7.91 / 7.36 / BE.490

Lecture #3

Mar. 2, 2004

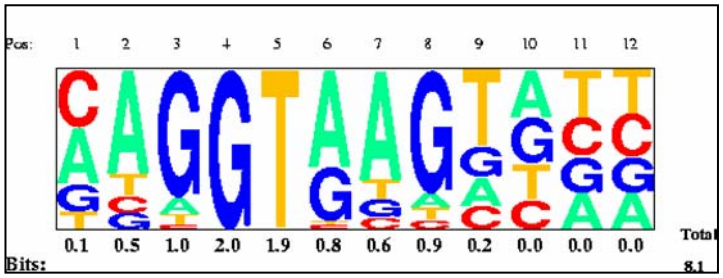
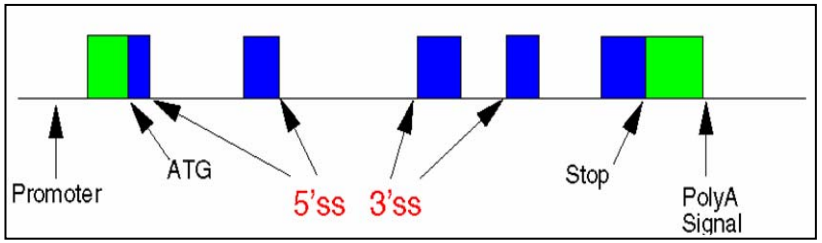
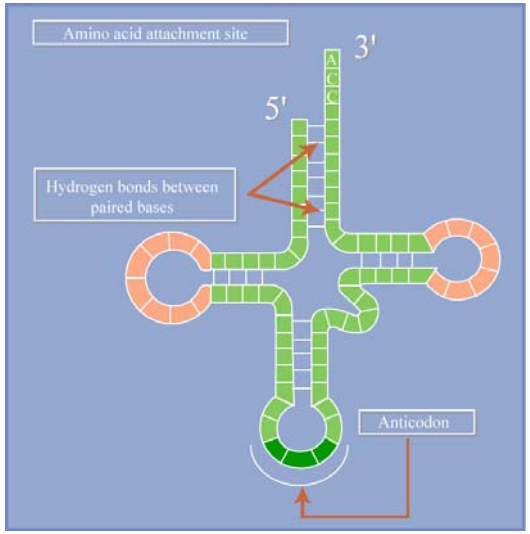
DNA Motif Modeling & Discovery

Chris Burge

Review of DNA Seq. Comparison/Alignment

- Target frequencies and mismatch penalties
- Eukaryotic gene structure
- Comparative genomics applications:
  - Pipmaker (2 species comparison)
  - Phylogenetic Shadowing (many species)
- Intro to DNA sequence motifs

Organization of Topics

Lecture	Object	Model	Dependence Structure
3/2		Weight Matrix Model	Independence
3/4		Hidden Markov Model	Local Dependence
3/9		Energy Model, Covariation Model	Non-local Dependence

DNA Motif Modeling & Discovery

- Review - WMMs for splice sites
- Information Content of a Motif
- The Motif Finding/Discovery Problem
- The Gibbs Sampler

The Gibbs Sampling Algorithm Multimedia Experience

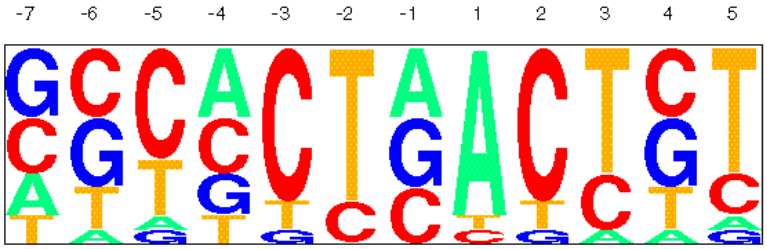
- Motif Modeling - Beyond Weight Matrices

See Ch. 4 of Mount

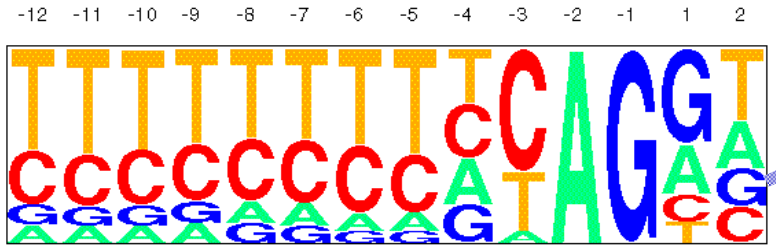
Splicing Model I

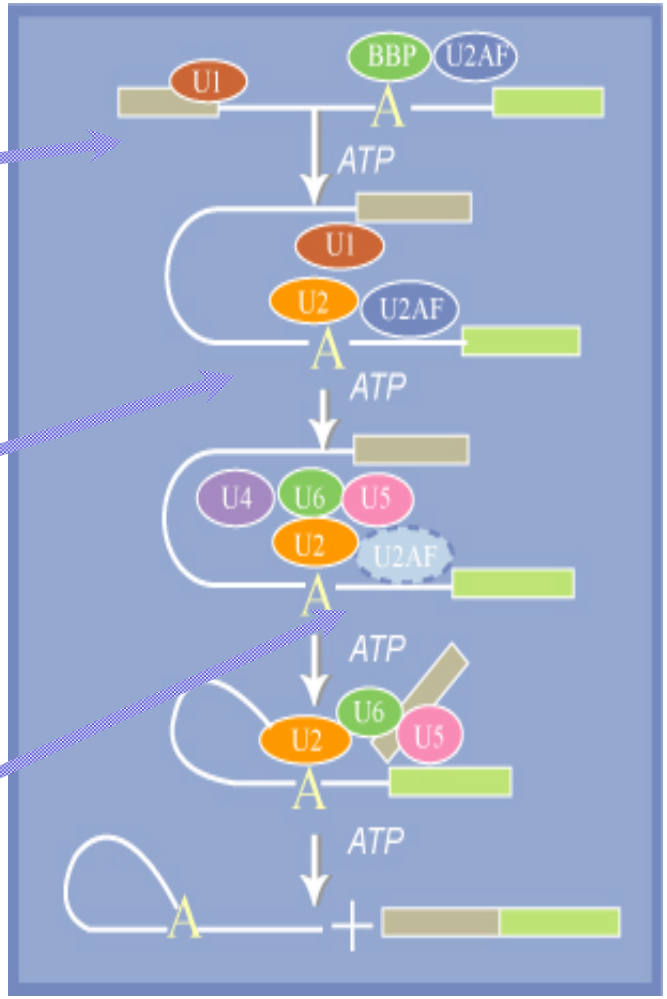
5′ splice site

branch site



3′ splice site





Weight Matrix Models II

5′ splice signal

Con: C A G … G T

Pos	-3	-2	-1	…	+5	+6
A	**0.3**	**0.6**	0.1	…	0.1	0.1
C	**0.4**	0.1	0.0	…	0.1	0.2
G	0.2	0.2	**0.8**	**…**	**0.8**	0.2
T	0.1	0.1	0.1	…	0.0	**0.5**

Background

Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$\text{Odds Ratio: } R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3)\,\bullet\bullet\bullet\, P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3)\,\bullet\bullet\bullet\, P_{bg}(S_8)P_{bg}(S_9)}$$

Background model homogenous, assumes independence

Weight Matrix Models III

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$\text{Odds Ratio: } R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3)\,\bullet\bullet\bullet\, P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3)\,\bullet\bullet\bullet\, P_{bg}(S_8)P_{bg}(S_9)}$$

$$= \prod_{k=1}^{k=9} P_{-4+k}(S_k)/\, P_{bg}(S_k)$$

$$\text{Score } \quad s = \log_2 R \;= \sum_{k=1}^{k=9} \log_2 \left(P_{-4+k}(S_k)/\, P_{bg}(S_k)\right)$$

