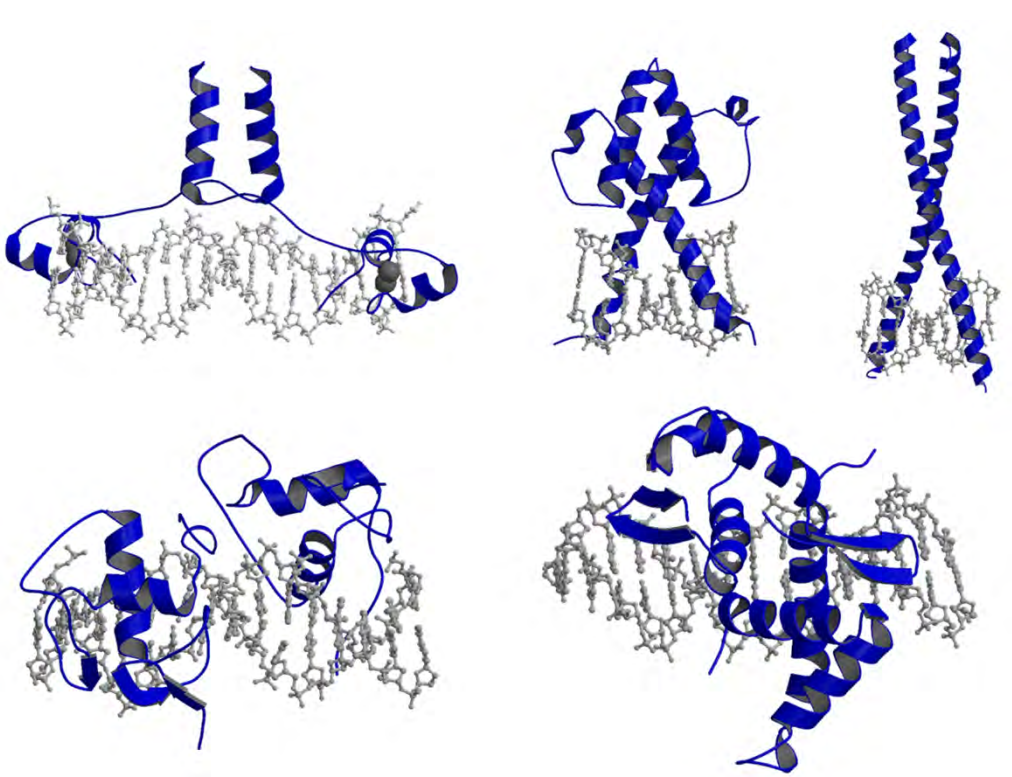
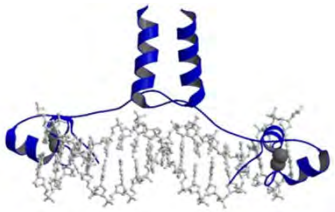
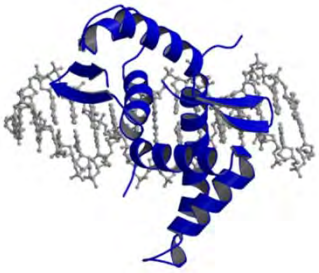
Transcription Factor Sequence Motifs



