第5-1章 Motif Finding

- Motif finding problem
- EM algorithm
- Markov chain Monte Carlo (Gibbs Sampler)

部分Slides来源于：
http://www.broadinstitute.org/annotation/winter_course_2006/index_files/Biological_Motif_Discovery.ppt
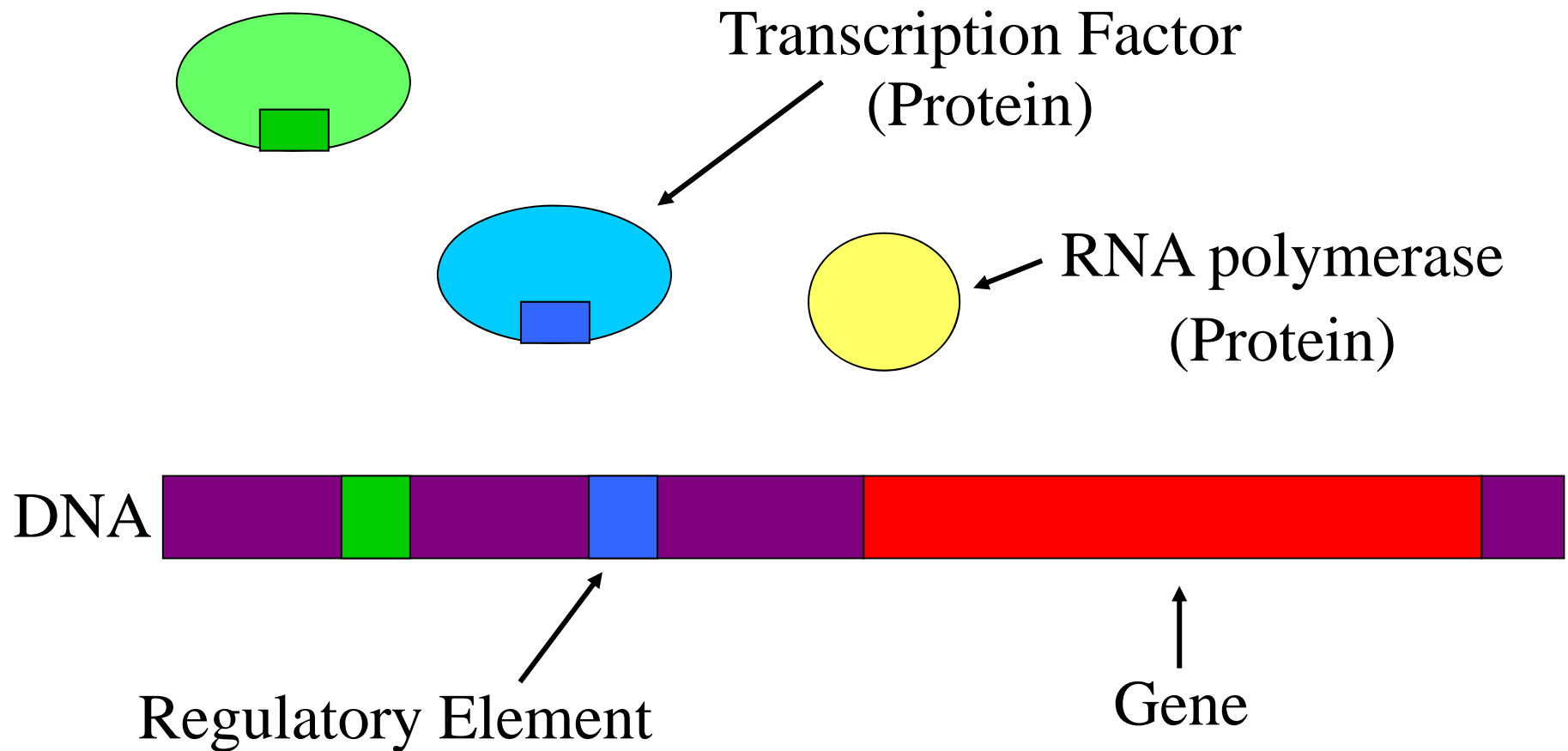http://ai.stanford.edu/~serafim/cs262/Slides/Lecture17.ppt
http://people.csail.mit.edu/manoli/6096/Lecture4_GibbsSampling3.ppt
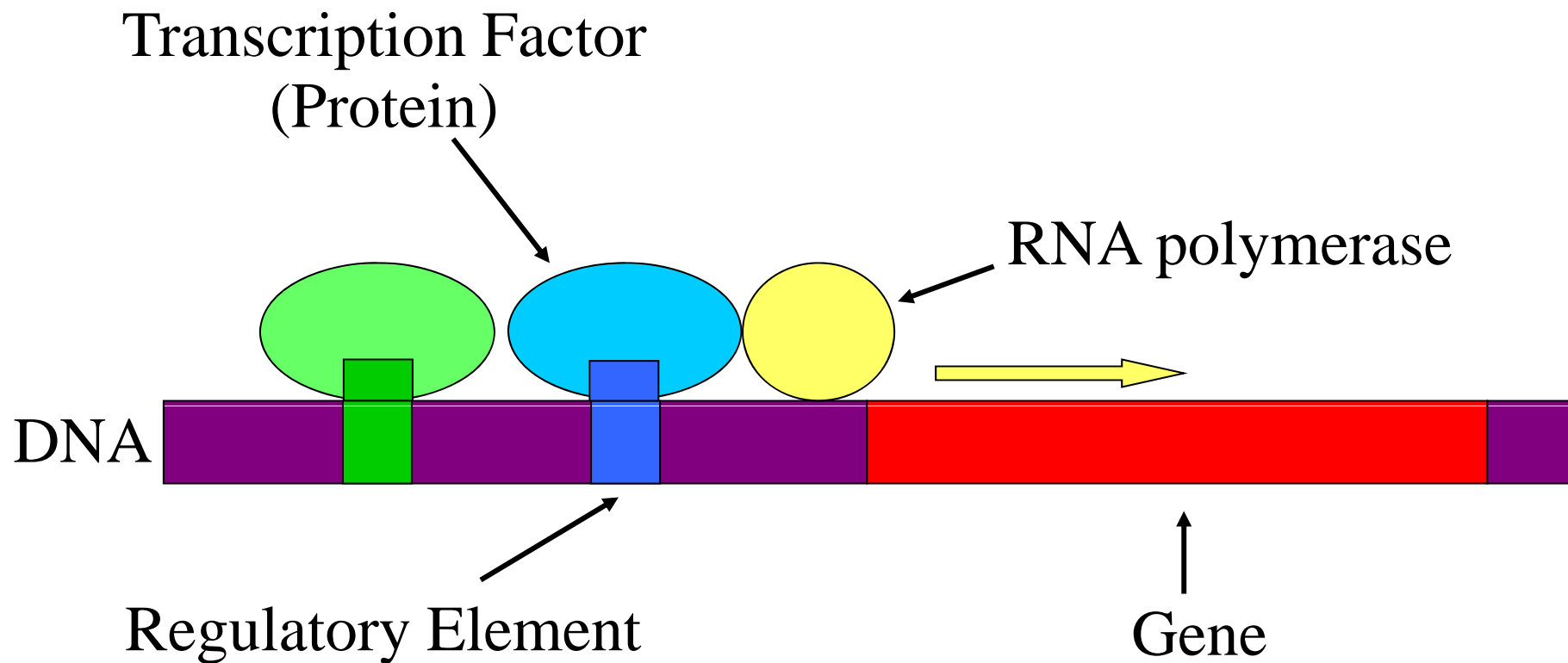
Transcriptional Regulation

- The transcription of each gene is controlled by a regulatory region of DNA relatively near the transcription start site (TSS).

- two types of fundamental components
  - short DNA regulatory elements
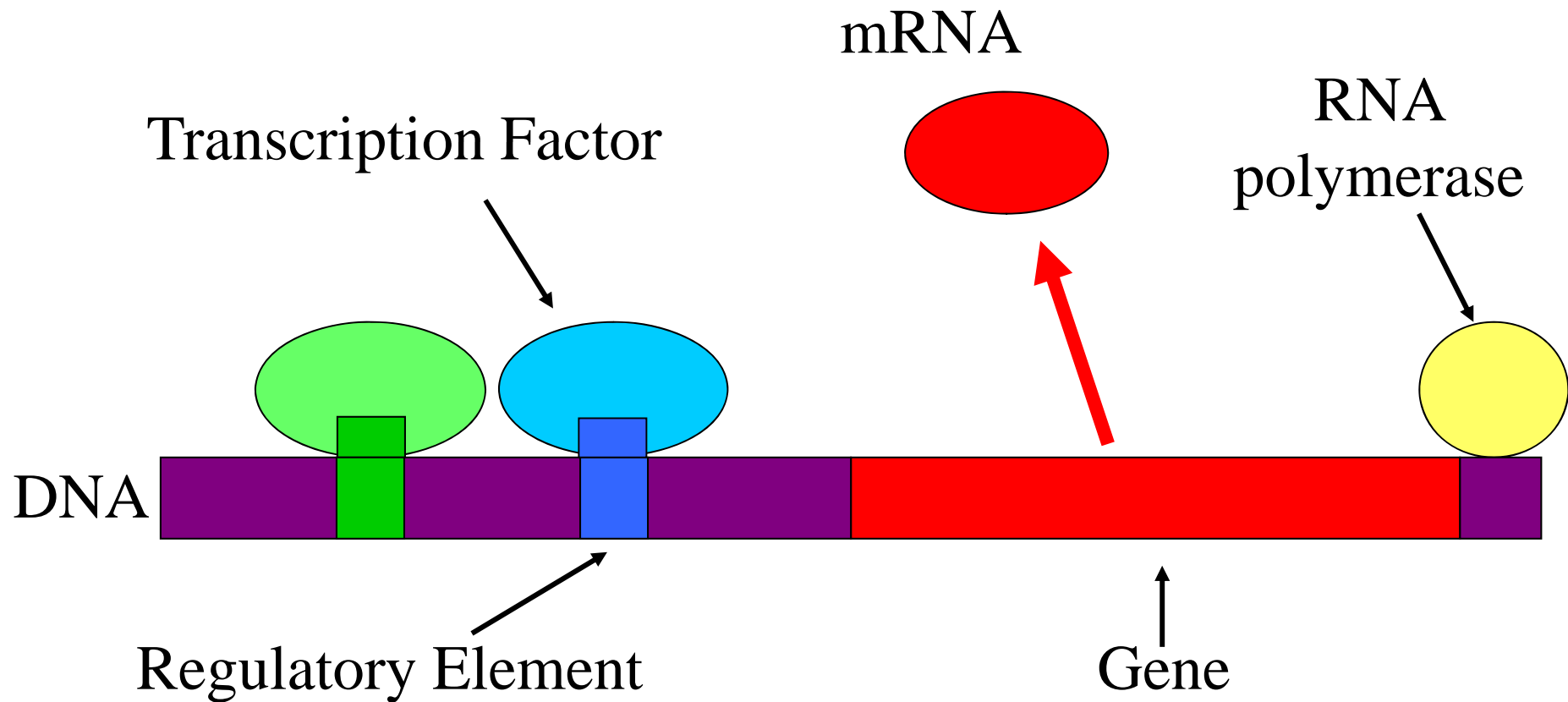 - *gene regulatory proteins* that recognize and bind to them.

