第5-1章 Motif Finding

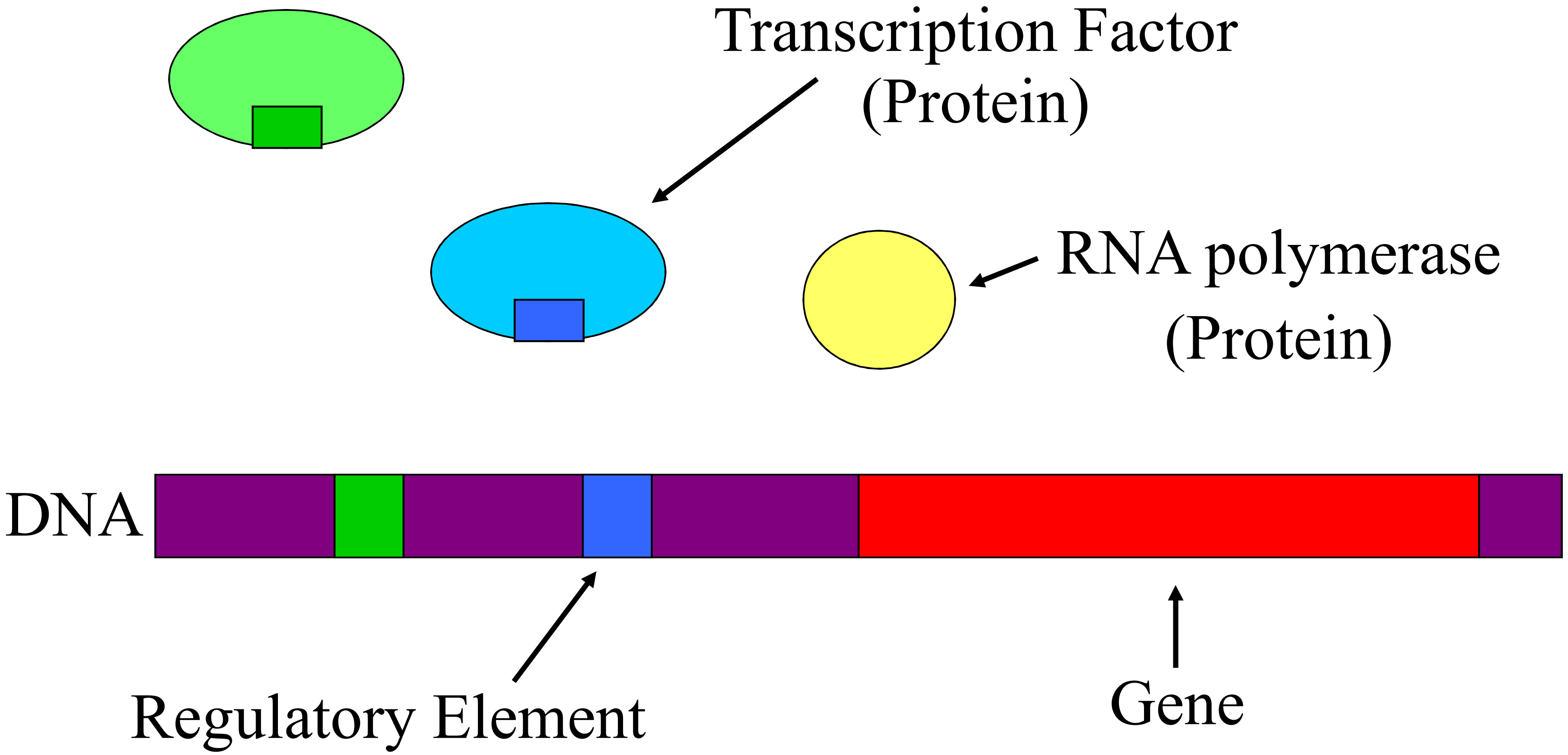
- Motif finding problem
- EM algorithm
- Markov chain Monte Carlo (Gibbs Sampler)
- Deep learning