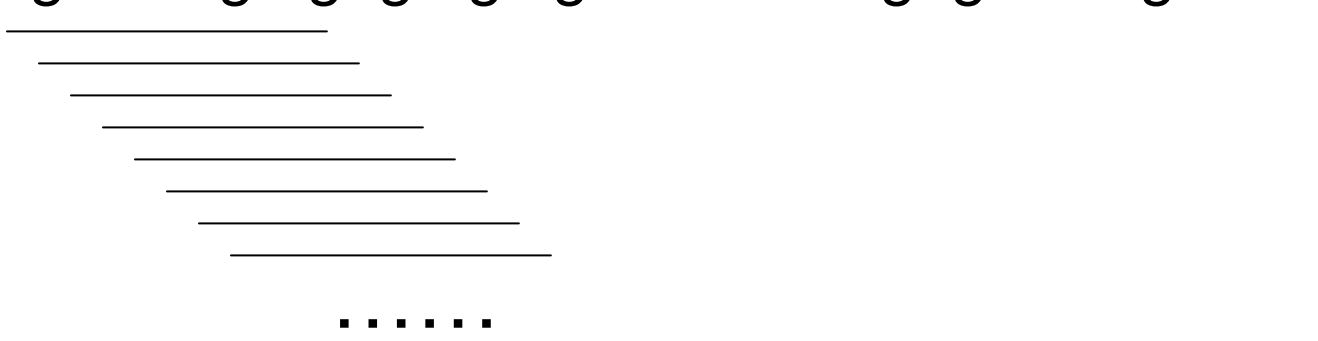
Neyman-Pearson Lemma:

Optimal decision rules are of the form $R > C$

Equiv.: $\log_2(R) > C'$ because log is a monotone function

Weight Matrix Models IV

Slide WMM along sequence:

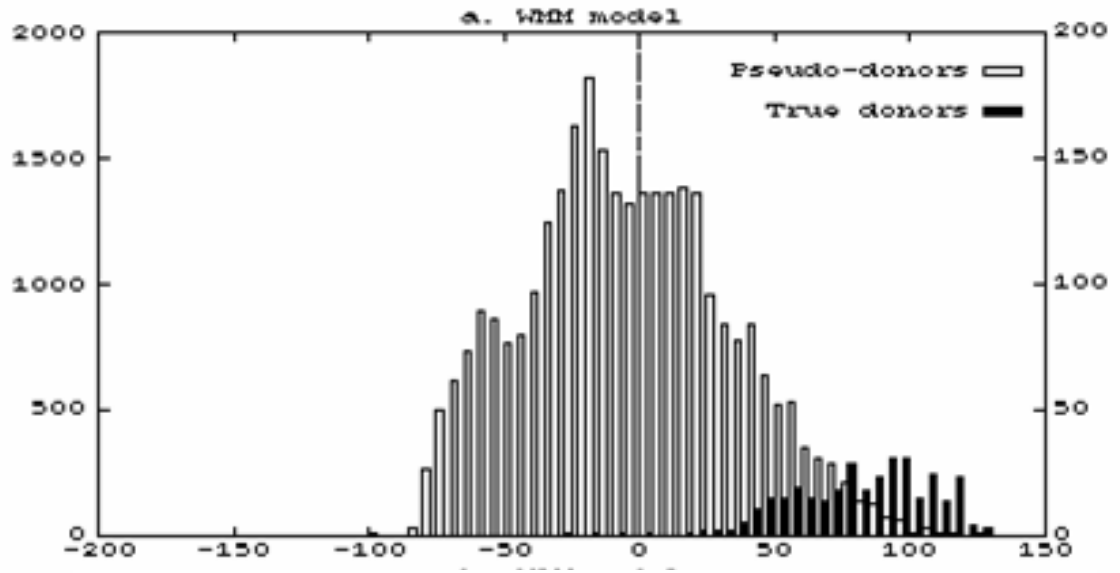


Assign score to each 9 base window.

Use score cutoff to predict potential 5′ splice sites

Histogram of 5′ss Scores

"Decoy" 5′ Splice Sites



True 5′ Splice Sites

Score (1/10 bit units)

Measuring Accuracy:

Sensitivity = % of true sites w/ score > cutoff

Specificity = % of sites w/ score > cutoff that are true sites

Sn:	20%	50%	90%
Sp:	50%	32%	7%

What does this result tell us?

A) Splicing machinery also uses other information besides 5'ss motif to identify splice sites;

OR

B) WMM model does not accurately capture some aspects of the 5'ss that are used in recognition

(or both)

This is a pretty common situation in biology

What is a DNA (RNA) Motif ?

A pattern common to a set of DNA (RNA) sequences that share a common biological property, such as being binding sites for a regulatory protein

Common motif adjectives:

exact/precise *versus* degenerate

strong *versus* weak (good *versus* lousy)

high information content *versus* low information content

Information Theory

So we end up with Shannon's famous formula:

$$H = -\sum_{i=1}^{20} Pi(\log_2 Pi)$$

Where H = the "Shannon Entropy"
In bits per position in the alignment

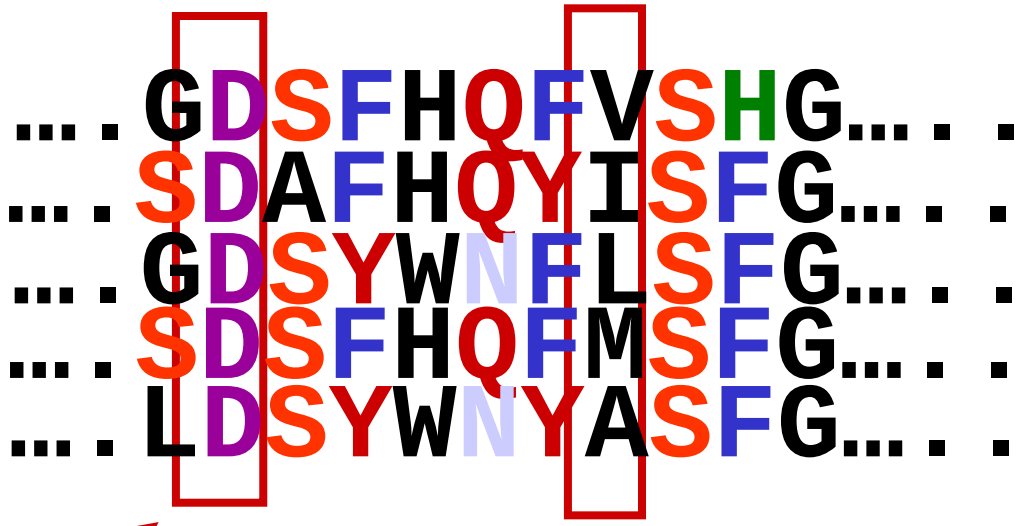
What does this mean???

H is a measure of entropy or randomness or disorder ….it tells us how much uncertainty there is for the different amino acid abundances at one position in a sequence motif

This slide courtesy of M. Yaffe

Information Theory

Courtesy of M. Yaffe



Assuming all 20 amino acids equally possible:

H_{before} = 4.32, H_{after}=0

Therefore, this position encodes 4.32-0 =4.32 bits of information!

Another position in the motif that contains all 20 amino acids…

H_{before} = 4.32, H_{after}=4.32

Therefore, this position encodes 4.32-4.32 =0 bits of information!