Regulation of Genes

Transcription Factor
(Protein)

RNA polymerase
(Protein)

DNA

Regulatory Element

Gene

source: M. Tompa, U. of Washington

3

Regulation of Genes

Transcription Factor (Protein)

RNA polymerase

DNA

Regulatory Element

Gene

source: M. Tompa, U. of Washington

4

Regulation of Genes

source: M. Tompa, U. of Washington

Transcriptional Binding Site

Wiki: DNA binding sites are a type of binding site found in DNA where other molecules may bind

- Small (6-20bp)
- Highly variable

Experimental Method (I)

- DNase footprinting assay: The method uses an enzyme, deoxyribonuclease (DNase, for short), to cut the radioactively end-labeled DNA, followed by gel electrophoresis to detect the resulting cleavage pattern
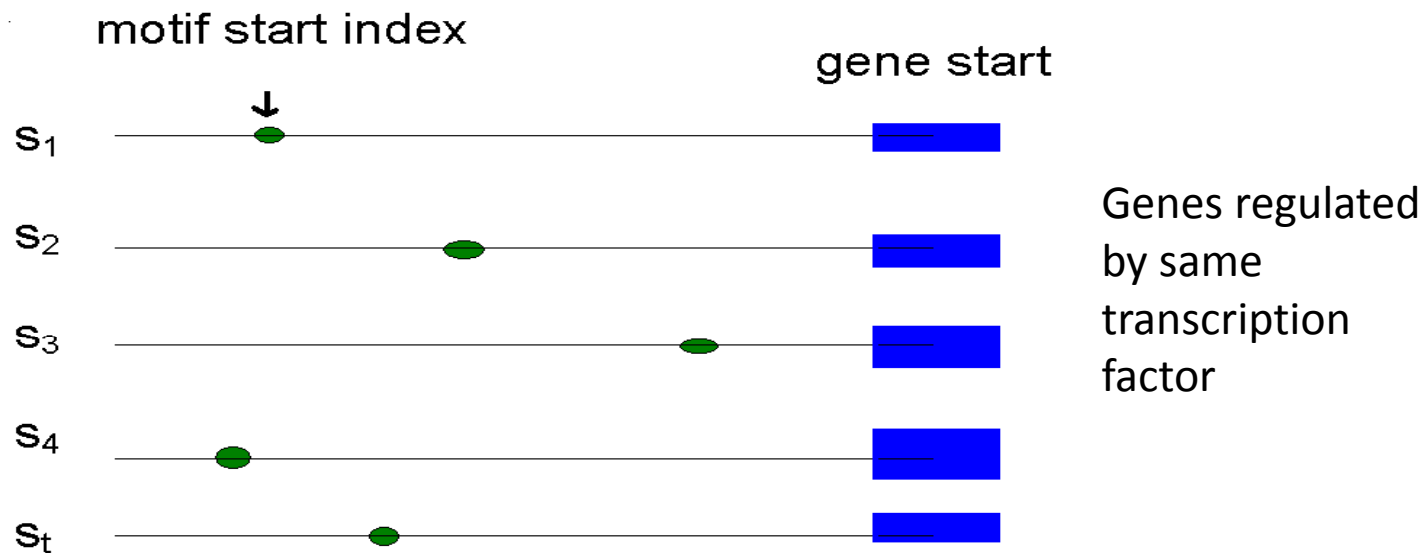
Experimental Method (II)

- ChIP-chip or ChIP-Seq: is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray (or sequencing) technology

http://en.wikipedia.org/wiki/ChIP-chip
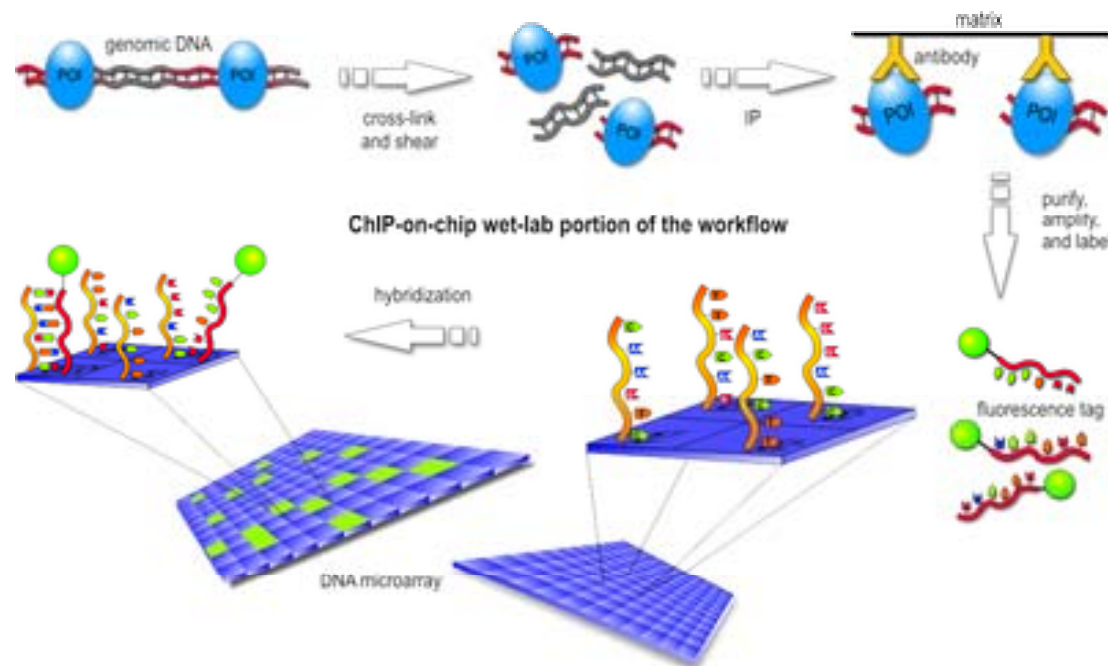
Motif Finding

- Find promoter motifs associated with **co-regulated** or **functionally related** genes

Genes regulated by same transcription factor

Input Sequences

- ChIP-chip experiment.

- Promoter sequences from a cluster of microarray data (or functional related genes)

- Conserved noncoding sequences among different species.

Essential Tasks

- Modeling motifs
- Visualization motifs
- Finding motif

Consensus

HEM13	CCCATTGTTCTC
HEM13	TTTCTGGTTCTC
HEM13	TCAATTGTTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTTG

YCHATTGTTCTC

Probabilistic Model

- Positional weighted matrix (PWM)
  - L x 4 matrix, where L is the length of the motif
  - Each position is a probabilty distribution (p(A), p(C), p(G), P(T))
  - Independence between different position

PWM

HEM13 CCCATT
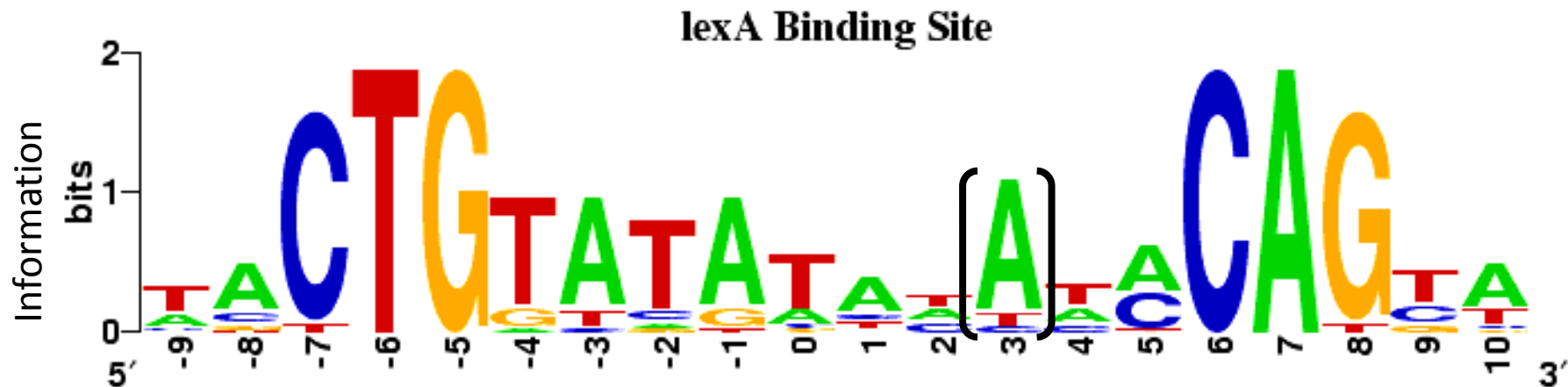HEM13 TTTCTG
HEM13 TCAATT
ANB1 CTCATT
ANB1 TCCATT
ANB1 CCTATT
ANB1 TCCATT
ROX1 CCAATT

	1	2	3	4	5	6
A	0	0	0.25	0.875	0	0
C	0.5	0.75	0.5	0.125	0	0.125
G	0	0	0	0	0	0.875
T	0.5	0.25	0.25	0	1.0	0

Motif Information

The height of a stack is often called the motif information at that position measured in bits

lexA Binding Site

$$\text{Motif Position Information} = 2 - \sum_{b=\{A,T,G,C\}} -p_b \log p_b$$

Why is this a measure of information?