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

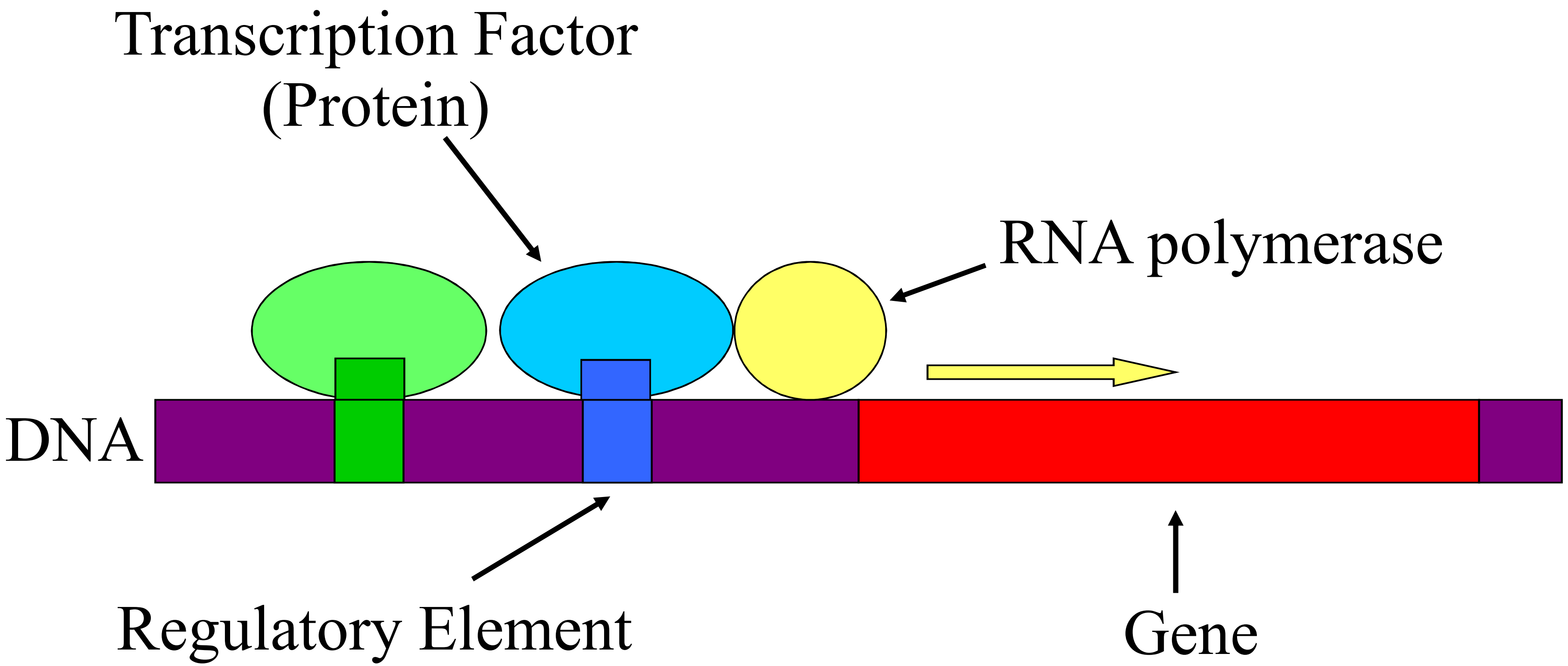
- The transcription of each gene is controlled by a regulatory region of DNA relatively near the transcription start site (TSS).
- two types of fundamental components
  - short DNA regulatory elements
  - *gene regulatory proteins* that recognize and bind to them.