Information Content of a DNA Motif

Information at position j: $I_j = H_{before} - H_{after}$

Motif probabilities: p_k (k = A, C, G, T)

Background probabilities: $q_k = \frac{1}{4}$ (k = A, C, G, T)

$$I_j = -\sum_{k=1}^{4} q_k \log_2 q_k - -\sum_{k=1}^{4} p_k \log_2 p_k = 2 - H_j$$

$$I_{motif} = \sum_{j=1}^{w} I_j = 2w - H_{motif}$$ (motif of width w bases)

Log base 2 gives entropy/information in ‘bits’

Mean Bit-score of a Motif

Bit-score: $\log_2(\frac{p_k}{q_k})$

Mean bit-score: (motif width w, $n = 4^w$, $q_k = \frac{1}{4^w}$)

$\sum_{k=1}^{n} p_k \log_2(\frac{p_k}{q_k})$ $= 2w - H_{motif} = I_{motif}$

Rule of thumb*: motif w/ m bits of information will occur about once every 2^m bases of random sequence

* True for regular expressions, approx. true for other motifs

The Motif Finding Problem

Unaligned

agggcactagcccatgtgagagggcaaggaccagcggaag
taattcagggccaggatgtatctttctcttaaaaataaca
tatcctacagatgatgaatgcaaatcagcgtcacgagctt
tggcgggcaaggtgcttaaaagataatatcgaccctagcg
attcgggtaccgttcataaaagtacgggaatttcgggtag
gttatgttaggcgagggcaaaagtcatatacttttaggtc
aagagggcaatgcctcctctgccgattcggcgagtgatcg
gatggggaaaatatgagaccaggggagggccacactgcag
ctgccgggctaacagacacacgtctagggctgtgaaatct
gtaggcgccgaggccaacgctgagtgtcgatgttgagaac
attagtccggttccaagagggcaactttgtatgcaccgcc
gcggcccagtgcgcaacgcacagggcaaggtttactgcgg
ccacatgcgagggcaacctccctgtgttgggcggttctga
gcaattgtaaaacgacggcaatgttcggtcgcctaccctg
gataaagagggggtaggaggtcaactcttccgtattaat
aggagtagagtagtgggtaaactacgaatgcttataacat
gcgagggcaatcgggatctgaaccttctttatgcgaagac
tccaggaggaggtcaacgactctgcatgtctgacaacttg
gtcatagaattccatccgccacgcggggtaatttggacgt
gtgccaacttgtgccgggggctagcagcttcccgtcaaa
cgcgtttggagtgcaaacatacacagcccgggaatataga
aagatacgagttcgatttcaagagttcaaaacgtgacggg
gacgaaacgagggcgatcaatgcccgataggactaataag
tagtacaaacccgctcacccgaaaggagggcaaatacctt
atatacagccaggggagacctataactcagcaaggttcag
cgtatgtactaattgtggagagcaaatcattgtccacgtg
…

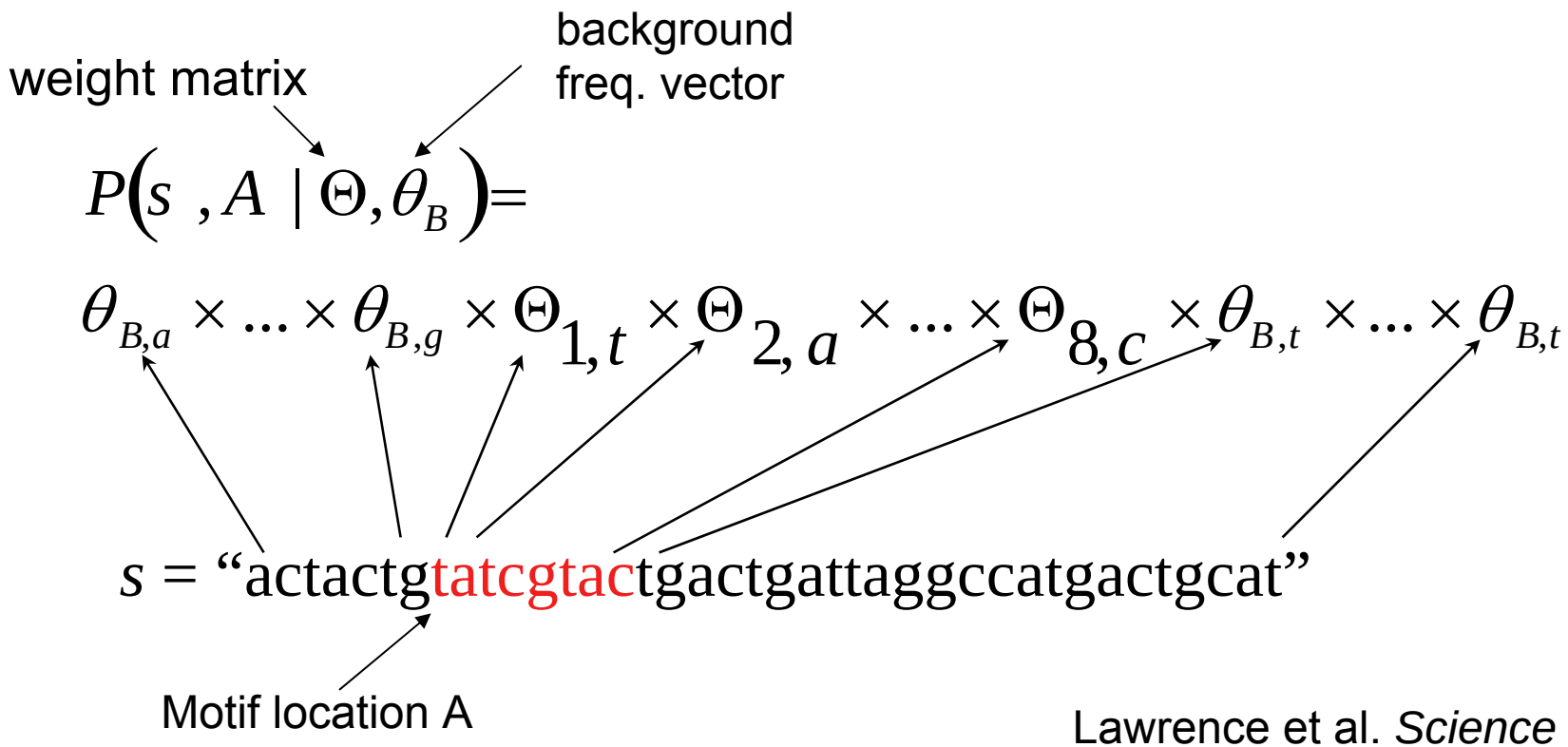
Aligned

gcggaagagggcactagcccatgtgagagggcaaggacca
atctttctcttaaaaataacataattcagggccaggatgt
gtcacgagctttatcctacagatgatgaatgcaaatcagc
taaaagataatatcgaccctagcgtggcgggcaaggtgct
gtagattcgggtaccgttcataaaagtacgggaatttcgg
tatacttttaggtcgttatgttaggcgagggcaaaagtca
ctctgccgattcggcgagtgatcgaagagggcaatgcctc
aggatggggaaaatatgagaccaggggagggccacactgc
acacgtctagggctgtgaaatctctgccgggctaacagac
gtgtcgatgttgagaacgtaggcgccgaggccaacgctga
atgcaccgccattagtccggttccaagagggcaactttgt
ctgcgggcggcccagtgcgcaacgcacagggcaaggttta
tgtgttgggcggttctgaccacatgcgagggcaacctccc
gtcgcctaccctggcaattgtaaaacgacggcaatgttcg
cgtattaatgataaagagggggtaggaggtcaactcttc
aatgcttataacataggagtagagtagtgggtaaactacg
tctgaaccttctttatgcgaagacgcgagggcaatcggga
tgcatgtctgacaacttgtccaggaggaggtcaacgactc
cgtgtcatagaattccatccgccacgcggggtaatttgga
tcccgtcaaagtgccaacttgtgccgggggctagcagct
acagcccgggaatatagacgcgtttggagtgcaaacatac
acgggaagatacgagttcgatttcaagagttcaaaacgtg
cccgataggactaataaggacgaaacgagggcgatcaatg
ttagtacaaacccgctcacccgaaaggagggcaaatacct
agcaaggttcagatatacagccaggggagacctataactc
gtccacgtgcgtatgtactaattgtggagagcaaatcatt
…

Motif Finding Example: The Gibbs Sampler

The Gibbs sampler is a Monte-Carlo method, which seeks to maximize a likelihood function over the input sequence data.