随机事件的信息量 (I)

- 如果说"明天的太阳会从东边升起"，你会觉得这是一句废话，因为没有得到任何信息。

- 反过来，如果说"明天会发生日食"，你会觉得很吃惊，感觉到得到了很多信息。

- 因此，信息量的多少与随机事件发生的概率有关，是概率的函数 f(p).

随机事件的信息量 (II)

- 相互独立的两个随机事件同时发生引起的信息量是分别引起的信息量之和。

$$f(pq) = f(p) + f(q)$$

- 什么函数具有上述性质？可以证明，唯有对数函数具有上述性质。

$$I(p) = -\log_2(p)$$

随机分布的信息量

- 定义为每个可能的随机事件的平均信息量。

- 若离散分布S有n个取值，p_i是相应取值的概率。则分布S的熵定义为

$$\text{Inf}(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$
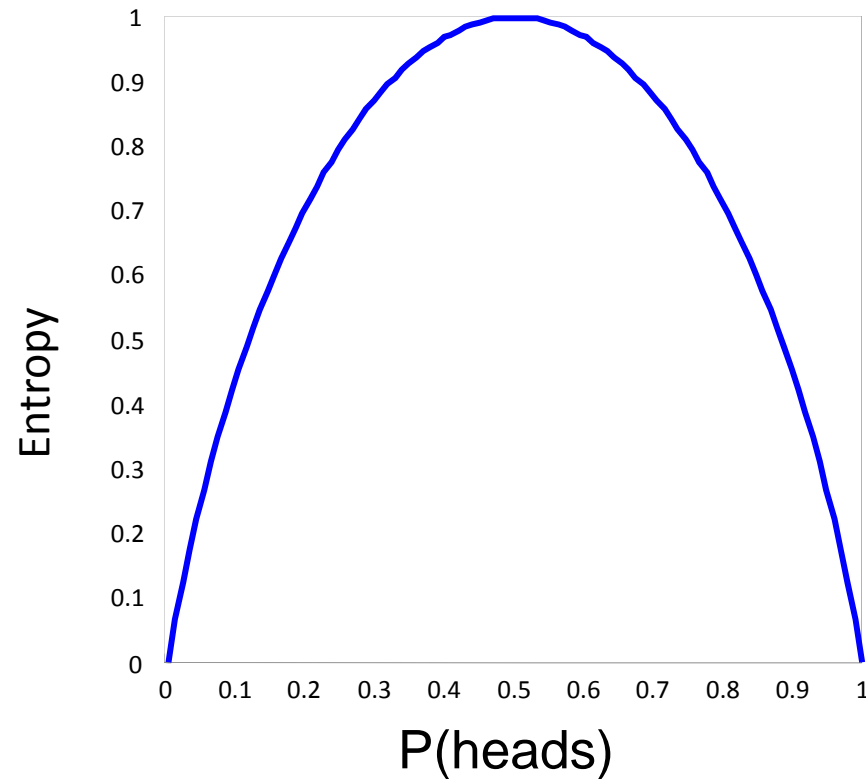
Entropy

Entropy measures average uncertainty

Entropy measures randomness

$$H(X) = -\sum_i p_i \log_2 p_i$$

If log is base 2, then the units are called bits

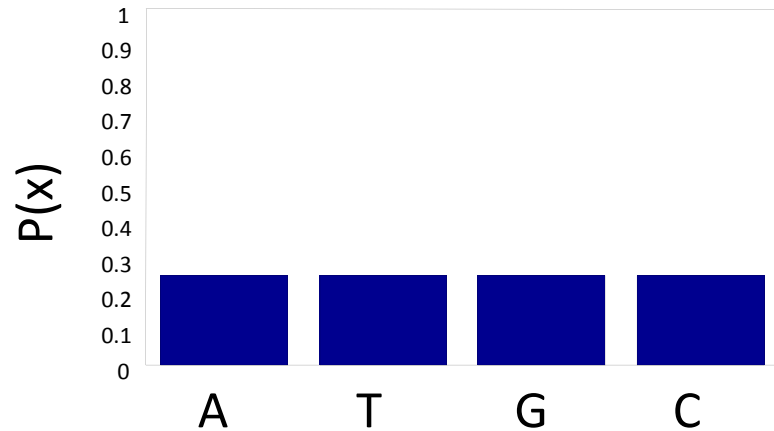# Entropy versus Randomness

Entropy is maximum at maximum randomness

Example: Coin Toss

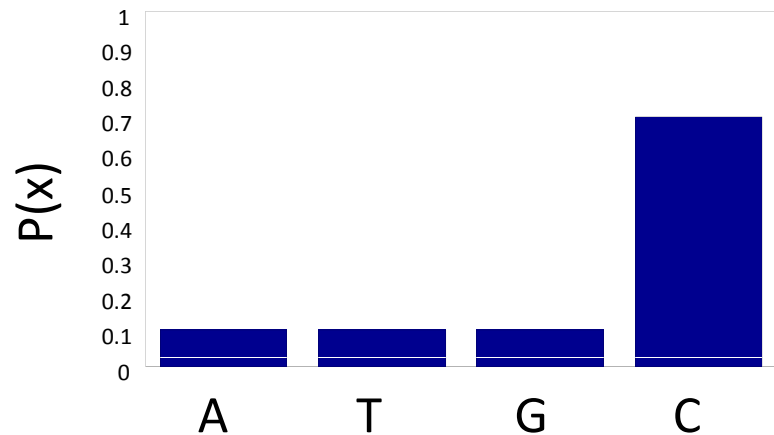P(heads)=0.1 Not very random
H(X)=0.47 bits

P(heads)=0.5 Completely random
H(X)=1 bits

Entropy Examples

$$H(X) = -[4 * 0.25 \log_2(0.25)]$$
$$= 2(bit)$$

$$H(X) = -[3 * 0.1 \log_2(0.1)$$
$$+ 0.7 \log_2(0.7)]$$
$$= 0.63(bit)$$

Motif Information

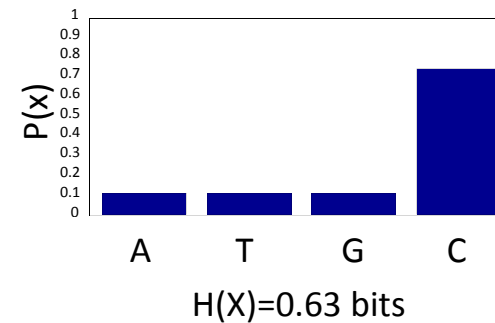Motif Position Information = $2.0 - \sum_{b=\{A,C,G,T\}} p_b \log_2 p_b$