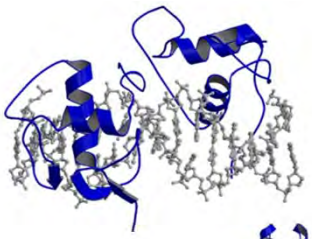


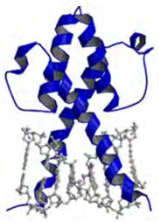
Transcription Factor Sequence Motifs

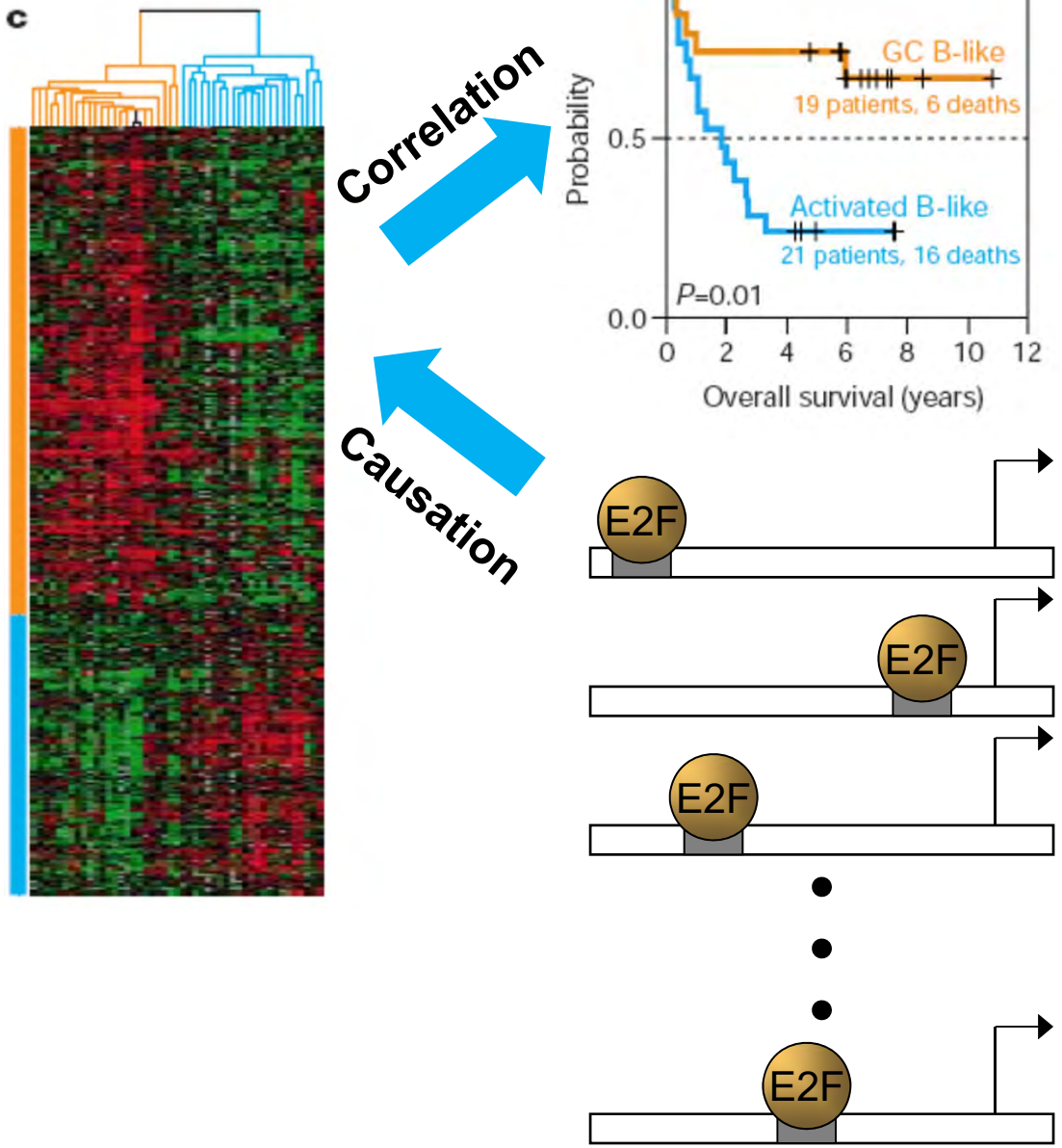


Outline

- Why look for TF binding sites?
- What is a motif?
- How do I compute a motif from bound sites?
- Is a motif enriched?
- How do you find a motif from unaligned seqs?