Regulation of Genes



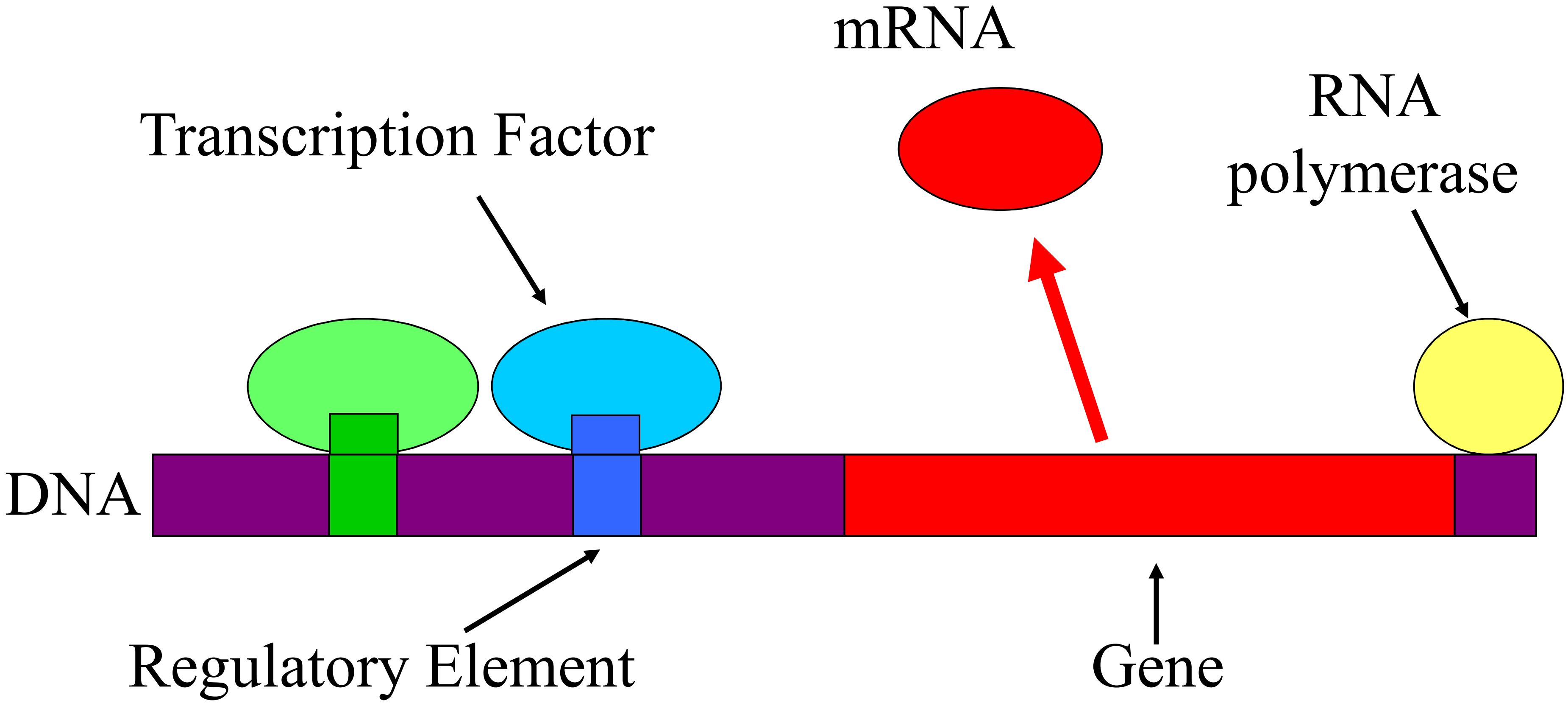
source: M. Tompa, U. of Washington

Regulation of Genes



source: M. Tompa, U. of Washington

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

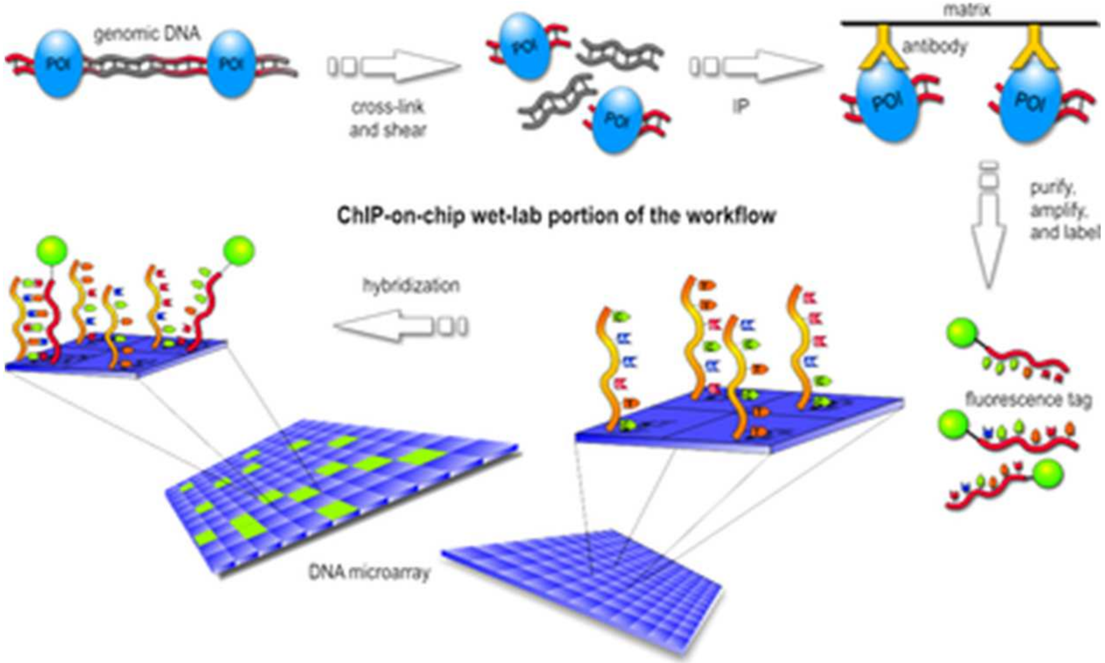
- Small (6-20bp)
- Highly variable