$H_{background}(X)$
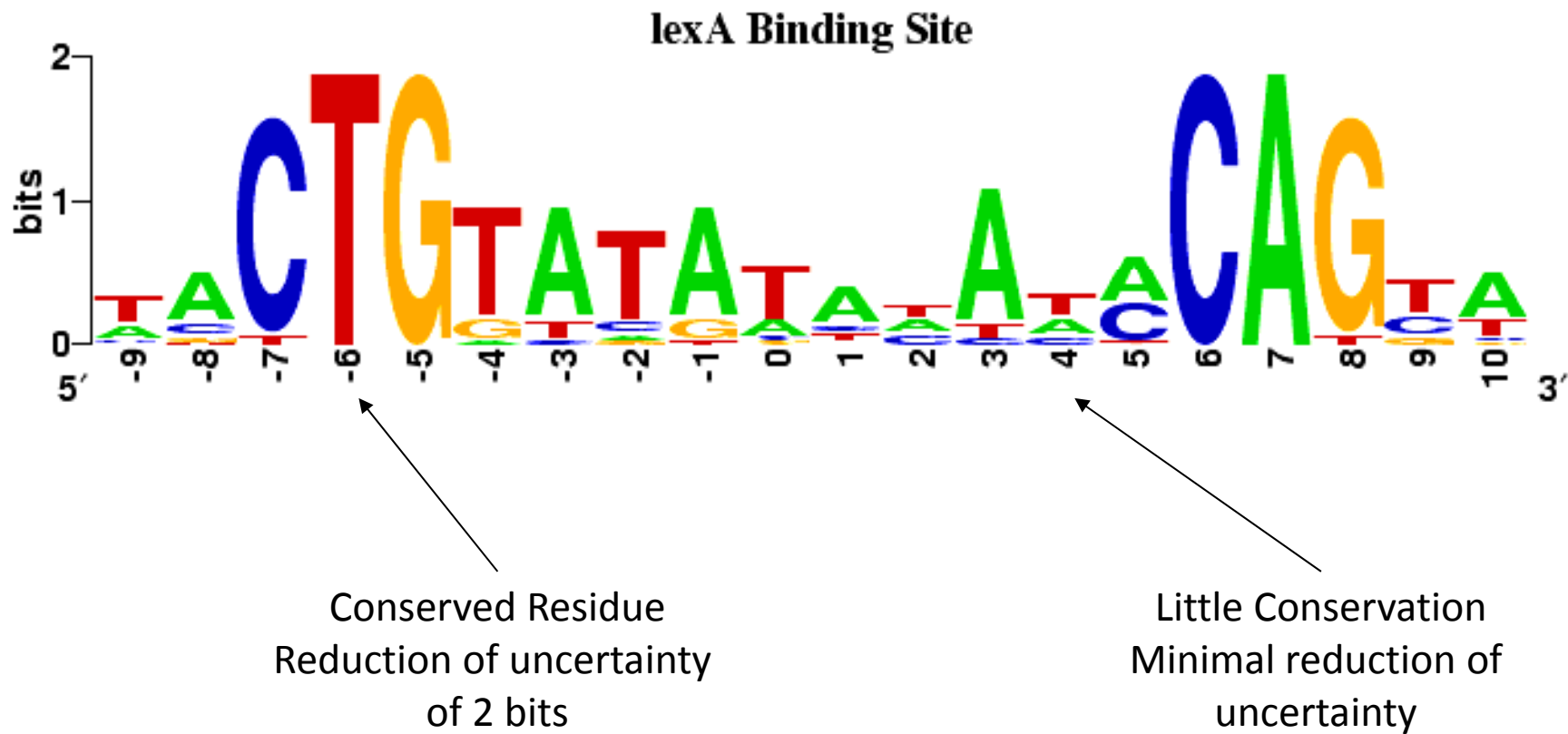
$H_{motif\_i}(X)$

Prior uncertainty about nucleotide
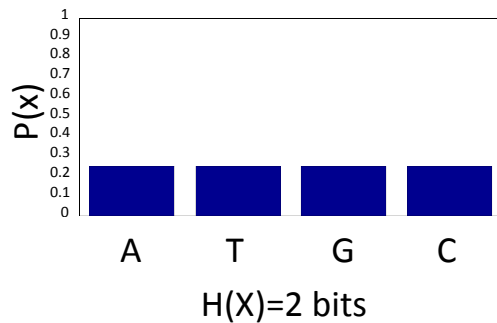
Uncertainty after learning it is position i in a motif

H(X)=2 bits

H(X)=0.63 bits

Uncertainty at this position has been reduced by 1.37 bits

Motif Logo

lexA Binding Site

Conserved Residue
Reduction of uncertainty
of 2 bits

Little Conservation
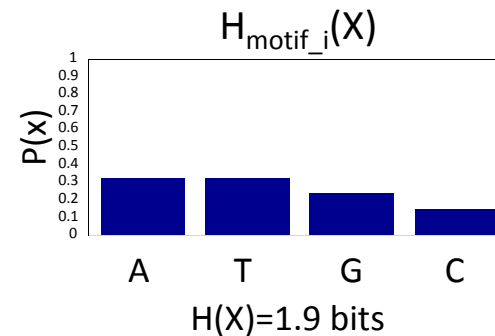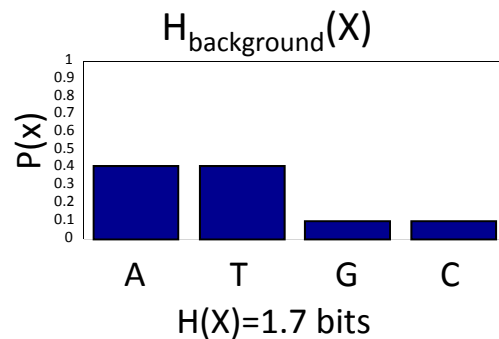Minimal reduction of
uncertainty

Background DNA Frequency

The definition of information assumes a uniform background DNA nucleotide frequency

What if the background frequency is not uniform?

(e.g. Plasmodium)

$H_{background}(X)$

H(X)=1.7 bits

$H_{motif\_i}(X)$

H(X)=1.9 bits

Motif Position Information = $1.7 - \sum_{b=\{A,C,G,T\}} p_b \log p_b = 0.2 (\text{bit})$

Some motifs could have negative information!

A Different Measure

- Relative entropy or Kullback-Leibler distance (divergence)

$$D_{KL}(P_{motif}||P_{bg}) = \sum_{b=\{A,C,G,T\}} P_{motif}(b) \log \frac{P_{motif}(b)}{P_{bg}(b)}$$
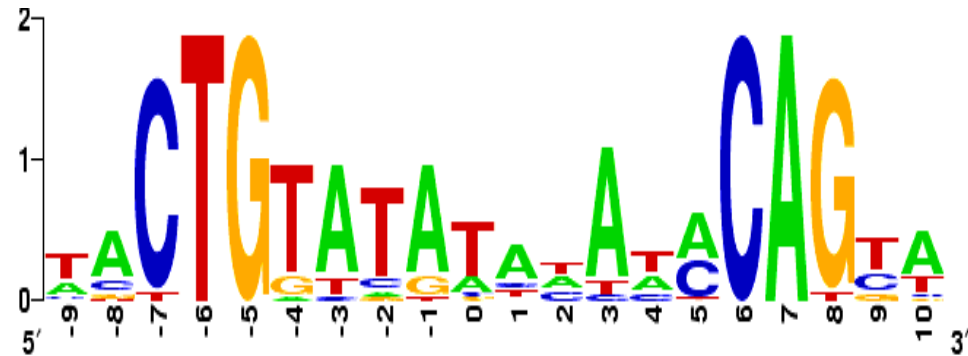
- Property

$$D_{KL} \geq 0$$
$$D_{KL} = 0 \Leftrightarrow P_{motif} = P_{bg}$$

Comparing Both Methods

Information assuming uniform background DNA

KL Distance assuming 20% GC content (e.g. Plasmodium)

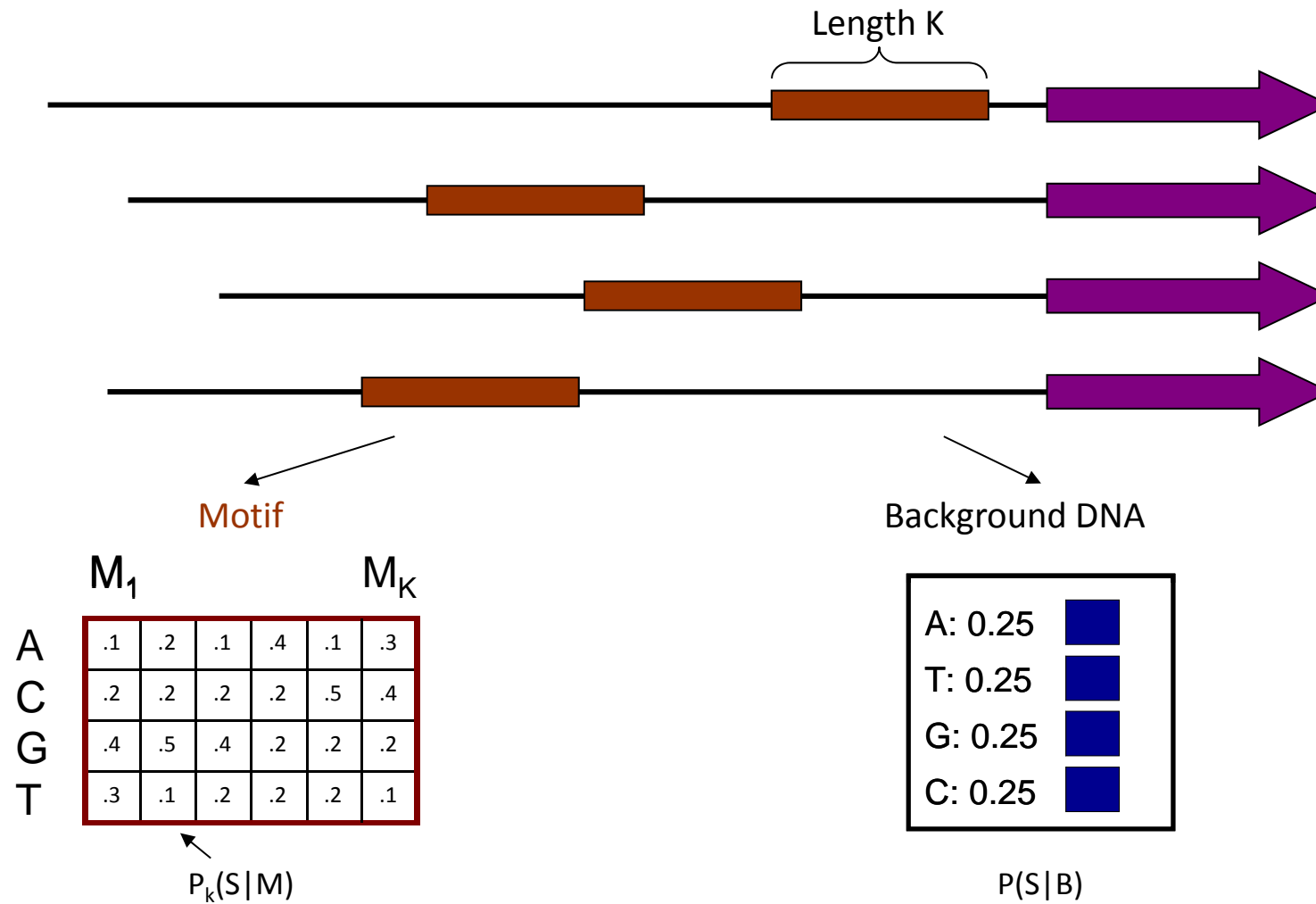# Finding New Motifs

Learning Motif Models

Motif Finding Problem

- Given a set of sequences, find the motif shared by all or most sequences, while its starting position in each sequence is unknown