The likelihood function for a sequence *s* with a motif in location *A*



Lawrence et al. *Science* 1993

Prepare Yourself for

The Gibbs Sampler Multimedia Experience

Featuring the Gibbs Sampling Algorithm in:

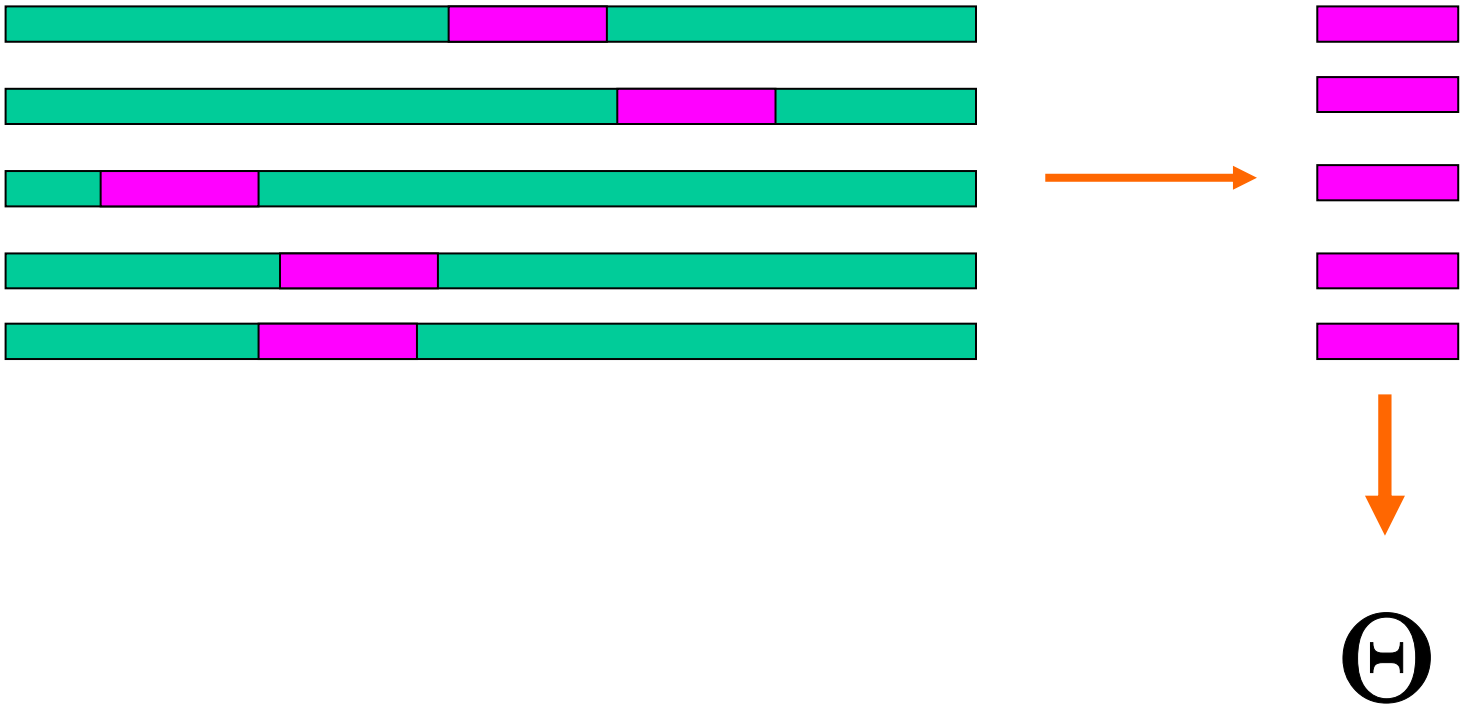
- Pictures
- Words
- Movies

Gibbs Sampling Algorithm I

1. Select a random position in each sequence

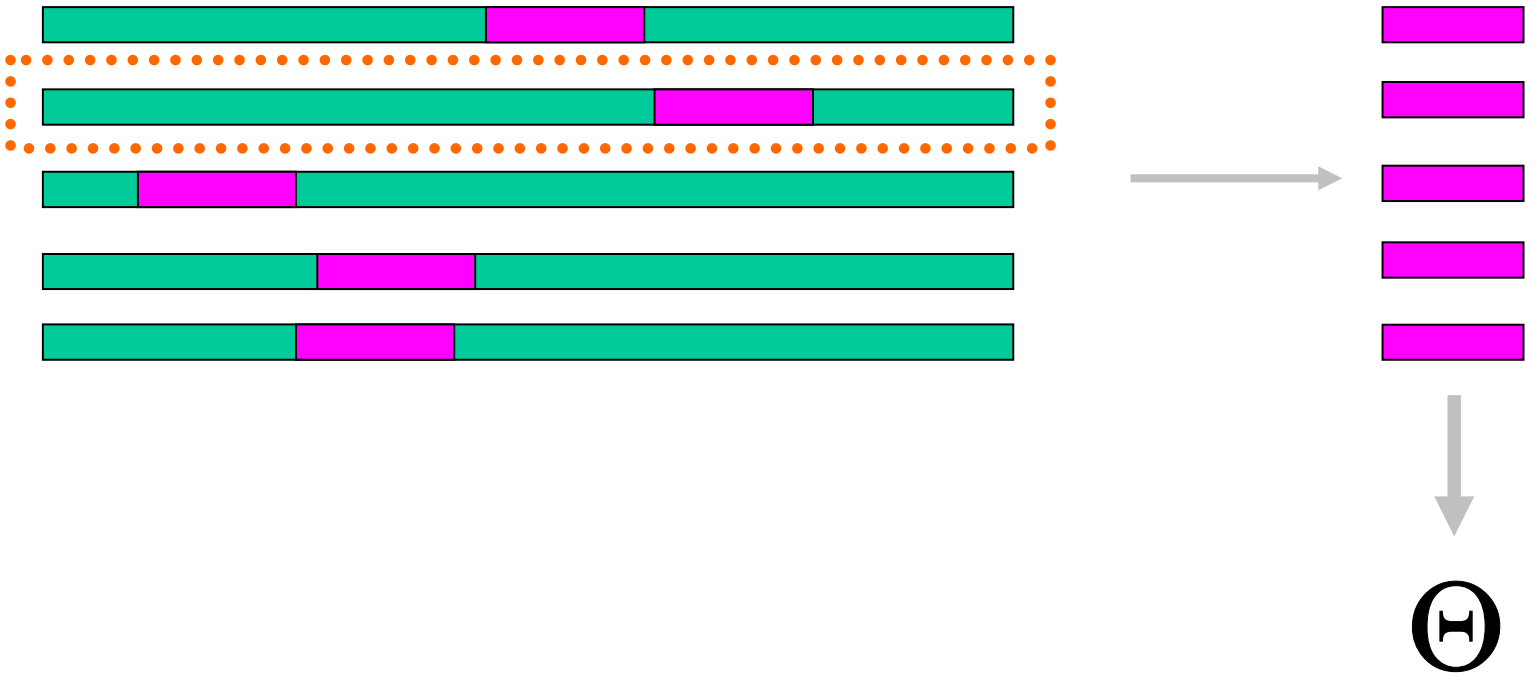
Gibbs Sampling Algorithm II

2. Build a weight matrix



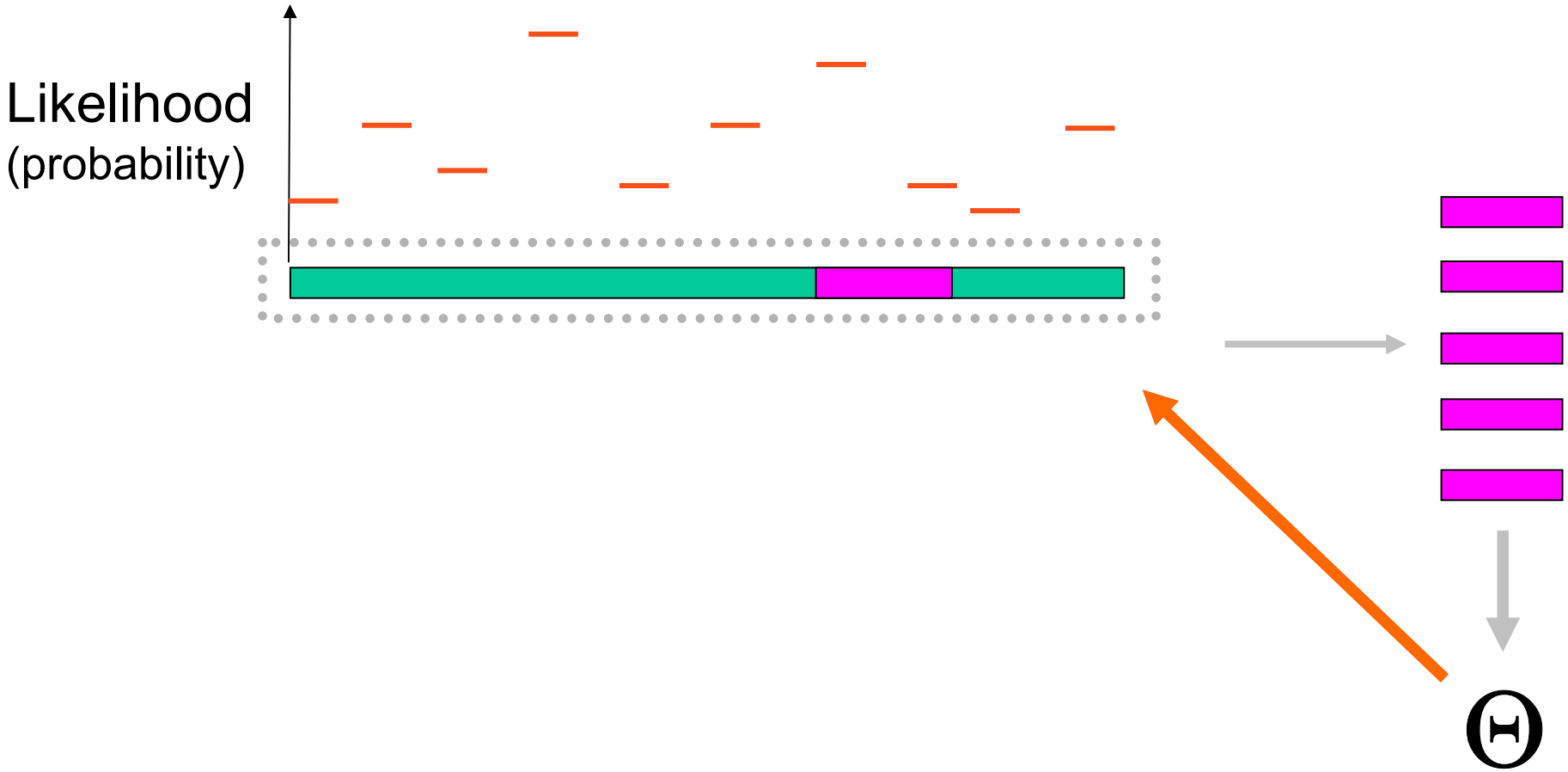
Gibbs Sampling Algorithm III

3. Select a sequence at random



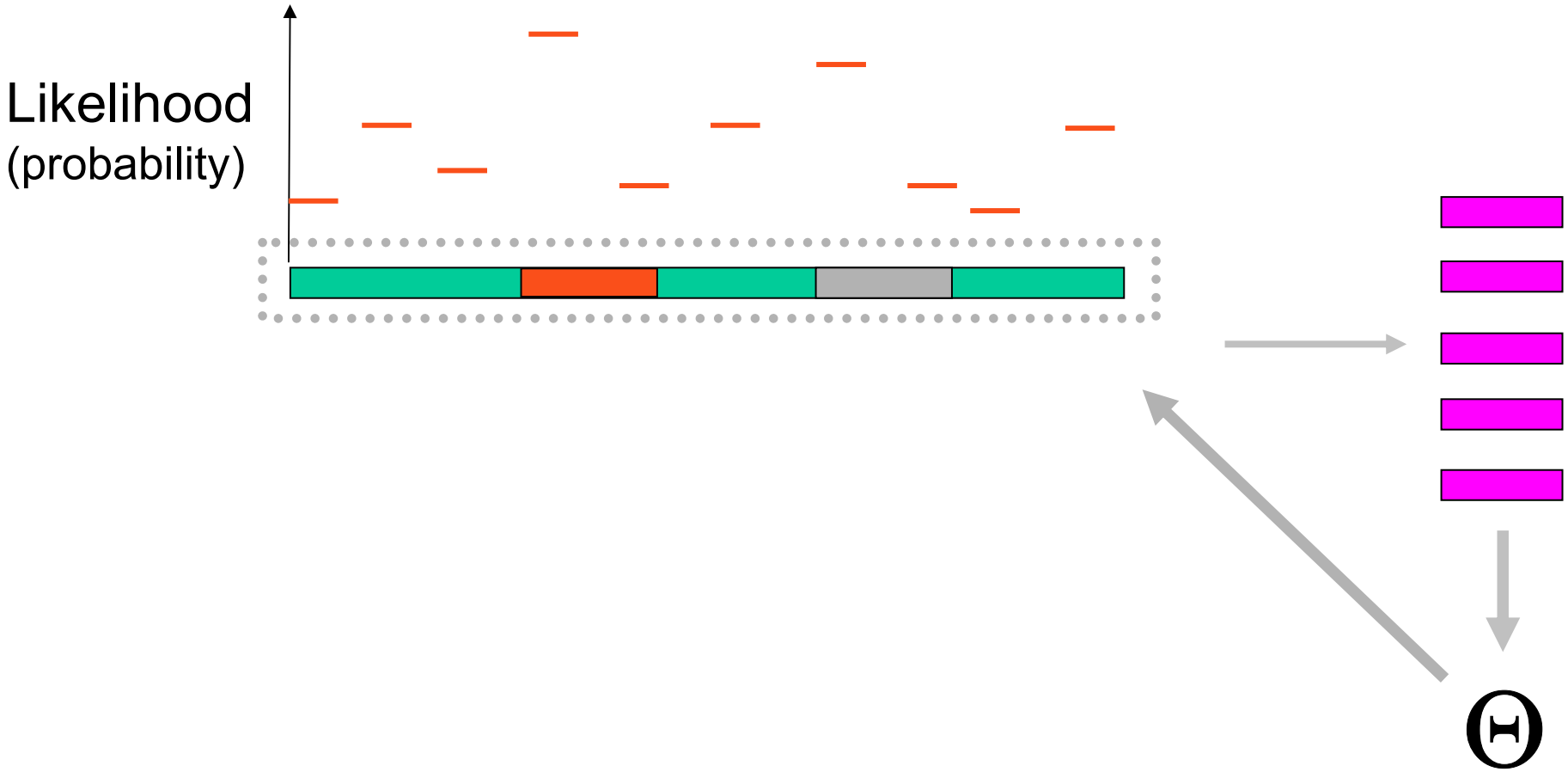
Gibbs Sampling Algorithm IV

4. Score possible sites in seq using weight matrix



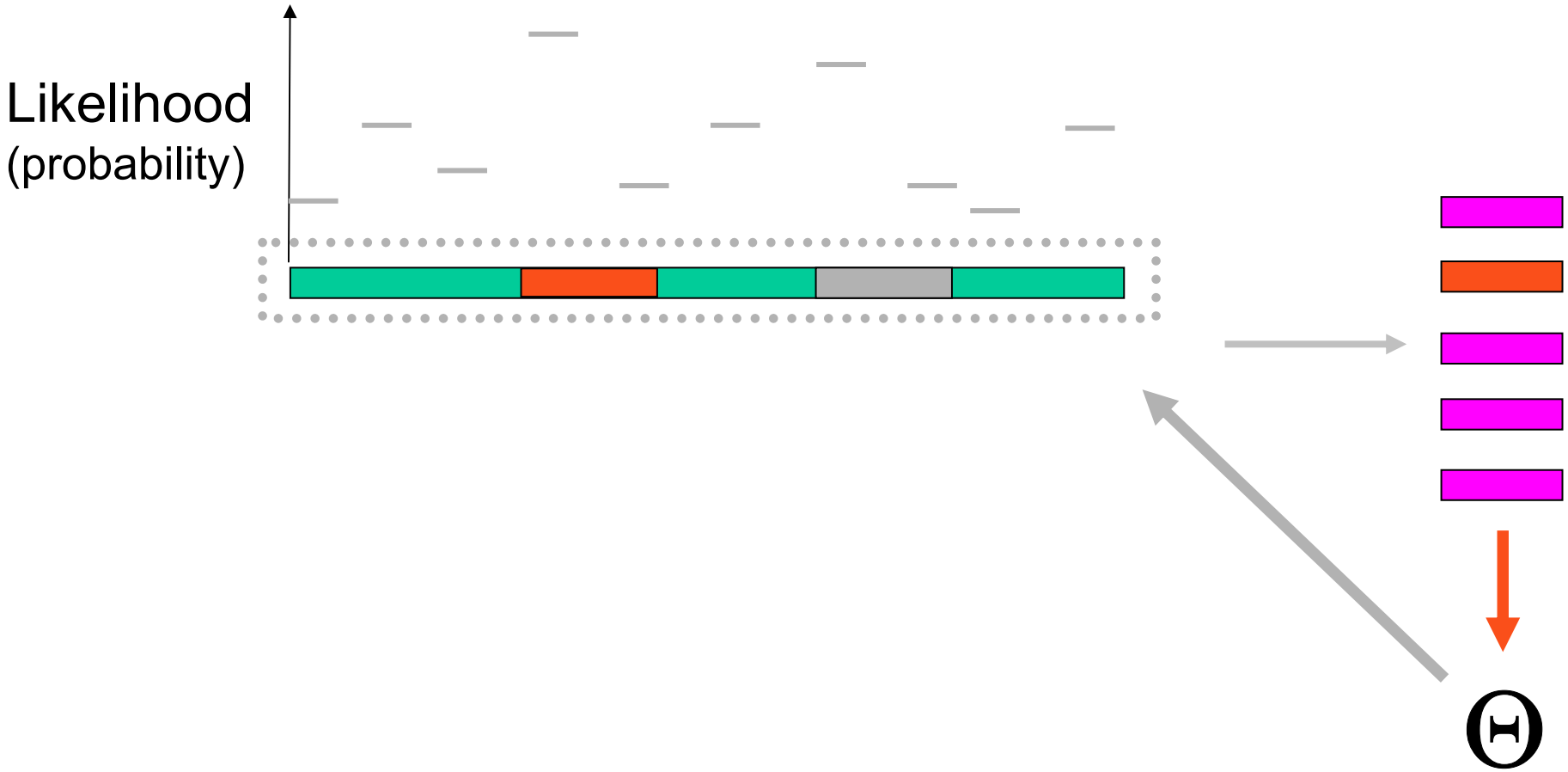