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)

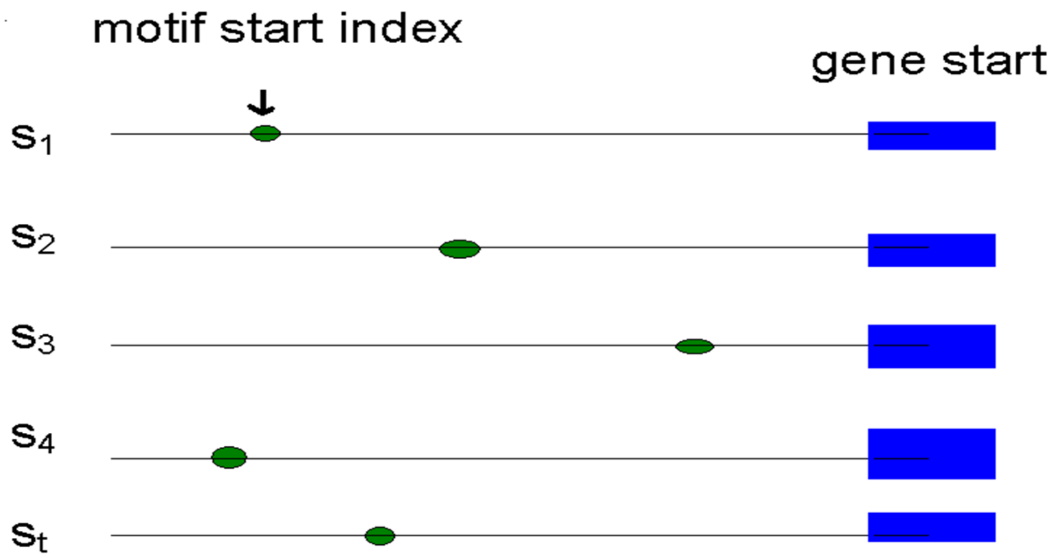
- ChIP-chip or ChIP-Seq: is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray (or sequencing) technology



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

Nature Biotechnology 24, 423 - 425 (2006)