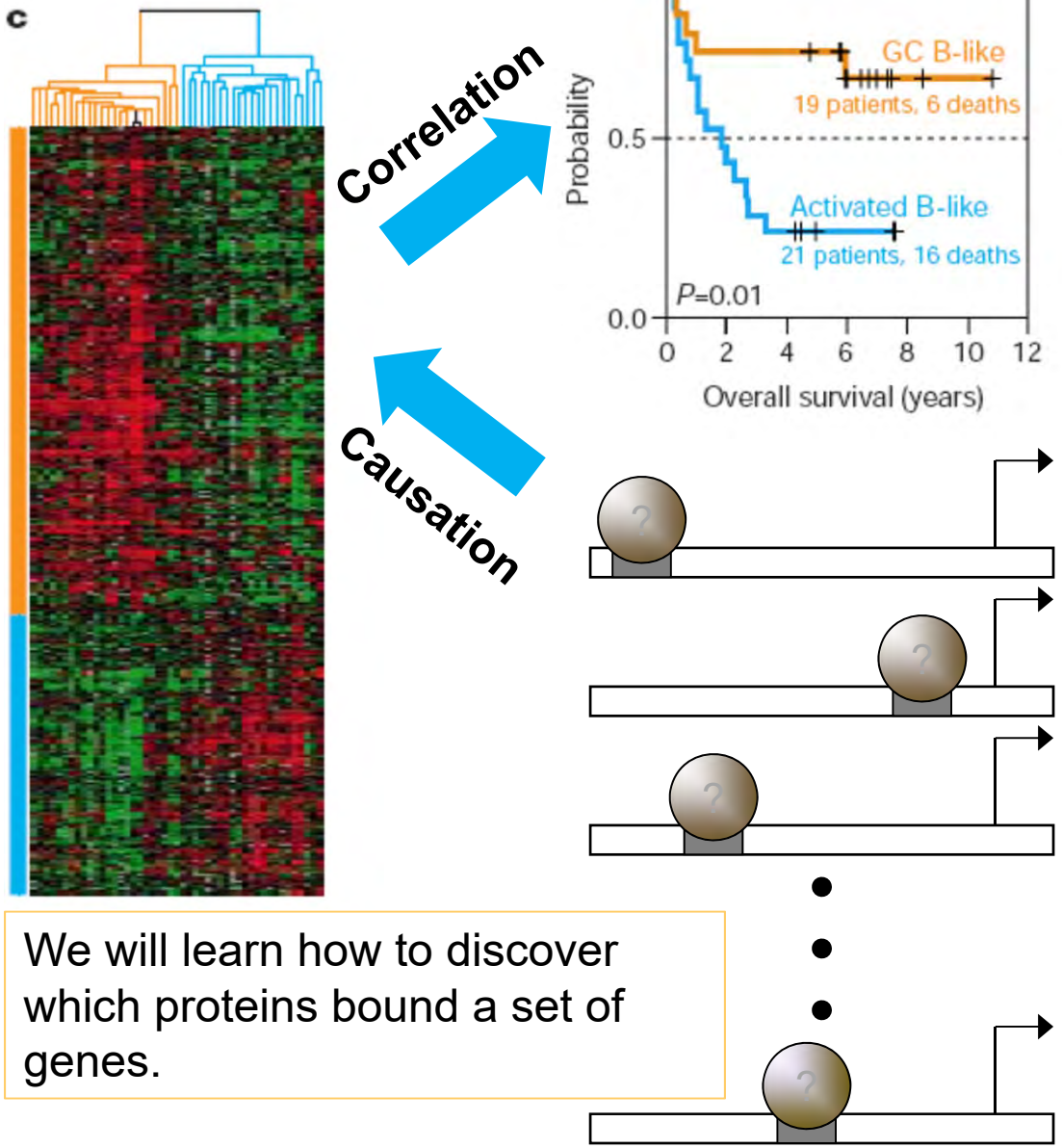




Why do these genes predict outcome?

The are regulated by the E2F family of cell-cycle regulatory proteins.

Tumors that divide more rapidly express more of these proteins and are more dangerous



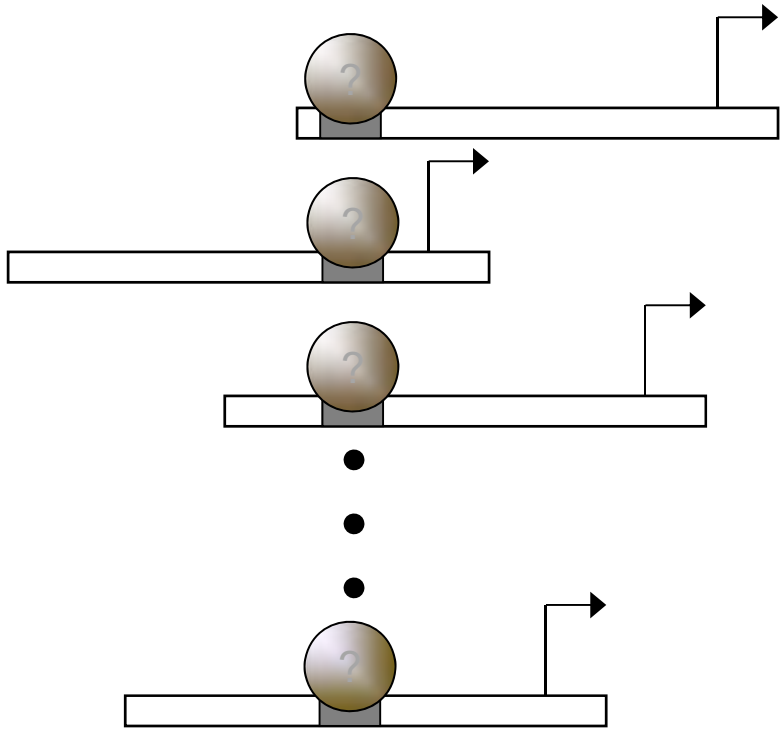
We will learn how to discover which proteins bound a set of genes.

Why do these genes predict outcome?