Gibbs Sampling Algorithm V

5. Sample a new site proportional to likelihood



Gibbs Sampling Algorithm VI

6. Update weight matrix



Gibbs Sampling Algorithm VII

7. Iterate until convergence (no change in sites/Θ)

The Gibbs Sampling Algorithm In Words I

Given **N** sequences of length **L** and desired motif width **W:**

Step 1) Choose a starting position in each sequence at random:

$\mathbf{a_1}$ in seq 1, $\mathbf{a_2}$ in seq 2, …, $\mathbf{a_N}$ in sequence **N**

Step 2) Choose a sequence at random from the set (say, seq 1).

Step 3) Make a weight matrix model of width **W** from the sites in all sequences *except* the one chosen in step 2.

Step 4) Assign a probability to each position in seq 1 using the weight matrix model constructed in step 3:

$\mathbf{p} = \{ \mathbf{p_1}, \mathbf{p_2}, \mathbf{p_3}, \ldots, \mathbf{p_{L-W+1}} \}$

Lawrence et al., *Science* 1993

The Gibbs Sampling Algorithm In Words II

Given **N** sequences of length **L** and desired motif width **W:**

Step 5) Sample a starting position in seq 1 based on this probability distribution and set $\mathbf{a_1}$ to this new position.

Step 6) Choose a sequence at random from the set (say, seq 2).

Step 7) Make a weight matrix model of width **W** from the sites in all sequences *except* the one chosen in step 6.

Step 8) Assign a probability to each position in seq 2 using the weight matrix model constructed in step 7.