- Assumption:
  - Each motif appears exactly once in one sequence
  - The motif has fixed length

Generative Model

- Suppose the sequences are aligned, the aligned regions are generated from a motif model

- Motif model is a PWM. A PWM is a position-specific multinomial distribution.
  - For each position i, a multinomial distribution on (A,C,G,T): $q_{iA}, q_{iC}, q_{iG}, q_{iT}$

- The unaligned regions are generated from a background model: p_A, p_C, p_G, p_T
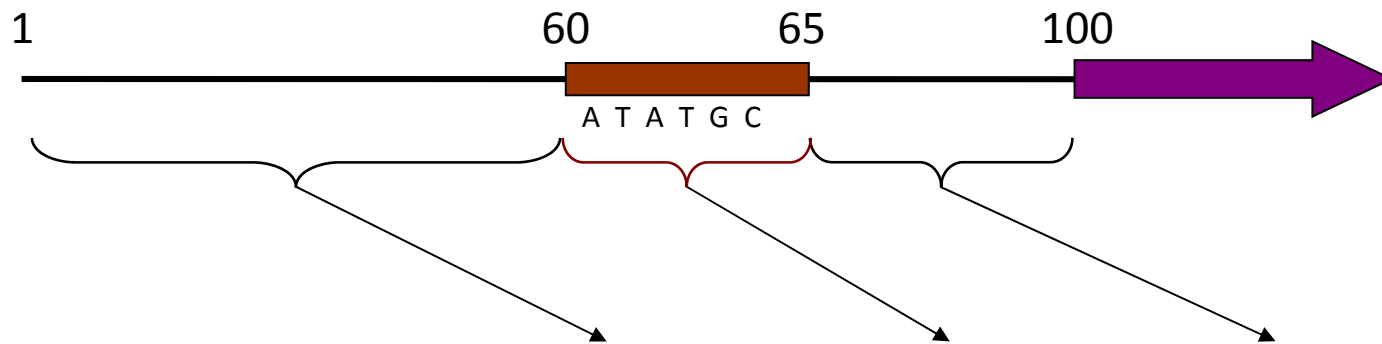
A Promoter Model

Length K

Motif

Background DNA

M_1 M_K

A	.1	.2	.1	.4	.1	.3
C	.2	.2	.2	.2	.5	.4
G	.4	.5	.4	.2	.2	.2
T	.3	.1	.2	.2	.2	.1

$P_k(S|M)$

A: 0.25

T: 0.25

G: 0.25

C: 0.25

$P(S|B)$

The same motif model in all promoters

Probability of a Sequence

Given a sequence(s), motif *model* and motif *location*

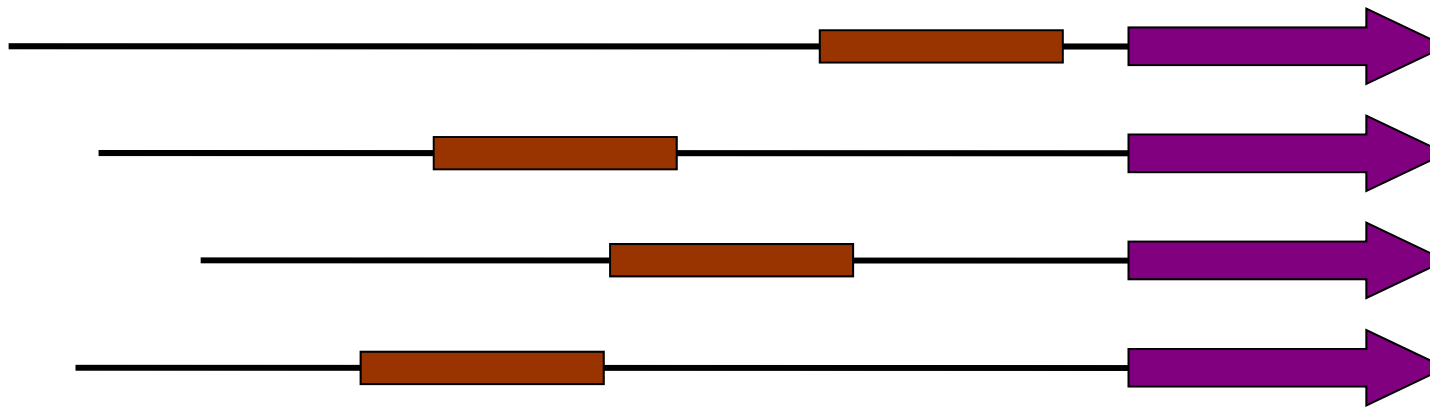$$P(Seq \mid Mstart = 10, Model) = \prod_{i=1}^{59} P(S_i \mid B) \prod_{k=1}^{6} P_k\left(S_{k+63} \mid M\right) \prod_{i=66}^{100} P(S_i \mid B)$$

S_i = nucleotide at position i in the sequence

	M₁					M_K
A	.1	.2	.1	.4	.1	.3
C	.2	.2	.2	.2	.5	.4
G	.4	.5	.4	.2	.2	.2
T	.3	.1	.2	.2	.2	.1

Parameterizing the Motif Model

Given multiple sequences and motif locations but no motif *model*

AATGCG
ATATGG
ATATCG
GATGCA

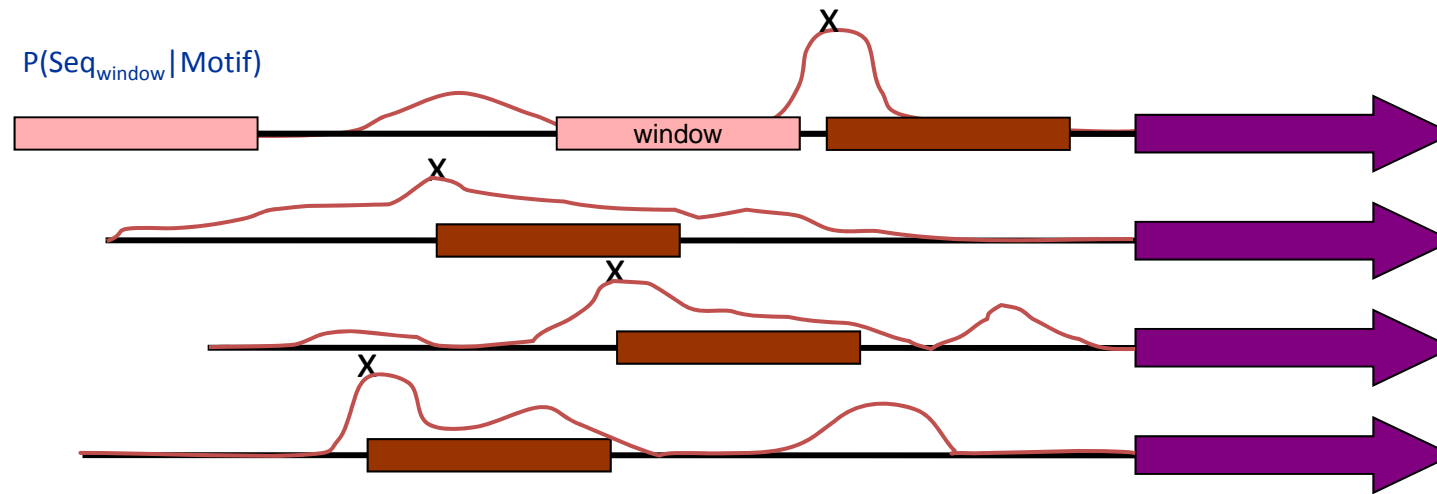Count Frequencies

Add pseudocounts

M_1 M_6

A	3/4	1/2	1/2			1/4
C					1/2	
G	1/4			1/2	1/2	3/4
T		1/2	1/2	1/2		

Finding Known Motifs

Given multiple sequences and motif model but no motif *locations*

Calculate $P(Seq_{window} | Motif)$ for every starting location

Choose best starting location in each sequence

The EM Approach

- EM is a family of algorithms for learning probabilistic models in problems that involve *hidden state*

- in our problem, the hidden state is where the motif starts in each training sequence
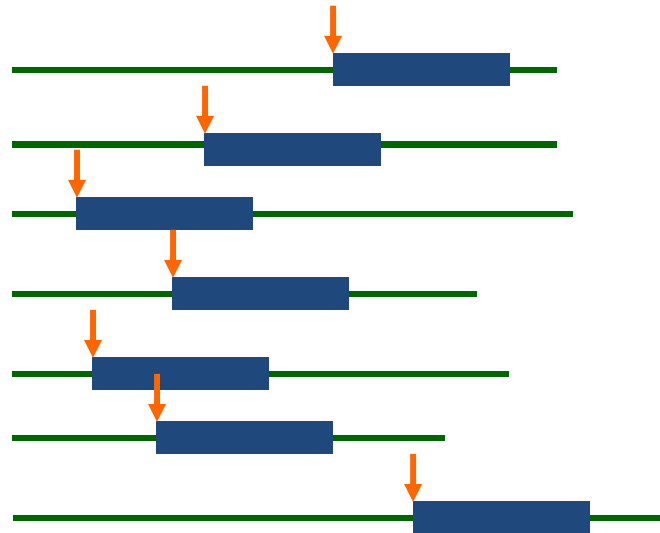
The MEME Algorithm