Probabilistic Model

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

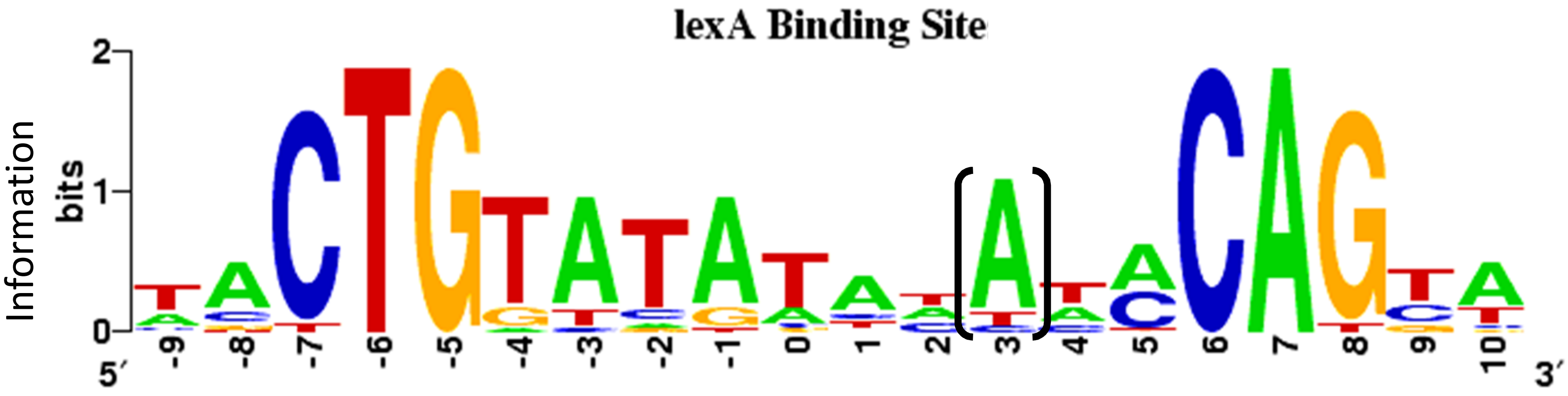
PWM

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



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

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

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

Entropy

Entropy measures average uncertainty

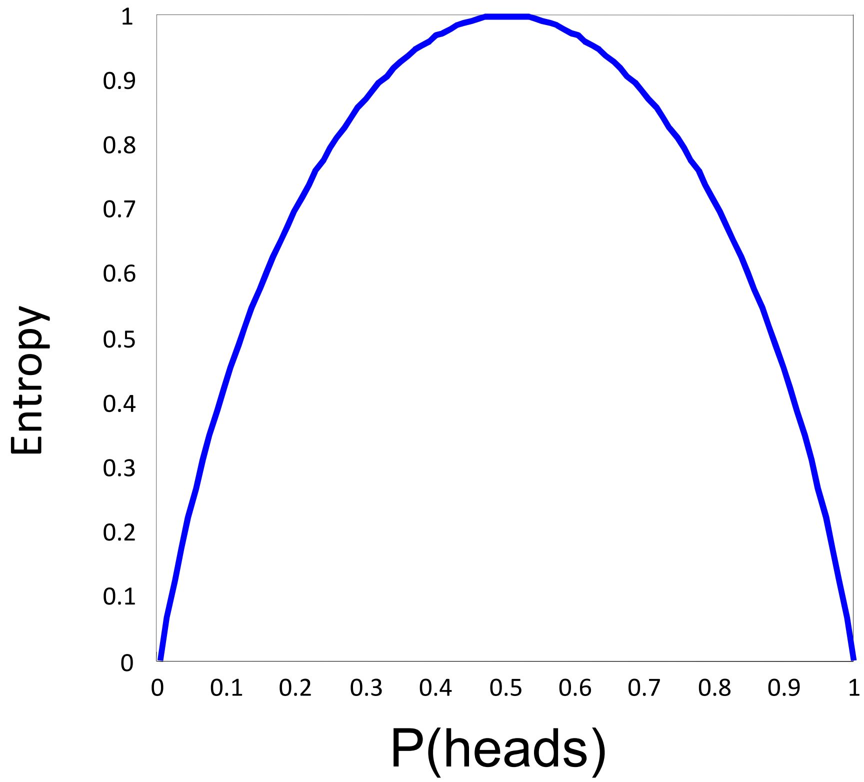
Entropy measures randomness

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

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

Entropy versus Randomness

Entropy is maximum at maximum randomness

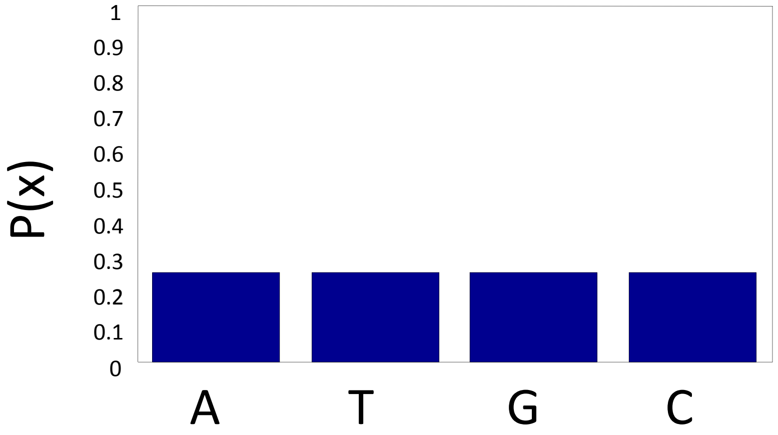


Example: Coin Toss

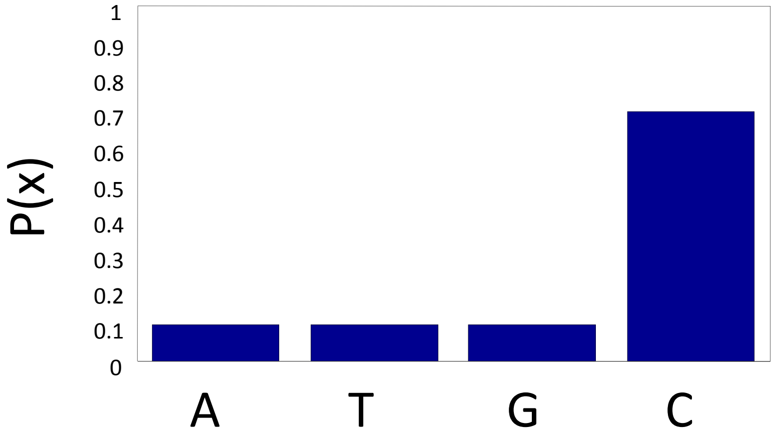
P(heads)=0.1 Not very random
H(X)=0.47 bits

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

Entropy Examples



$$\begin{aligned} H(X) &= -[4 * 0.25 \log_2(0.25)] \\ &= 2(bit) \end{aligned}$$



$$\begin{aligned} H(X) &= -[3 * 0.1 \log_2(0.1) \\ &\quad + 0.7 \log_2(0.7)] \\ &= 0.63(bit) \end{aligned}$$

Motif Information

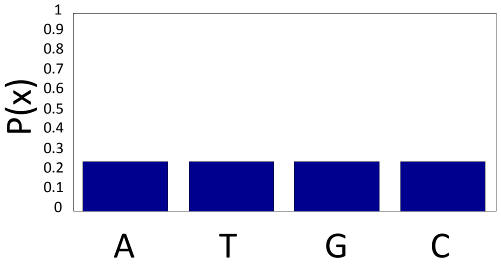
Motif Position Information = $2.0 - \sum_{b=\{A,C,G,T\}} p_b \log_2 p_b$