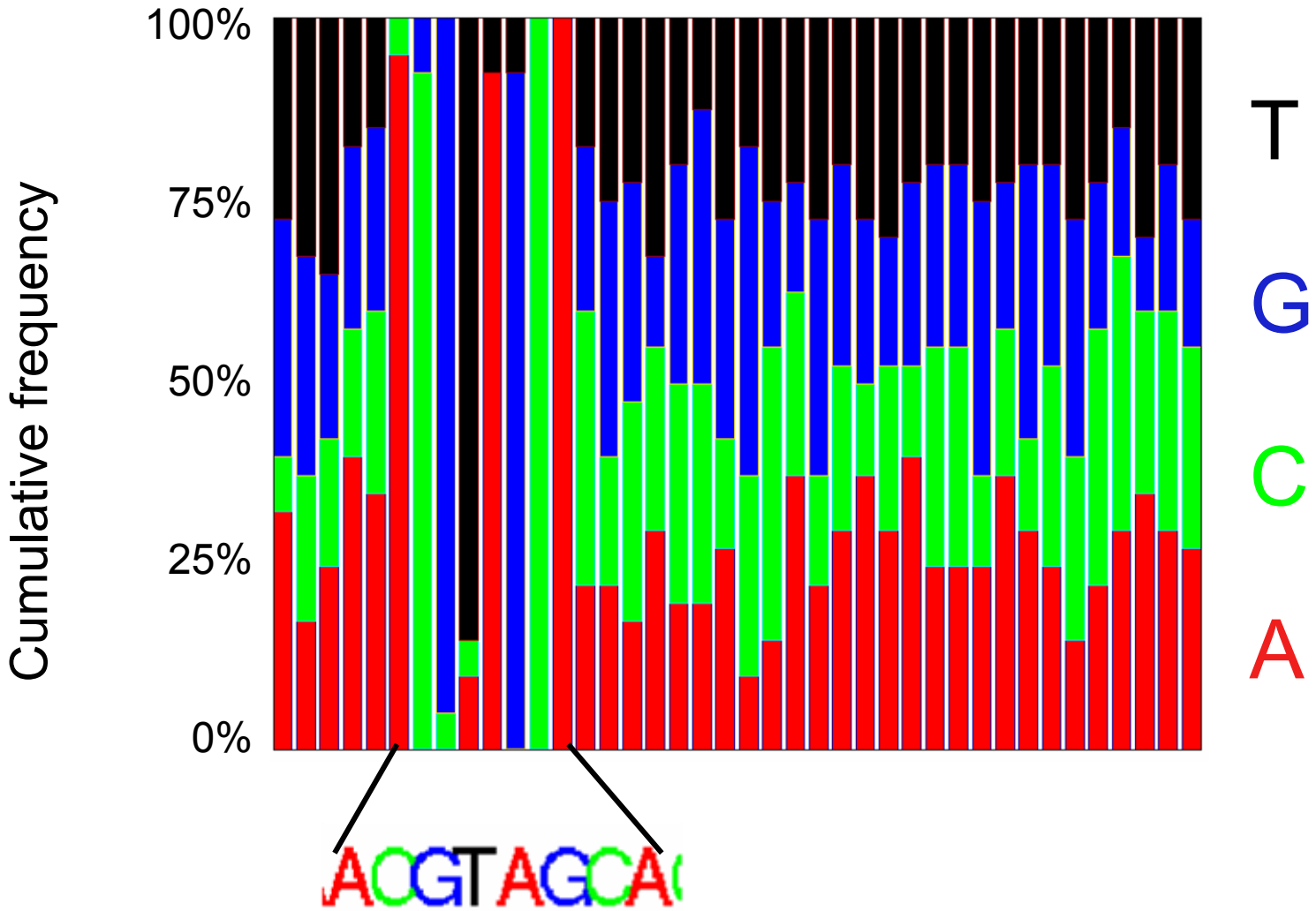
Step 9) Sample a starting position in seq 2 based on this dist.

Step 10) Repeat until convergence

Lawrence et al., *Science* 1993

What, if anything, does this algorithm accomplish (besides keeping your computer busy)?

Input Sequences with Strong Motif



Input Sequences (Weak Motif)

gcggaagagggcactagcccatgtgagagggcaaggacca
atctttctcttaaaaataacataattcagggccaggatgt
gtcacgagctttatcctacagatgatgaatgcaaatcagc
taaaagataatatcgaccctagcgtggcgggcaaggtgct
gtagattcgggtaccgttcataaaagtacgggaatttcgg
tatacttttaggtcgttatgttaggcgagggcaaaagtca
ctctgccgattcggcgagtgatcgaagagggcaatgcctc
aggatggggaaaatatgagaccaggggagggccacactgc
acacgtctagggctgtgaaatctctgccgggctaacagac
gtgtcgatgttgagaacgtaggcgccgaggccaacgctga
atgcaccgccattagtccggttccaagagggcaactttgt
ctgcgggcggcccagtgcgcaacgcacagggcaaggttta
tgtgttgggcggttctgaccacatgcgagggcaacctccc
gtcgcctaccctggcaattgtaaaacgacggcaatgttcg
cgtattaatgataaagagggggtaggaggtcaactcttc
aatgcttataacataggagtagagtagtgggtaaactacg
tctgaaccttctttatgcgaagacgcgagggcaatcggga
tgcatgtctgacaacttgtccaggaggaggtcaacgactc
cgtgtcatagaattccatccgccacgcggggtaatttgga
tcccgtcaaagtgccaacttgtgccgggggctagcagct
acagcccgggaatatagacgcgtttggagtgcaaacatac
acgggaagatacgagttcgatttcaagagttcaaaacgtg
cccgataggactaataaggacgaaacgagggcgatcaatg
ttagtacaaacccgctcacccgaaaggagggcaaatacct
agcaaggttcagatatacagccaggggagacctataactc
gtccacgtgcgtatgtactaattgtggagagcaaatcatt
…

Gibbs Sampler Summary

- A stochastic (Monte Carlo) algorithm for motif finding
- Works by ‘stumbling’ onto a few motif instances, which bias the weight matrix, which causes it to sample more motif instances, which biases the weight matrix more, … until convergence
- Not guaranteed to converge to same motif every time - run several times, compare results
- Works for protein, DNA, RNA motifs

MEME - Multiple EM for Motif Elicitation

- Another popular motif finding algorithm - optimizes a similar likelihood function using an algorithm called ‘expectation maximization’ (EM)

- Unlike Gibbs Sampler, MEME is deterministic

Bailey & Elkan, Proc. ISMB, 1994

Weight Matrix Models II

5′ splice signal

Con: C A G … G T

Pos	-3	-2	-1	…	+5	+6
A	**0.3**	**0.6**	0.1	…	0.1	0.1
C	**0.4**	0.1	0.0	…	0.1	0.2
G	0.2	0.2	**0.8**	**…**	**0.8**	0.2
T	0.1	0.1	0.1	…	0.0	**0.5**

Background

Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

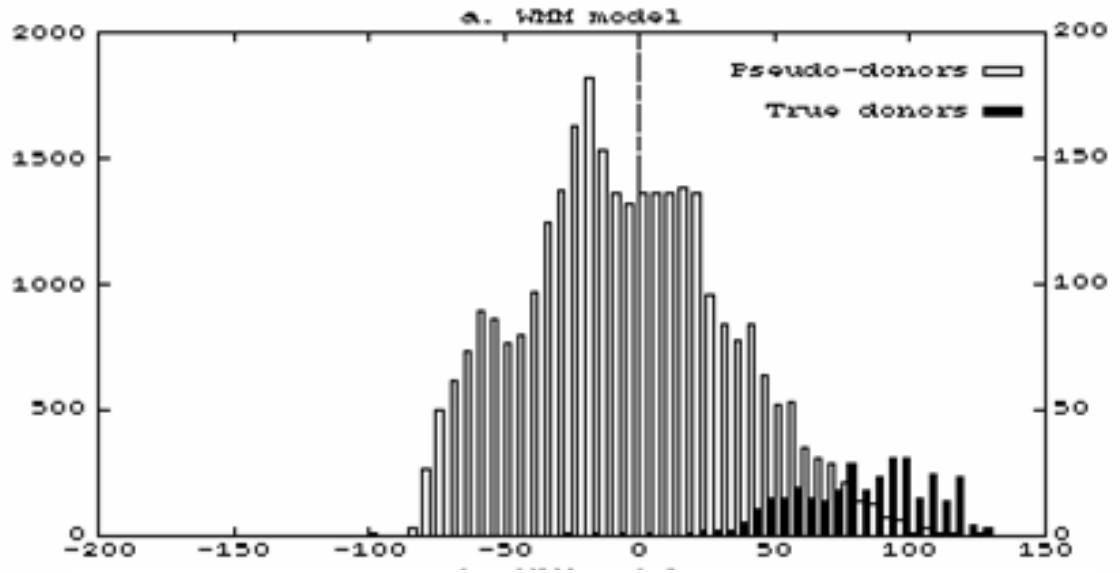
$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$\text{Odds Ratio: } R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3)\,\bullet\bullet\bullet\, P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3)\,\bullet\bullet\bullet\, P_{bg}(S_8)P_{bg}(S_9)}$$

Background model homogenous, assumes independence

Histogram of 5′ss Scores

"Decoy" 5′ Splice Sites



True 5′ Splice Sites

Score (1/10 bit units)

Measuring Accuracy:

Sensitivity = % of true sites w/ score > cutoff

Specificity = % of sites w/ score > cutoff that are true sites

Sn:	20%	50%	90%
Sp:	50%	32%	7%

What does this result tell us?