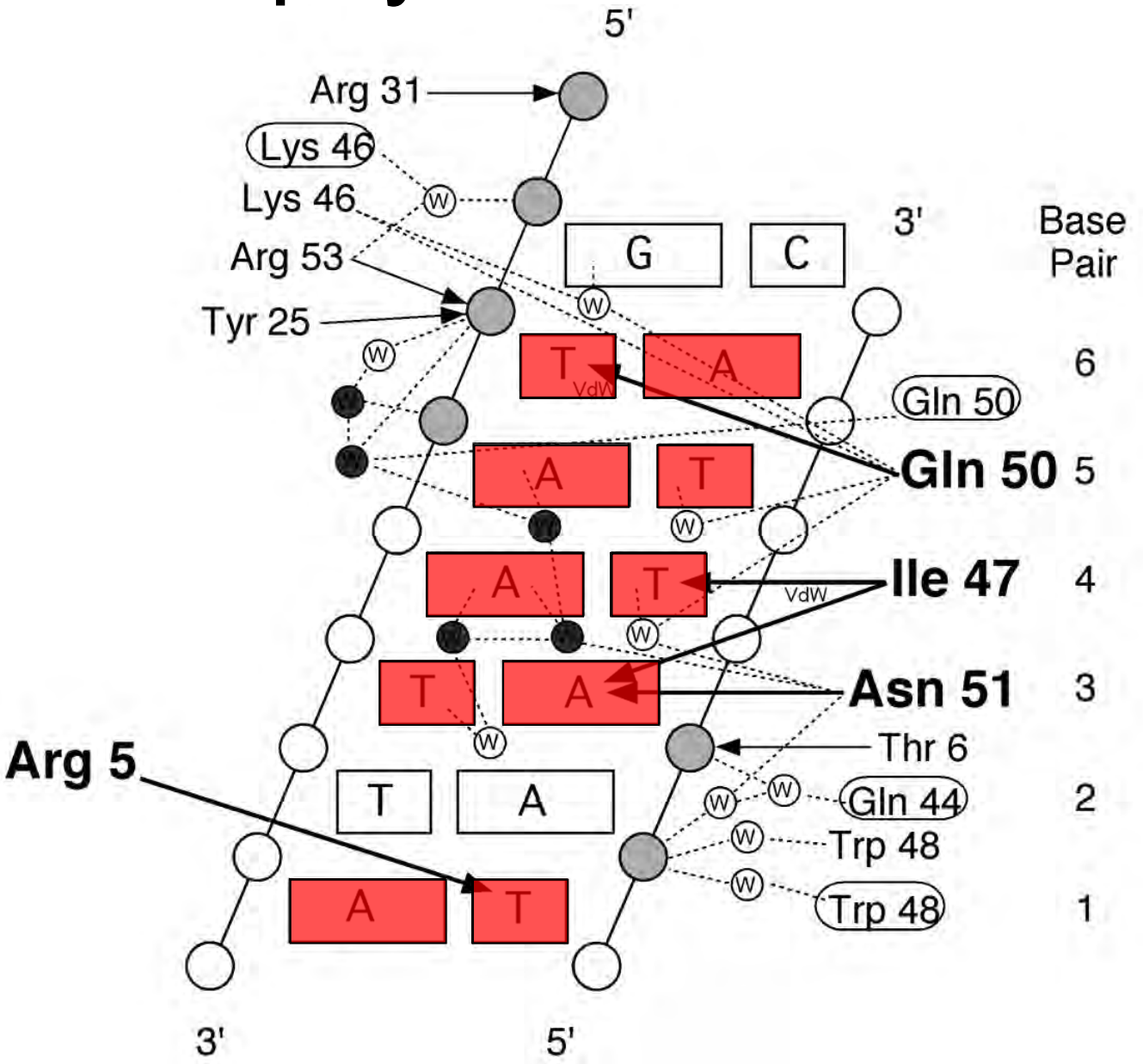
The are regulated by the E2F family of cell-cycle regulatory proteins.

Tumors that divide more rapidly express more of these proteins and are more dangerous

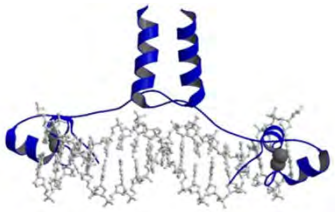
What would I expect if I lined up the sequences of binding sites?



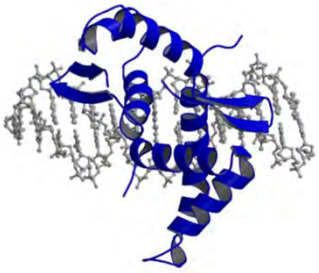
Biophysics determines binding



Some base pairs are more critical than others

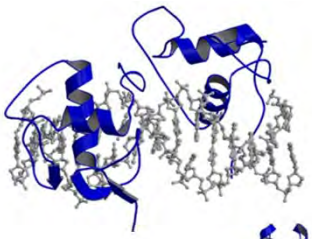


Transcription Factor Sequence Motifs



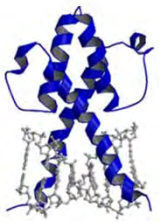
Outline

Why look for TF binding sites?



What is a motif?

How do I compute a motif from bound sites?



Is a motif enriched?

How do you find a motif from unaligned seqs?

Motifs

- Measure the *odds* that a sequence was *generated* by the motif model after compared to the *background* frequencies of each nucleotide

Is a region a valid binding site?

Match?

