D K G L S T D I D W  RDKDIHGNQI
gi|13541914|ref|NP_11160     58  1.61e-31  DQGPEWRAFD  S E Q N E S R A R A G S P M T F T I H D K G L S T D I S W  KNKDSYGRSI
gi|16081986|ref|NP_39439     58  1.61e-31  DQGPEWRAFD  S E Q N E S R A R A G S P M T Y T I H D K G L S T D I S W  KNKDSYGRSI
gi|479487|pir||S34116        49  4.56e-31  DMGPEWRAFD  A S Q R E R R S R T G A P E S I L L H D K G L S T E I G I  DRSLSGLMRE
gi|18977749|ref|NP_57910     49  4.56e-31  DMGPEWRAFD  A S Q R E R R S R T G A P E S I L L H D K G L S T E I G I  DRSLSGLMRE
```

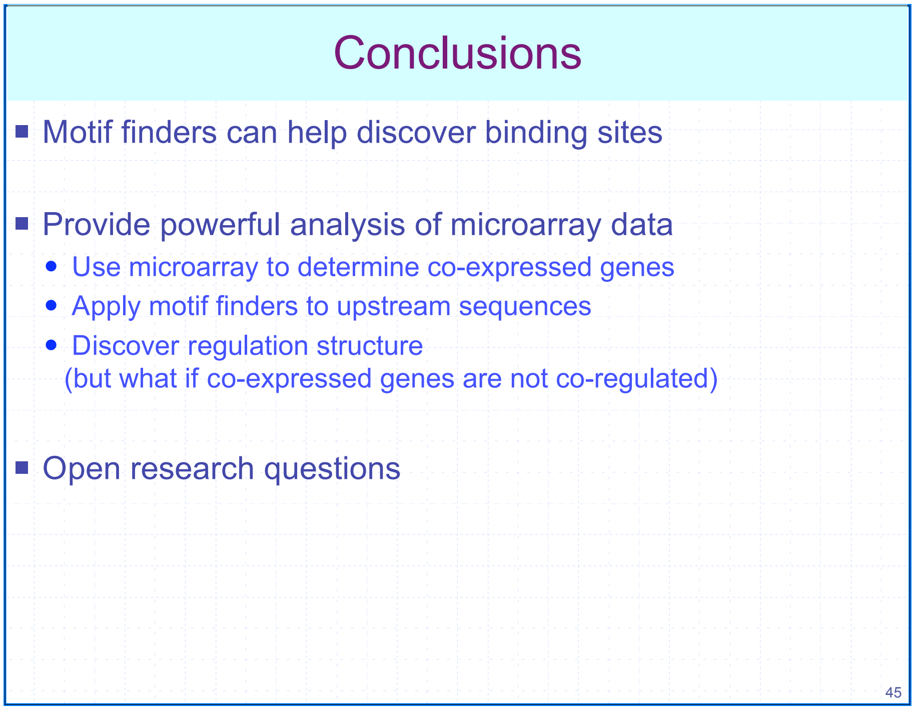
# Notes

- EM is essentially a gradient optimization search
  - May settle on a local maximum
  - Result may depend on initial conditions
- Artificial example: $X_1, \ldots, X_n$; (n = 200) k=6; suppose initial selection of k-mers
  - 99 words "AAAAAA"
  - 99 words "CCCCCC"
  - 2 words "ACACAC"

| | | | | | | |
|---|---|---|---|---|---|---|
| A | .51 | .49 | .51 | .49 | .51 | .49 |
| C | .49 | .51 | .49 | .51 | .49 | .51 |

Which motif will be discovered?

- Challenges
  - Handling 0, or 2+ binding sites
  - Gapped binding sites



# Conclusions

- Motif finders can help discover binding sites
- Provide powerful analysis of microarray data
  - Use microarray to determine co-expressed genes
  - Apply motif finders to upstream sequences
  - Discover regulation structure
    (but what if co-expressed genes are not co-regulated)
- Open research questions