$H_{background}(X)$

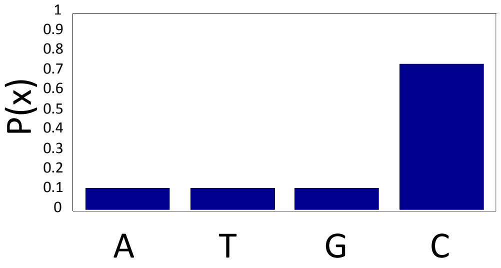
$H_{motif\_i}(X)$

Prior uncertainty about nucleotide

Uncertainty after learning it is position i in a motif



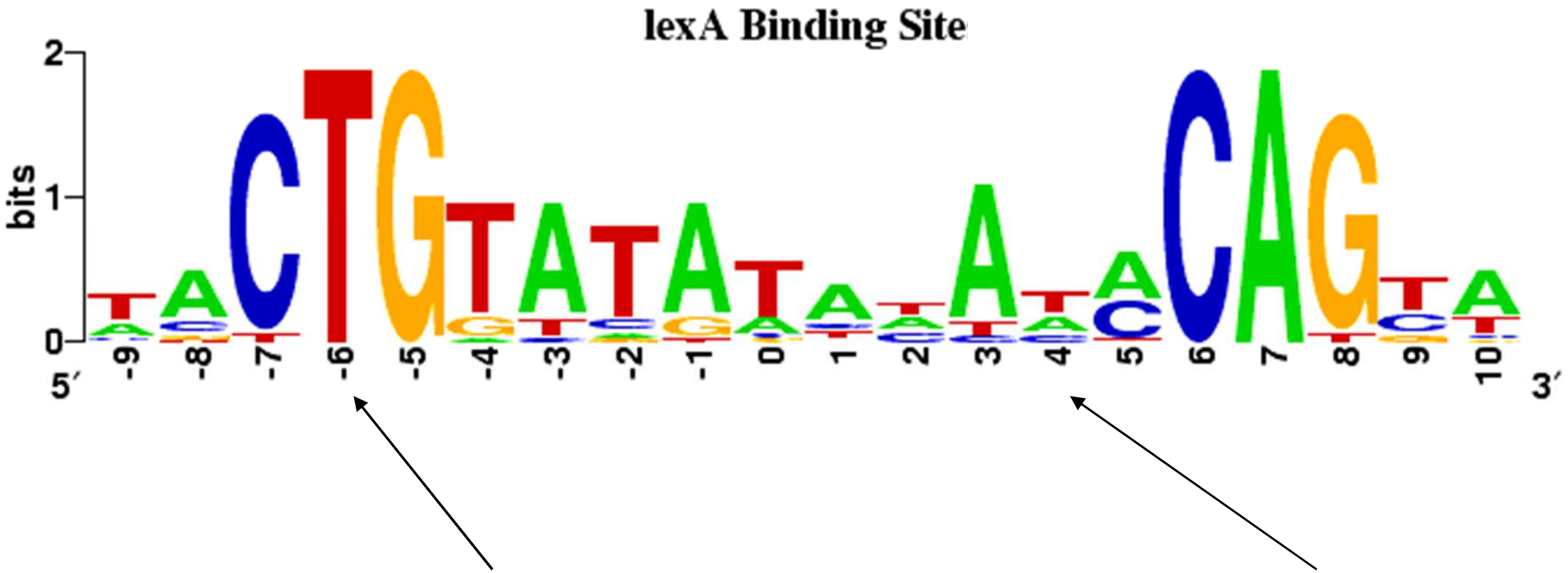
H(X)=2 bits



H(X)=0.63 bits

Uncertainty at this position has been reduced by 1.37 bits

Motif Logo



Conserved Residue
Reduction of uncertainty
of 2 bits

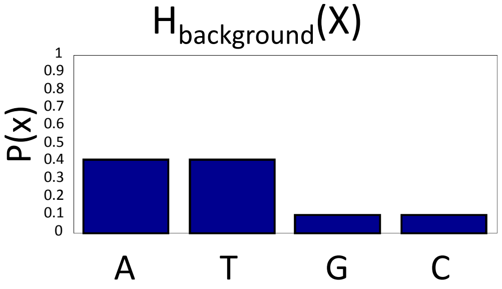
Little Conservation
Minimal reduction of
uncertainty

Background DNA Frequency

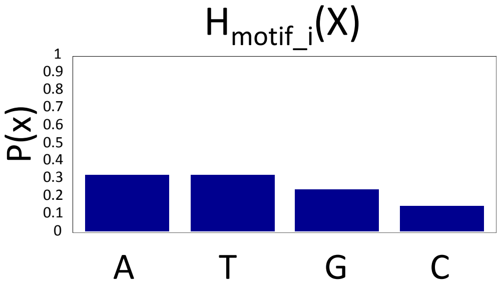
The definition of information assumes a uniform background DNA nucleotide frequency

What if the background frequency is not uniform?

(e.g. Plasmodium)



H(X)=1.7 bits



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)

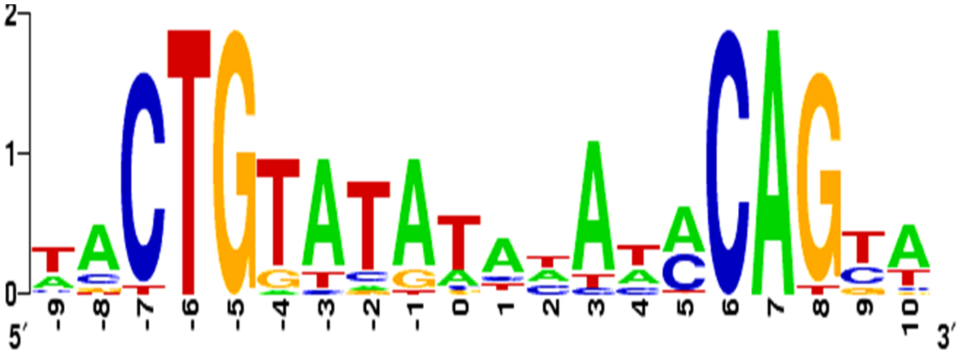
$$D_{KL}(P_{motif}||P_{bg}) = \sum_{b=\{A,C,G,T\}} P_{motif}(b) \log \frac{P_{motif}(b)}{P_{bg}(b)}$$

- Property

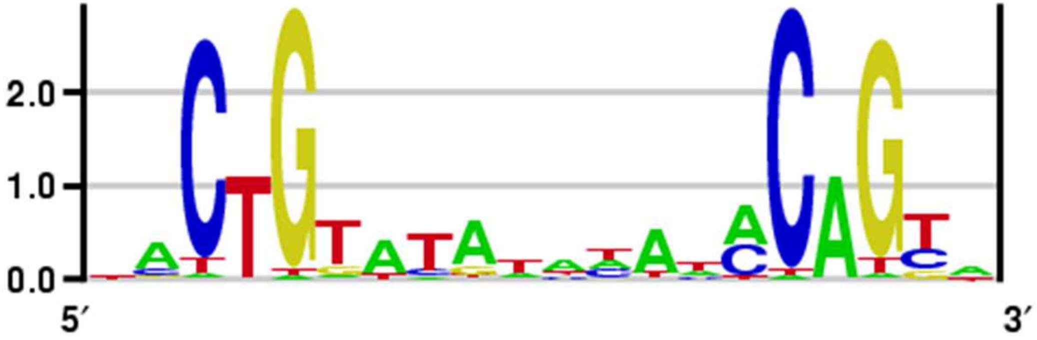
$$D_{KL} \geq 0$$
$$D_{KL} = 0 \Leftrightarrow P_{motif} = P_{bg}$$