ACGTAGATCGATCCCCTGATCAAAATCGTGTTGAGCGCGCGTAATATCGCTAGCTAGCAAATTCCGATA

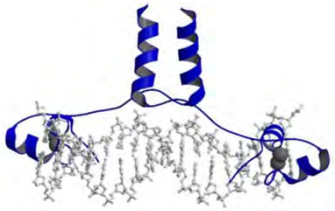
- Steps:
  1. Define a mathematical model for matching sequences
  2. Define a model for sequences that don't match
  3. Quantitatively compare the two hypotheses

$$odds = \frac{Probability_{binding\_site}}{Probability_{not\_a\_binding\_site}}$$

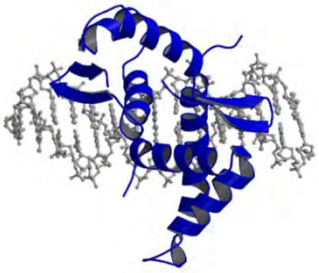
Define motif model → Define background model → Compare the models

Sequence Logos

- A letter's height indicates the information it contains (i.e. reduction in uncertainty)
- The top letter at each position can be read to obtain the *consensus sequence*

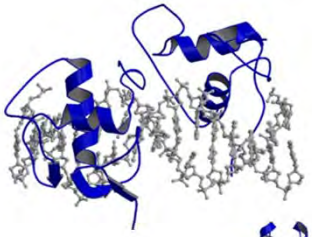


Transcription Factor Sequence Motifs



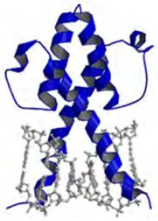
Outline

Why look for TF binding sites?



What is a motif?

How do I compute a motif from bound sites?



Is a motif enriched?

How do you find a motif from unaligned seqs?

If I had found these sites using ChIP-Seq, how would I describe the specificity?

TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA

A

If I had found these sites using ChIP-Seq, how would I describe the specificity?

```
TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA
```

Position Frequency Matrix (PFM)

A:	0	0	10	0	2	3	9
C:	0	0	0	10	0	7	1
G:	0	9	0	0	0	0	0
T:	10	1	0	0	8	0	0

If I had found these sites using ChIP-Seq, how would I describe the specificity?

```
TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA
```

Position Frequency Matrix (PFM)

A:	0	0	10	0	2	3	9
C:	0	0	0	10	0	7	1
G:	0	9	0	0	0	0	0
T:	10	1	0	0	8	0	0

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

Define motif model → Define background model → Compare the models

Is a region a valid binding site?

- Steps:
  1. Define a mathematical model for matching sequences

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b, i)$$

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

Is a region a valid binding site?

- Steps:
  1. Define a mathematical model for matching sequences

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b,i)$$

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

  2. Define a model for sequences that don't match: $P_{background} = 0.25$

Is the sequence more probably a motif or a random genomic region?

- Steps:
  3. Quantitatively compare the two hypotheses

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b,i)$$

$$Background\_prob = \prod_{i=1}^{w} p_{\text{background}}(b)$$

Odds ratio

$$\frac{Model\_prob}{Background\_prob} = \prod_{i=1}^{w} \frac{p_{\text{model}}(b,i)}{p_{\text{background}}(b)} = \prod_{i=1}^{w} odds(b,i)$$

B

Define motif model | Define background model | Compare the models

$$\frac{Model\_prob}{Background\_prob} = \prod_{i=1}^{w} \frac{p_{model}(b,i)}{p_{background}(b)} = \prod_{i=1}^{w} odds(b,i)$$

Position Probability Matrix (PPM) = *Model_prob*

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

Assume that all bases are equally likely in the rest of the genome

$P_{background} = 0.25$

Odds Ratio

A:	0	0	4.000	0	0.800	1.200	3.600
C:	0	0	0	4.000	0	2.800	0.400
G:	0	3.600	0	0	0	0	0
T:	4.000	0.400	0	0	3.200	0	0

Define motif model → Define background model → Compare the models

Likelihood = Prob(sequence | model)

What is the likelihood of TGACTCA?
What is the likelihood of AGACTCA?

Odds Ratio

A:	0	0	4.000	0	0.800	1.200	3.600
C:	0	0	0	4.000	0	2.800	0.400
G:	0	3.600	0	0	0	0	0
T:	4.000	0.400	0	0	3.200	0	0

Laplace's rule takes into account rare events

$$F(base = b) = \frac{C(base = b)}{Total\_counts}$$

$$F(base = b) = \frac{C(base = b) + 1}{Total\_counts + 4}$$

Pseudocounts:

add a small frequency ρ to all bases

$$F(base = b) = \frac{C(base = b)}{Total\_counts}$$

$$F(base = b) = \frac{C(base = b) + 1}{Total\_counts + 4}$$

$$F(base = b) = \frac{C(base = b) + p}{Total\_counts + 4p}$$

C

In next slides p=0.25% of Total_Counts

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

Odds Ratio

A:	0	0	4.000	0	0.800	1.200	3.600
C:	0	0	0	4.000	0	2.800	0.400
G:	0	3.600	0	0	0	0	0
T:	4.000	0.400	0	0	3.200	0	0

With Pseudocounts

D

Position Probability Matrix (PPM)