- Bailey & Elkan, 1993
- uses EM algorithm to find multiple motifs in a set of sequences
- first EM approach to motif discovery: Lawrence & Reilly 1990
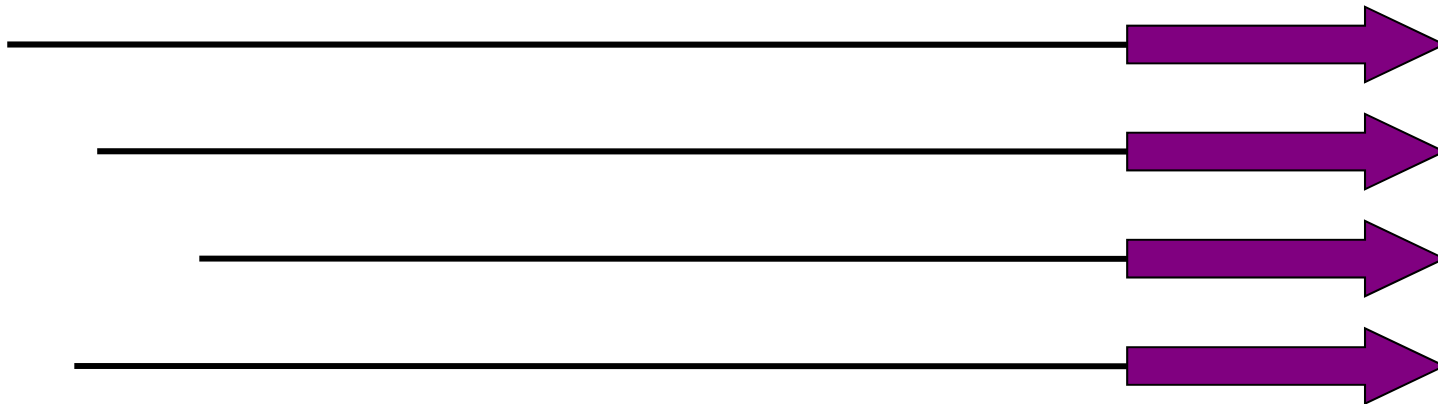
EM Algorithm for Motif Discovery

1. Start with random motif model

2. E Step: estimate probability of motif positions for each sequence

3. M Step: use estimate to update motif model

4. Iterate (to convergence)

At each iteration, P(Sequences|Model) guaranteed to increase

Demo: Initialization

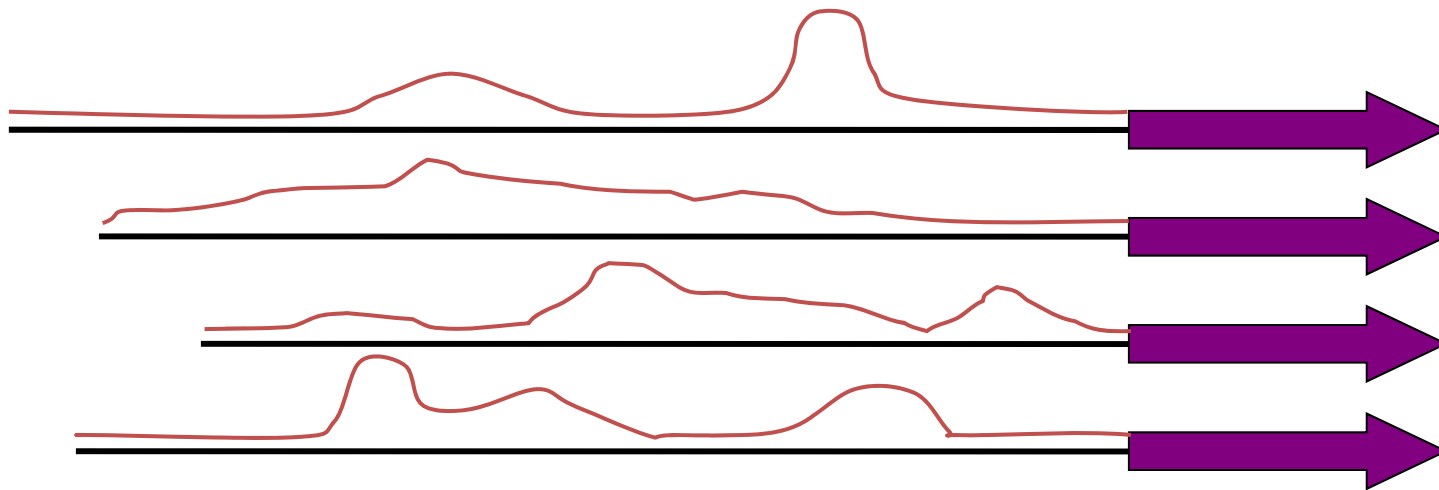- Given a random motif model

	0.1	0.2	0.1	0.4	0.1	0.3
A	0.1	0.2	0.1	0.4	0.1	0.3
C	0.2	0.2	0.2	0.2	0.5	0.4
G	0.4	0.5	0.4	0.2	0.2	0.2
T	0.3	0.1	0.2	0.2	0.2	0.1

Demo: E-Step

- E Step: estimate probability of motif positions for each sequence

Demo: M-Step

- M Step: use estimate to update motif model

A	0.1	0.1	0.1	0.1	0.1	0.3
C	0.2	0.3	0.2	0.2	0.5	0.1
G	0.4	0.5	0.4	0.5	0.2	0.1
T	0.3	0.1	0.2	0.2	0.2	0.1

Basic EM Approach

- we'll need to calculate the probability of a training sequence given a hypothesized starting position:

$$\Pr(X_i \mid Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k,0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k,k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} p_{c_k,0}}_{\text{after motif}}$$

X_i is the *i*th sequence

Z_{ij} is 1 if motif starts at position *j* in sequence *i*

c_k is the character at position *k* in sequence *i*

Example

$$X_i = \text{G C } \boxed{\text{T G T}} \text{ A G}$$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$p = $

$$\Pr(X_i \mid Z_{i3} = 1, p) =$$

$$p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

The E-step: Estimating Z

- To estimate the starting positions in *Z* at step *t*

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p^{(t)}) \Pr(Z_{ij} = 1)}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, p^{(t)}) \Pr(Z_{ik} = 1)}$$

- This comes from Bayes' rule applied to

$$\Pr(Z_{ij} = 1 \mid X_i, p^{(t)})$$

The E-step: Estimating Z

- Assume that it is equally likely that the motif will start in any position

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p^{(t)}) \cancel{\Pr(Z_{ij} = 1)}}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, p^{(t)}) \cancel{\Pr(Z_{ik} = 1)}}$$

Example: Estimating Z

$X_i =$ G C T G T A G

$$p = \begin{array}{c c c c c} & 0 & 1 & 2 & 3 \\ A & 0.25 & 0.1 & 0.5 & 0.2 \\ C & 0.25 & 0.4 & 0.2 & 0.1 \\ G & 0.25 & 0.3 & 0.1 & 0.6 \\ T & 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$$\vdots$$

- Then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$

The M-step: Estimating p

- recall $p_{c,k}$ represents the probability of character c in position k ; values for position 0 represent the background

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

pseudo-counts

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1}=c\}} Z_{ij} & k > 0 \\ n_c - \sum_{j=1}^{W} n_{c,j} & k = 0 \end{cases}$$

total # of c's in data set

Example: Estimating p

A C A G C A

$Z_{1,1} = 0.1, Z_{1,2} = 0.7, Z_{1,3} = 0.1, Z_{1,4} = 0.1$

A G G C A G

$Z_{2,1} = 0.4, Z_{2,2} = 0.1, Z_{2,3} = 0.1, Z_{2,4} = 0.4$

T C A G T C

$Z_{3,1} = 0.2, Z_{3,2} = 0.6, Z_{3,3} = 0.1, Z_{3,4} = 0.1$

$$p_{A,1} = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \ldots + Z_{3,3} + Z_{3,4} + 4}$$

MEME

- **MEME** - implements EM for motif discovery in DNA and proteins

- **MAST** – search sequences for motifs given a model

- References

1. Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", Nucleic Acids Research, 37:W202-W208, 2009.

http://meme.sdsc.edu/meme/

P(Seq|Model) Landscape

EM searches for parameters to increase P(seqs|parameters)