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

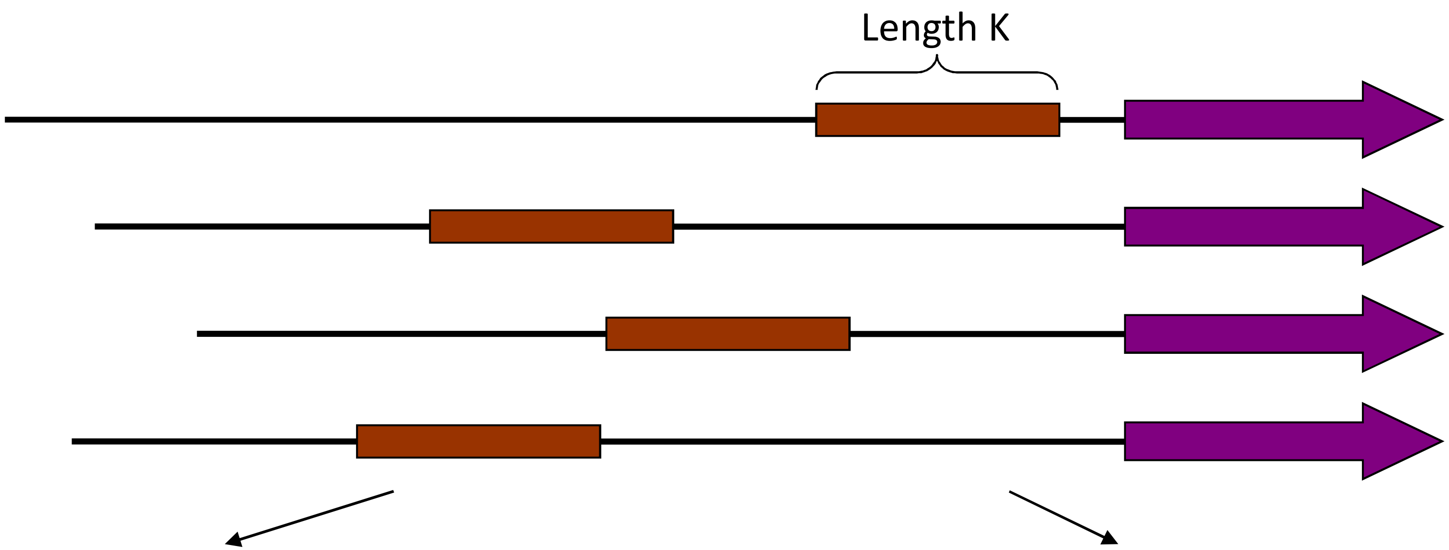
Generative Model

- 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



Motif

	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

$P_k(S|M)$

Background DNA

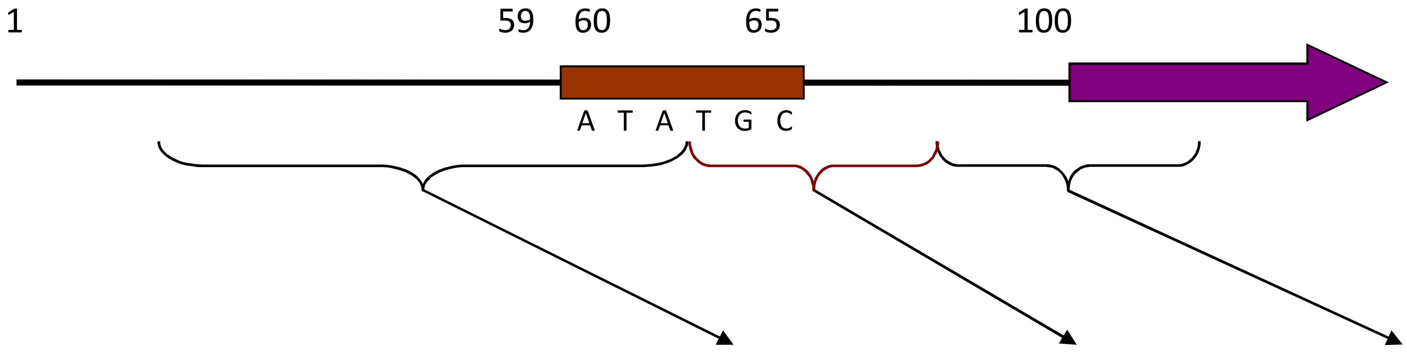
A: 0.25 🟦
T: 0.25 🟦
G: 0.25 🟦
C: 0.25 🟦

$P(S|B)$

The same motif model in all promoters

Probability of a Sequence

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



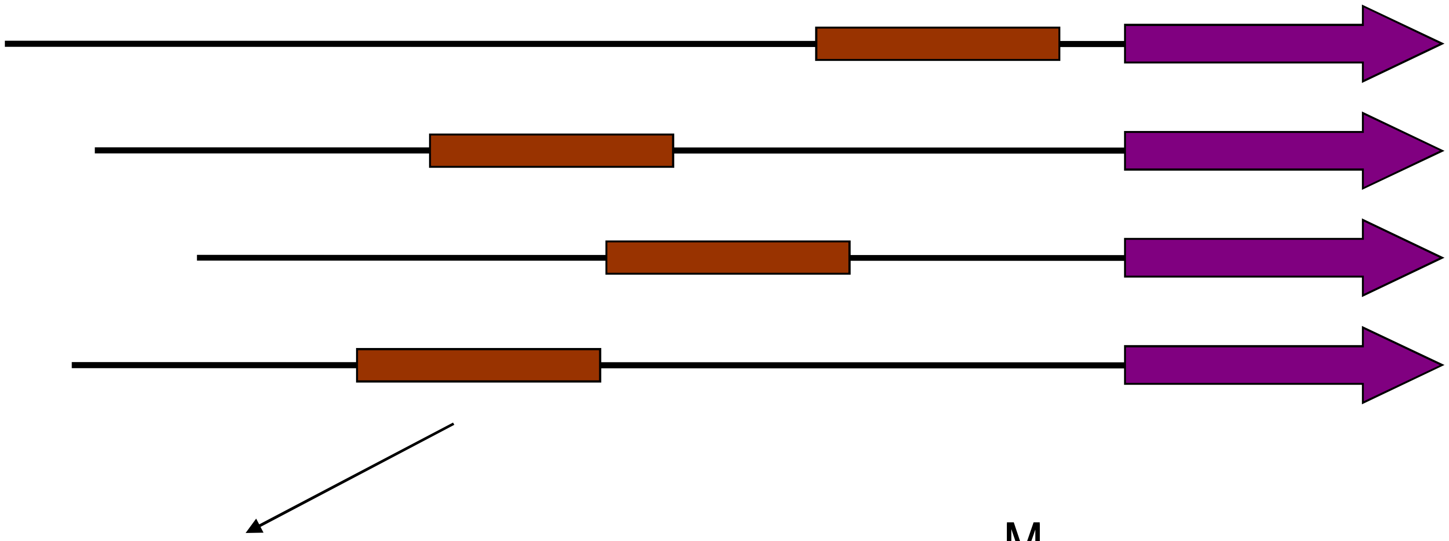