A:	0.0025	0.0025	0.9926	0.0025	0.2005	0.2995	0.8936
C:	0.0025	0.0025	0.0025	0.9926	0.0025	0.6955	0.1015
G:	0.0025	0.8936	0.0025	0.0025	0.0025	0.0025	0.0025
T:	0.9926	0.1015	0.0025	0.0025	0.7946	0.0025	0.0025

Odds Ratio

A:	0.0099	0.0099	3.9703	0.0099	0.8020	1.1980	3.5743
C:	0.0099	0.0099	0.0099	3.9703	0.0099	2.7822	0.4059
G:	0.0099	3.5743	0.0099	0.0099	0.0099	0.0099	0.0099
T:	3.9703	0.4059	0.0099	0.0099	3.1782	0.0099	0.0099

Log-Odds

$$\frac{Model\_prob}{Background\_prob} = \prod_{i=1}^{w} \frac{p_{model}(b,i)}{p_{background}(b)} = \prod_{i=1}^{w} odds(b,i)$$

$$\log\left[\frac{Model\_prob}{Background\_prob}\right] = \log\left[\prod_{i=1}^{w} odds(b,i)\right] = \sum_{i=1}^{w} \log\left(odds(b,i)\right)$$

E

Define motif model | Define background model | Compare the models

Log-odds matrix is often called PWM position weight matrix or PSSM position-specific scoring matrix

$$log\left[\frac{P_{model}}{P_{background}}\right] = log[P_{model}] - log[P_{background}]$$

Log Odds

```
A: -6.658 | -6.658 |  1.989 | -6.658 | -0.318 |  0.261 |  1.838 |
C: -6.658 | -6.658 | -6.658 |  1.989 | -6.658 |  1.476 | -1.301 |
G: -6.658 |  1.838 | -6.658 | -6.658 | -6.658 | -6.658 | -6.658 |
T:  1.989 | -1.301 | -6.658 | -6.658 |  1.668 | -6.658 | -6.658
```

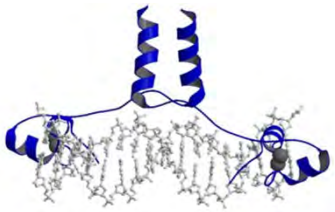
Define motif model → Define background model → Compare the models

Is a region a valid binding site?

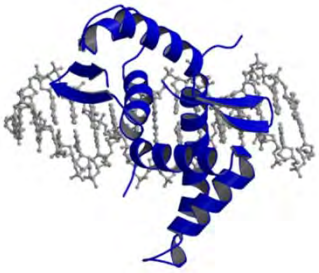
- Steps:
  1. Define a mathematical model for matching sequences
  2. Define a model for sequences that don't match
  3. Quantitatively compare the two hypotheses

ACGTAGATCGATCCCCTGATCAAAATCGTGTTGAGCGCGCGTAATATCGCTAGCTAGCAAATTCCGATA

Match?

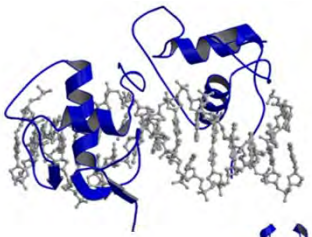


Transcription Factor Sequence Motifs



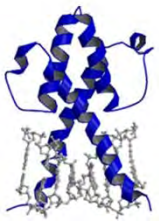
Outline

Why look for TF binding sites?



What is a motif?

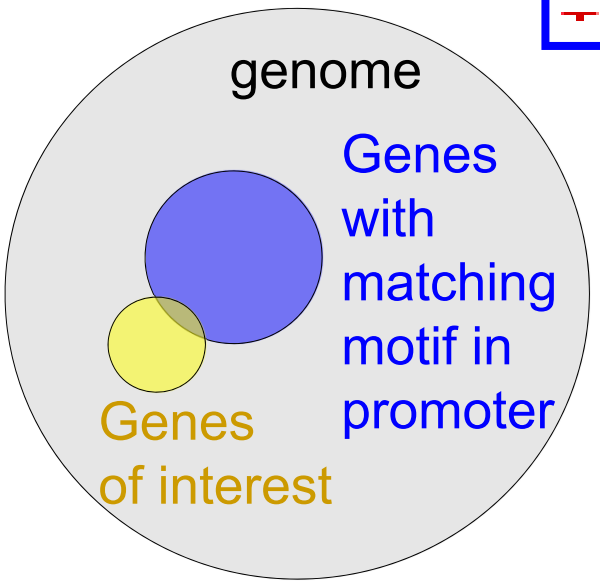
How do I compute a motif from bound sites?