Useful to think of
P(seqs|parameters)
as a function of parameters

EM starts at an initial set of
parameters 🟢

And then "climbs uphill" until it
reaches a local maximum 🔴
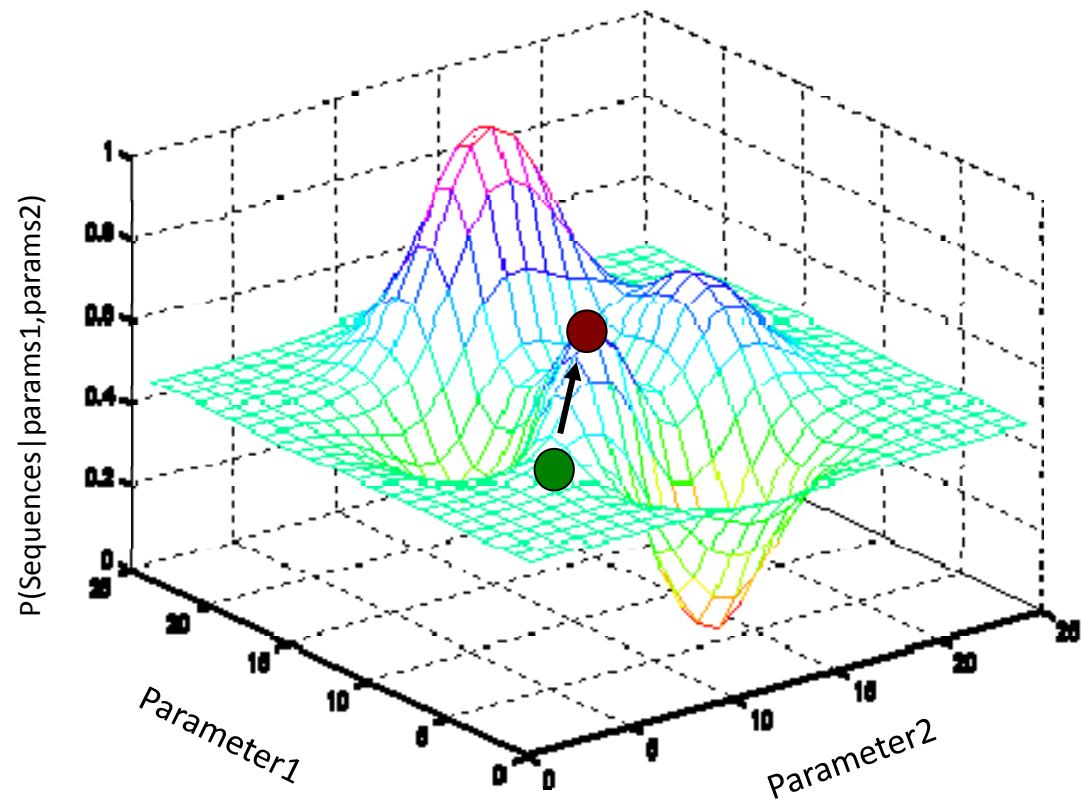
Where EM starts can make a big difference
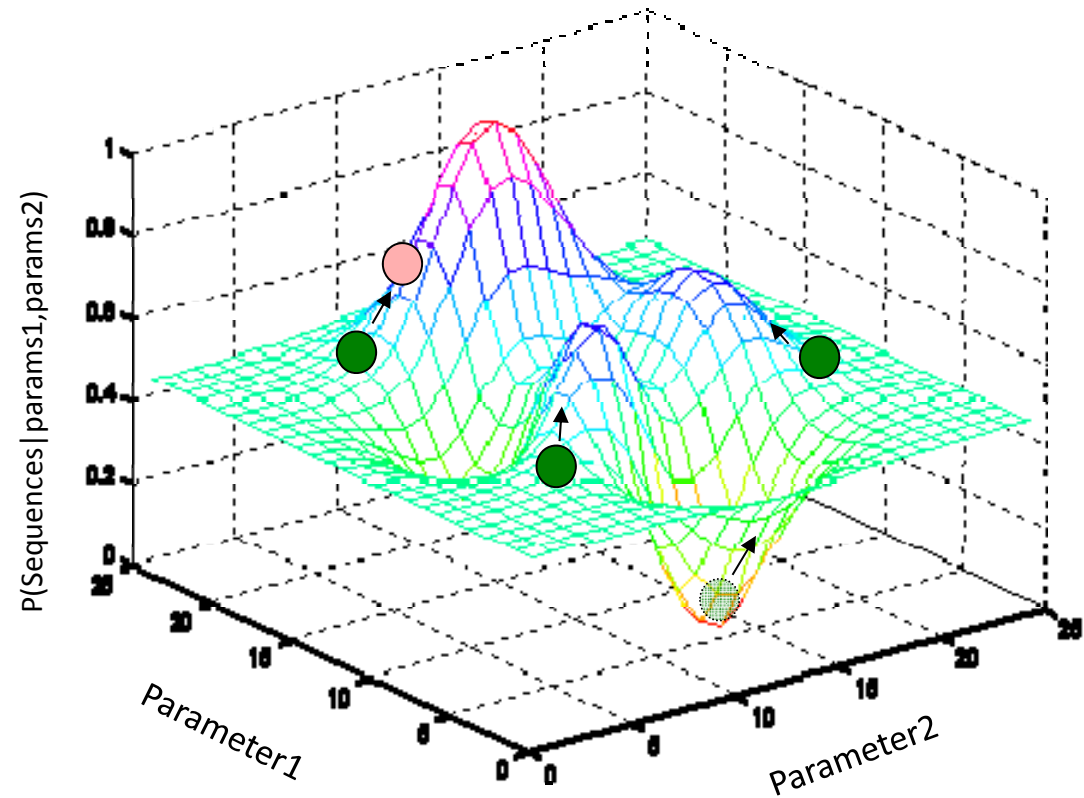
Search from Many Different Starts

To minimize the effects of local maxima, you should search
multiple times from different starting points

MEME uses this idea

Start at many points

Run for one iteration

Choose starting point that got
the "highest" and continue

Gibbs Sampler

- A stochastic version of EM that differs from deterministic EM in two key ways

- At each iteration, we only update the motif position of a single sequence

- We may update a motif position to a "suboptimal" new position

Algorithm: Gibbs Sampler

1. Start with random motif locations and calculate a motif model

2. Randomly select a sequence, remove its motif and recalculate tempory model

3. With temporary model, calculate probability of motif at each position on sequence

4. Select new position based on this distribution

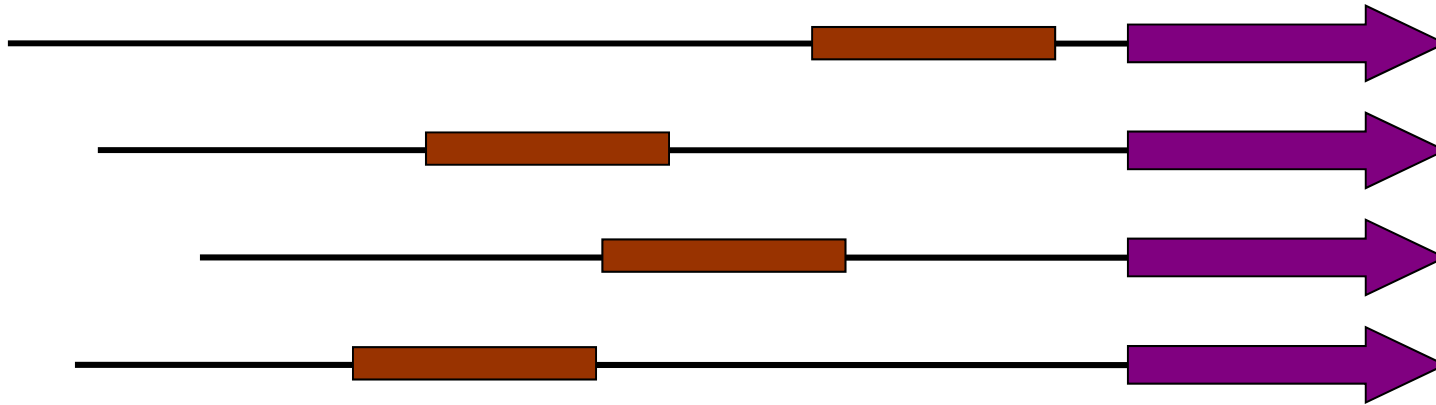5. Update model and Iterate

Sampling New Motif Positions

- For each possible starting position, $a_i = j$, compute a weight (likelihood ratio)

$$A_j = \frac{\displaystyle\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}{\displaystyle\prod_{k=j}^{j+W-1} p_{c_k, 0}}$$

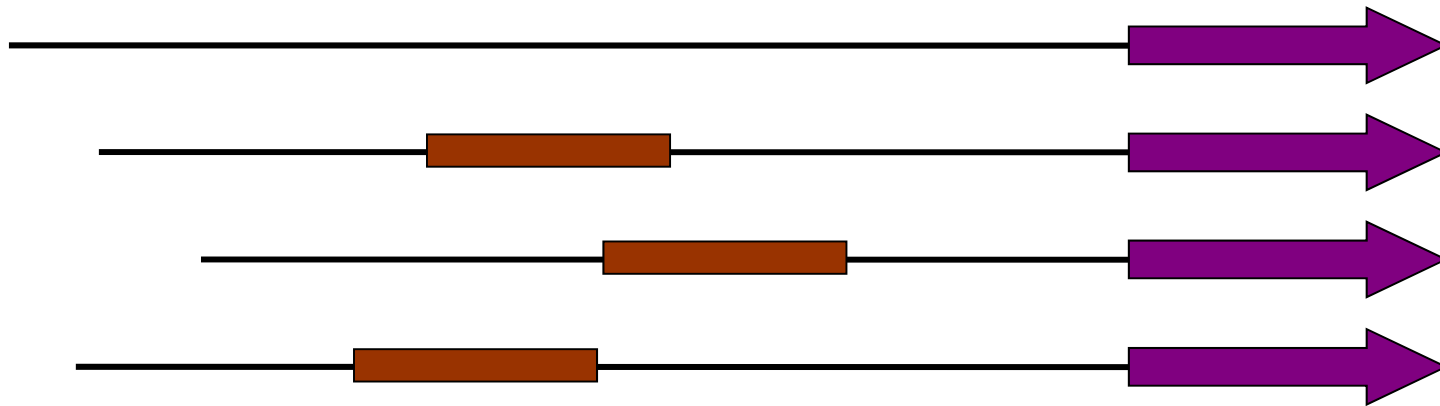- Randomly select a new starting position a_i according to these weights