A) Splicing machinery also uses other information besides 5'ss motif to identify splice sites;

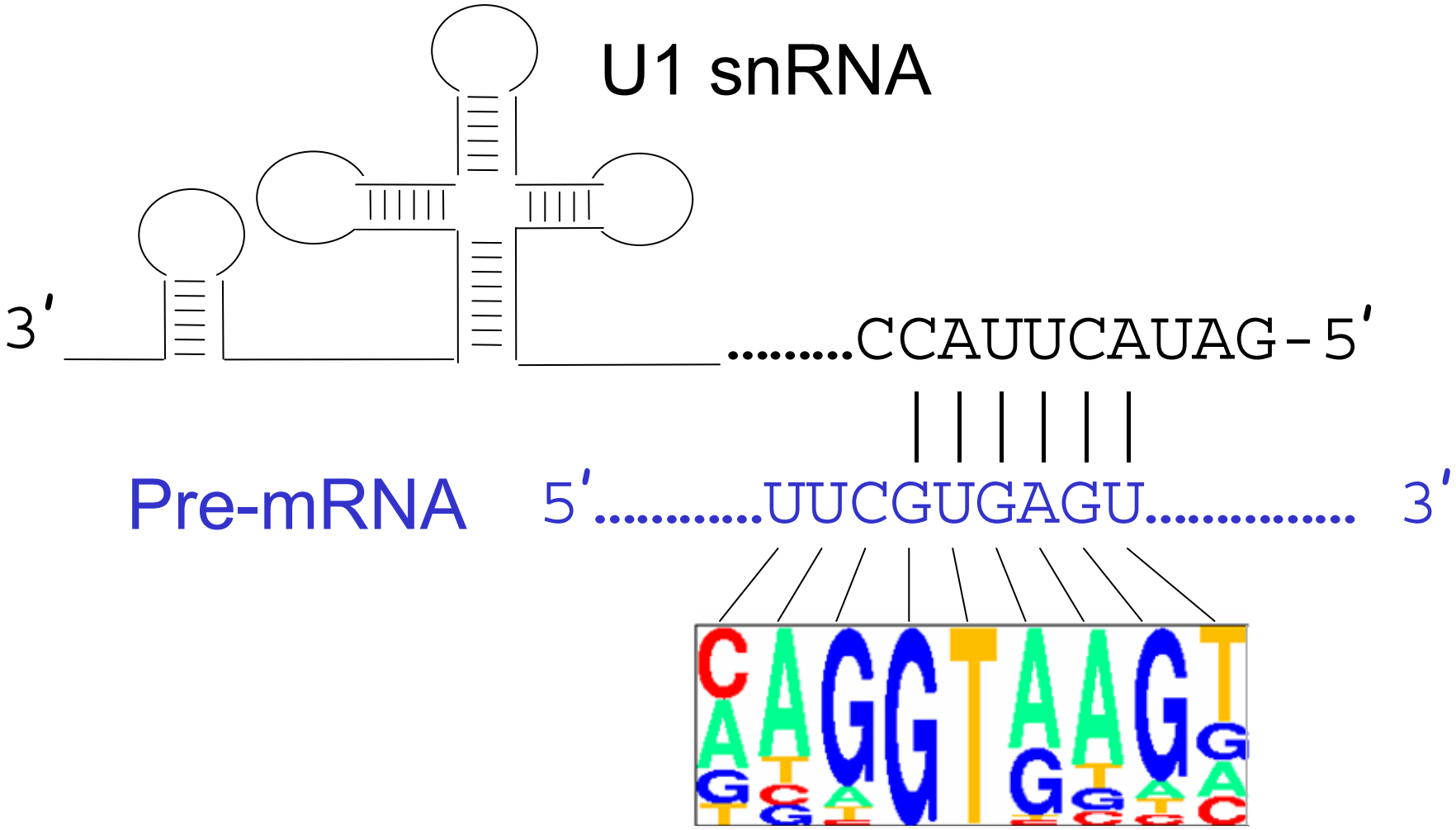
OR

B) WMM model does not accurately capture some aspects of the 5'ss that are used in recognition

(or both)

This is a pretty common situation in biology

How is the 5′ss recognized?

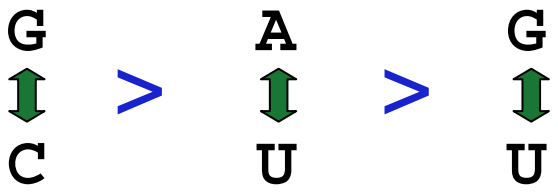


RNA Energetics I

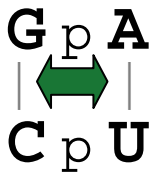
```
   ...CCAUUCAUAG-5'
      ||||||
5'...CGUGAGU.....  3'
```

Free energy of helix formation derives from:

- base pairing:



- base stacking:



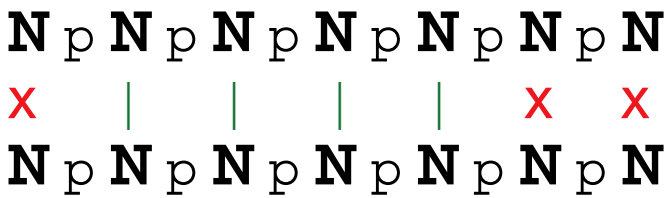
Doug Turner's Energy Rules:

```
                5' --> 3'
                   UX
                   AY
                3' <-- 5'

                      X
Y       A         C         G         U
A       .         .         .     -1.30
C       .         .     -2.40         .
G       .     -2.10         .     -1.00
T   -0.90         .     -1.30         .
```

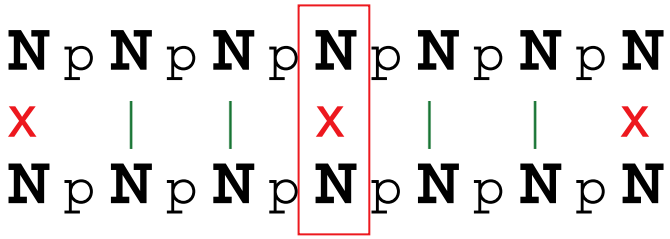
RNA Energetics II

A)



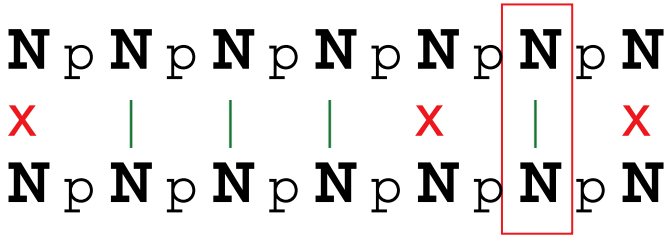
Lots of consecutive base pairs - good

B)



Internal loop - bad

C)



Terminal base pair not stable - bad

Generally A will be more stable than B or C

Conditional Frequencies in 5′ss Sequences

5′ss which have G at +5

Pos	-1		+3	+4	+6
A	9		**44**	**75**	14
C	4		3	4	18
G	**78**		**51**	13	19
T	9		3	9	**49**

5′ss which lack G at +5

Pos	-1		+3	+4	+6
A	2		**81**	**51**	22
C	1		3	28	20
G	**97**		**15**	9	30
T	0		2	12	**28**

Data from Burge, 1998 “Computational Methods in Molecular Biology”

What kind of model could incorporate interactions between positions?