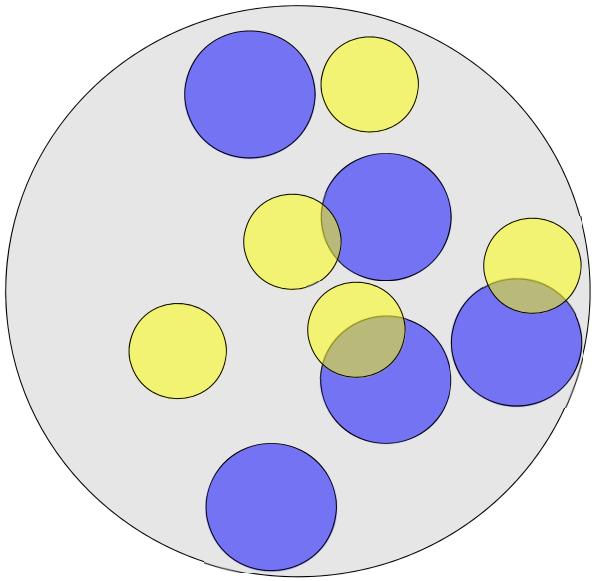


Is a motif enriched?

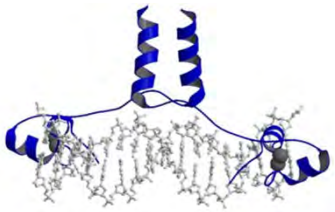
How do you find a motif from unaligned seqs?

Compute Significance With Hypergeometric Distribution

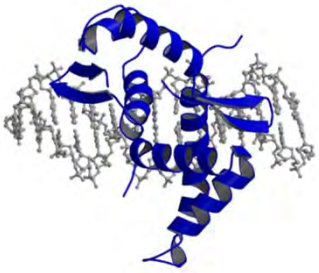




- **Rule-of-thumb**: *60% of the maximum-possible* LLR score is a reasonable threshold for determining a match to a *PWM motif*

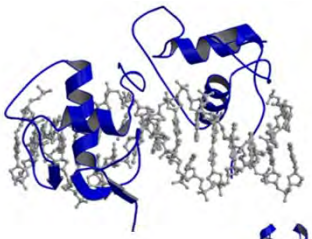


Transcription Factor Sequence Motifs



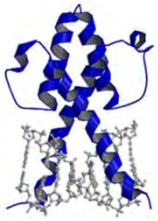
Outline

Why look for TF binding sites?



What is a motif?

How do I compute a motif from bound sites?



Is a motif enriched?

How do you find a motif from unaligned seqs?

Motif Discovery

- **Given**: a set of sequences
- **Find**: the PWM for an over-represented motif

```
ACGTGTCTGCTACAAAATGCAAATACGATGATAAATGCAGCAATTGT
ACGTAAATGCAATTACGATGATAAATGCAGCAACCGTTATCGACTTG
ATCTTACTAGCATGGCCATCATCAACATGCAAAGCAGGTTGTGCCCT
ATAAATGCCCAATTGATTTGTCTCCACTACATAATGCAAATACCATG
```

Motif Discovery

- **Given**: a set of sequences
- **Find**: the PWM for an over-represented motif