Demo: Initialization
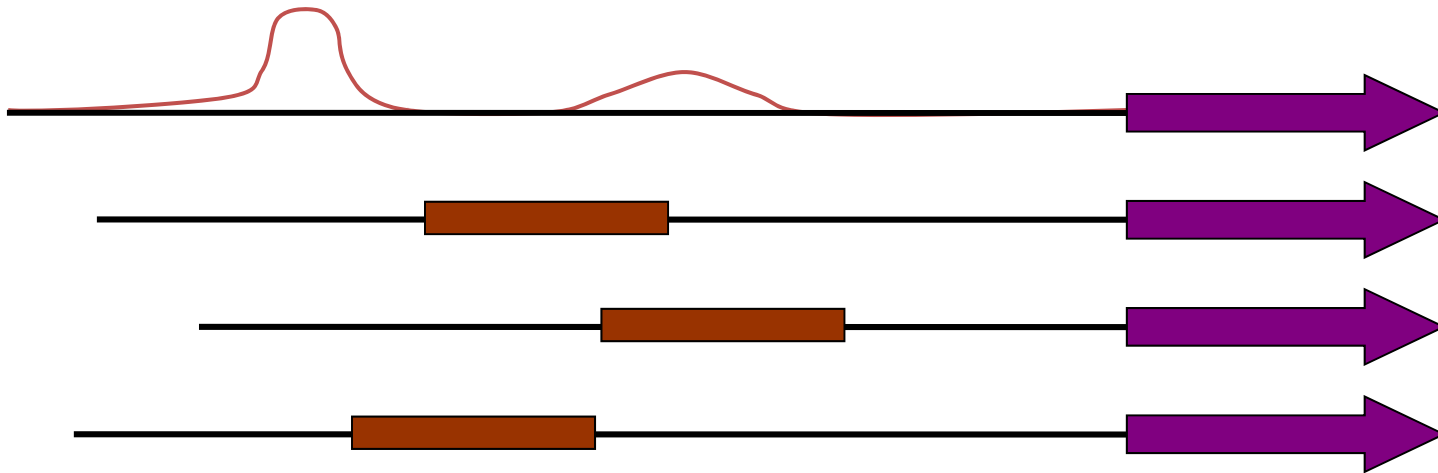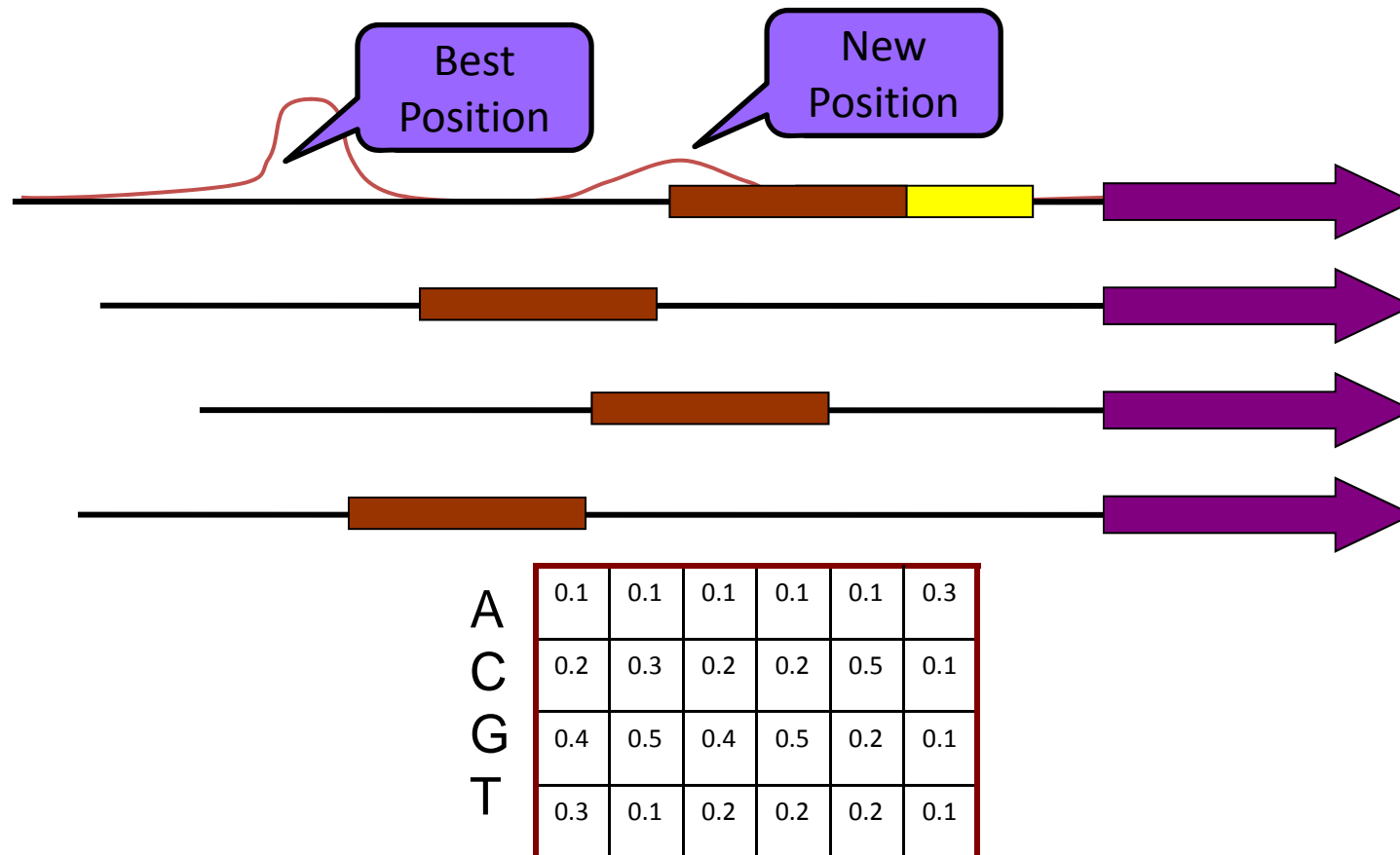
- Random choose motif location

Demo: Step 3

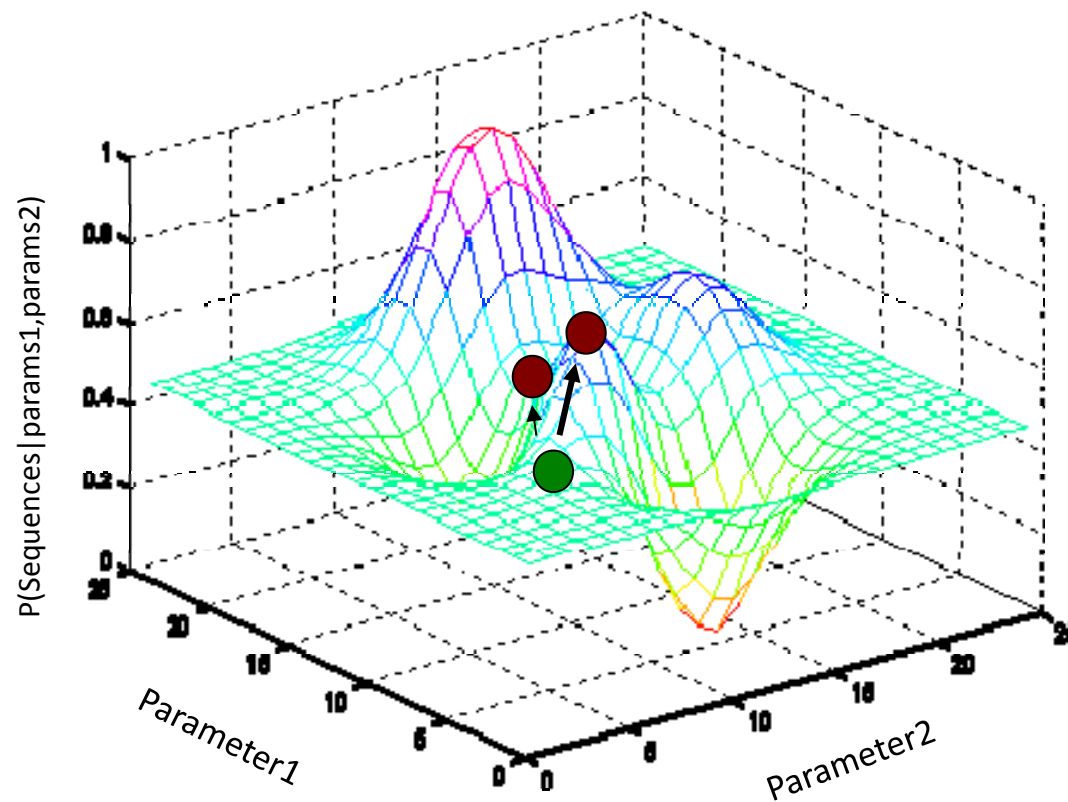- Calculate Probability of motif at each position on sequence

Demo: Step 4

- Demo: Select New Position, Update Motif Model

Gibbs Sampling and Climbing

Because gibbs sampling does always choose the best new location
it can move to another place not directly uphill

In theory, Gibbs Sampling less likely to get stuck a local maxima

AlignACE

- **Implements Gibbs sampling for motif discovery**

- **ScanAce – look for motifs in a sequence given a model**

- **CompareAce – calculate "similarity" between two motifs (i.e. for clustering motifs)**

Reference
1. Roth, F.R., Hughes, J.D., Estep, P. E. & G.M. Church. Finding DNA Regulatory Motifs within Unaligned Non-Coding Sequences Clustered by Whole-Genome mRNA Quantitation. *Nature Biotechnology* 16, 939 - 945 (1998)
2. Hughes, JD, Estep, PW, Tavazoie S & GM Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae, *Journal of Molecular Biology* 2000 Mar 10;296(5):1205-14.

http://arep.med.harvard.edu/mrnadata/mrnasoft.html

References

- Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting Subtle Sequence Signals: a Gibbs Sampling Strategy for Multiple Alignment. Science 1993 Oct 8;262(5131):208-14.