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTGT

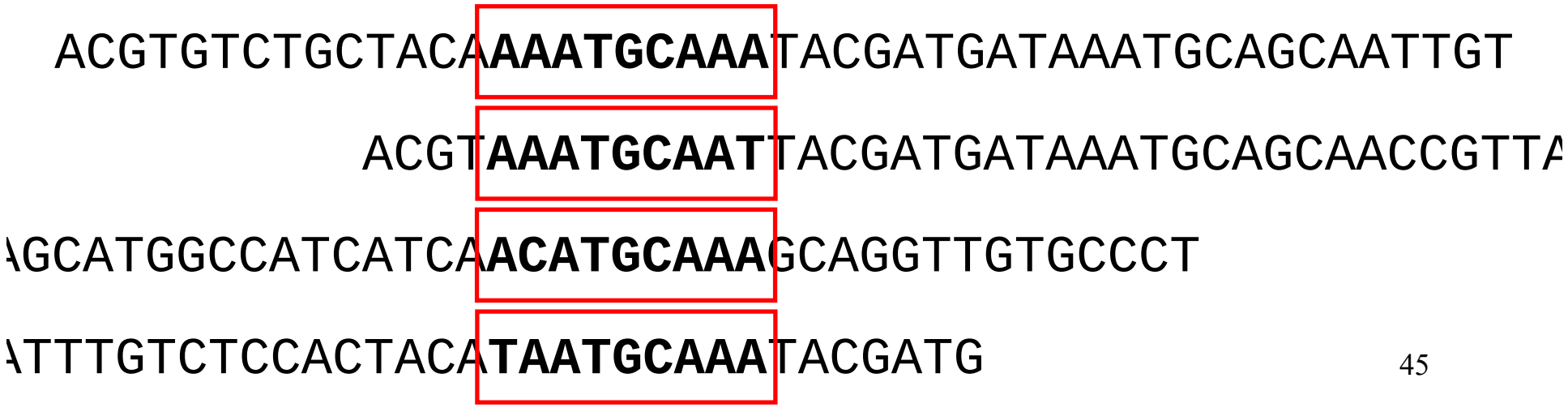
ACGT**AAATGCAAT**TACGATGATAAATGCAGCAACCGTTATCGACTTG

ATCTTACTAGCATGGCCATCATCA**ACATGCAAA**GCAGGTTGTGCCCT

ATAAATGCCCAATTGATTTGTCTCCACTACA**TAATGCAAA**TACGATG

Motif Discovery

- **Note 1:**
  - If you know the PWM, you can easily align the sequences
- **Note 2:**
  - If the sequences are aligned, you can easily find the PWM



The Expectation Maximization (EM) Algorithm

- When we begin
  - we don't know the PWM
  - we don't know the location of the binding sites
- We iteratively:
  - assume we know the motif and look for the most likely binding site
  - assume we know the binding site and compute the best motif

Expectation Maximization

- E step – calculate expected motif locations given the current motif

Given our current best guess about the motif,
Where do we think the protein is binding?

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTGT

ACGT**TCATGTATT**TACGATGATAAATGCAGCAACCGTTATCGACTTG

ATCTTACTAGCATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCT

ATAAATGCCCAATTGATTTGTCTCCACTACA**AAATGCAAT**TACGATG

Expectation Maximization

- M step – re-estimate the motif to maximize likelihood

Given our expectation about where binding occurs,
What is the most likely motif model?

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTG

GACATTTTGTACGT**TCATGTATT**TACGATGATAAATGCAGCAACCG

CATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCGGTTTACTGA

TTGTCTCCACTACA**AAATGCAAT**TACGATGAGAGGGTGATGGCACT

Expectation Maximization

- M step – re-estimate the motif to maximize likelihood

Old motif

New motif

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTG

GACATTTTGTACGT**TCATGTATT**TACGATGATAAATGCAGCAACCG

CATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCGGTTTACTGA

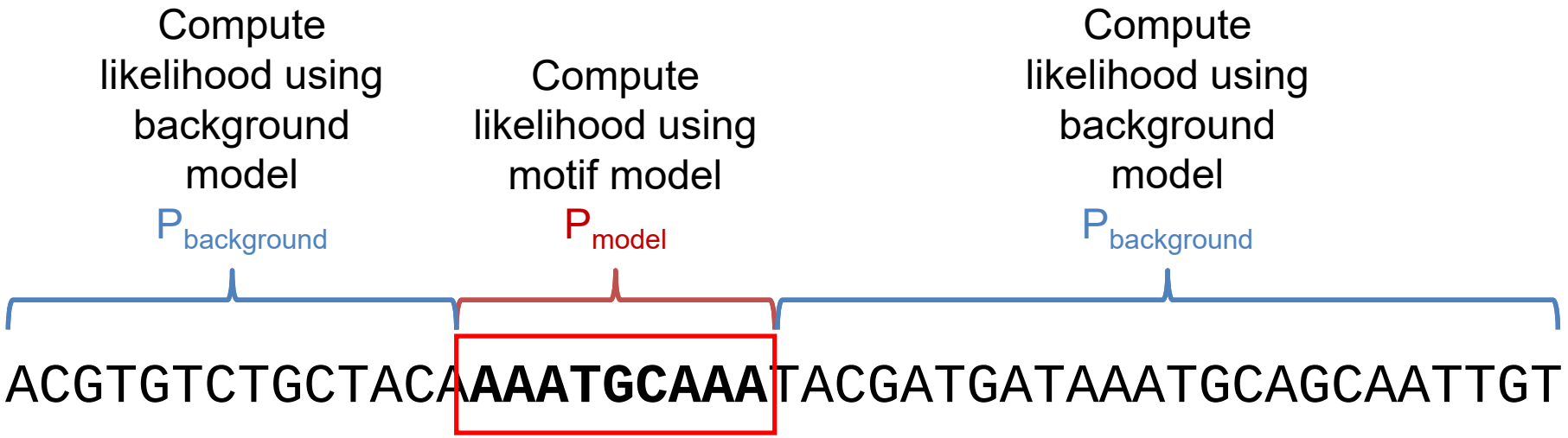
TTGTCTCCACTACA**AAATGCAAT**TACGATGAGAGGGTGATGGCACT

Properties of the EM algorithm

- EM is guaranteed to converge
  - at each step our overall score improves
- EM is not guaranteed to give the right answer
  - had we started with a different initial guess, we might have found a better answer

What do we maximize?

- We maximize the likelihood of the full sequences given our current motif model.



- Remember that each element of the motif is

$$log\left[\frac{P_{model}}{P_{background}}\right] = log[P_{model}] - log[P_{background}]$$