$$Pr(Seq, MS = 60|Model) = \prod_{i=1}^{59} P(S_i|Bg) \prod_{k=1}^{6} P_k(S_{k+59}|M) \prod_{i=66}^{100} P(S_i|Bg)$$

S_i = nucleotide at position i in the sequence

	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

Parameterizing the Motif Model

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



```
AATGCG
ATATGG
ATATCG
GATGCA
```

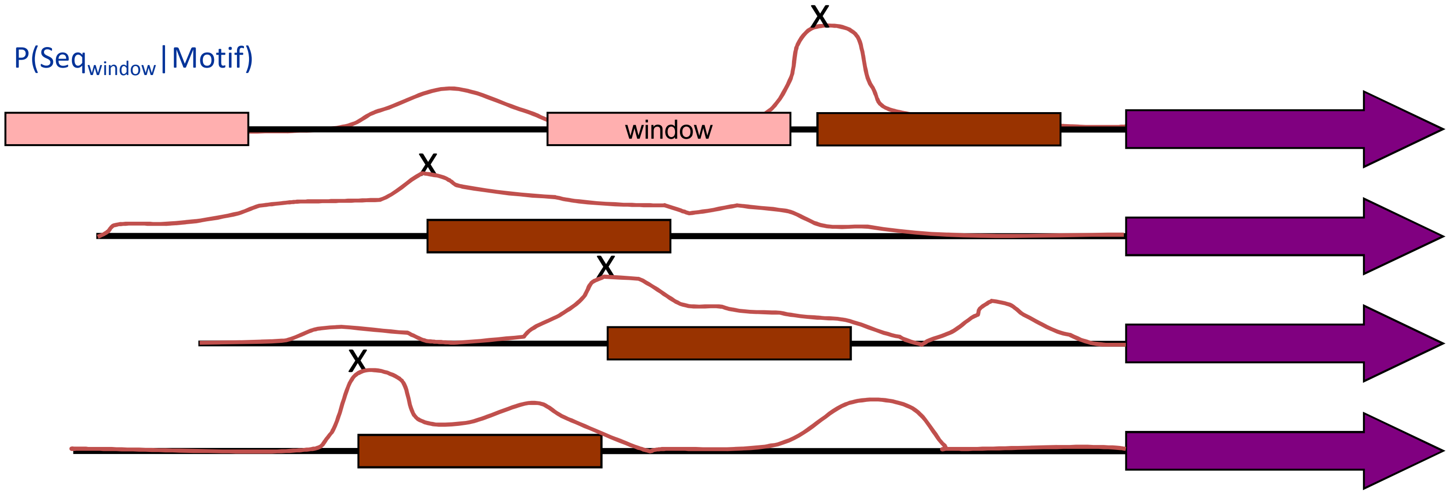
Count Frequencies

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*

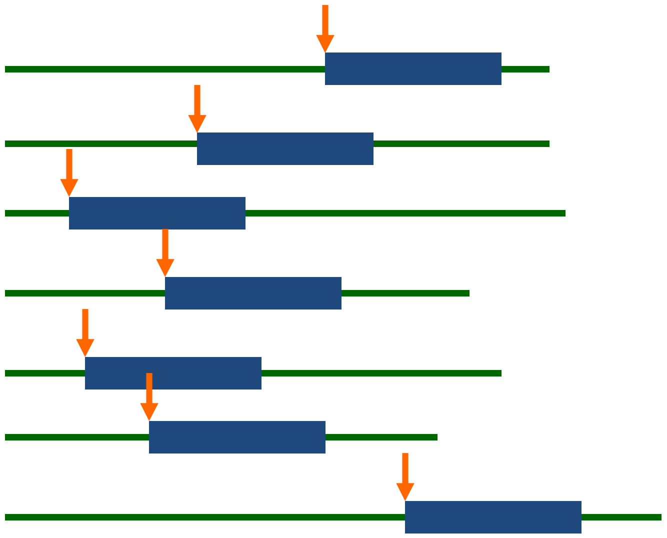


Calculate P(Seq_{window}|Motif) for every starting location

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

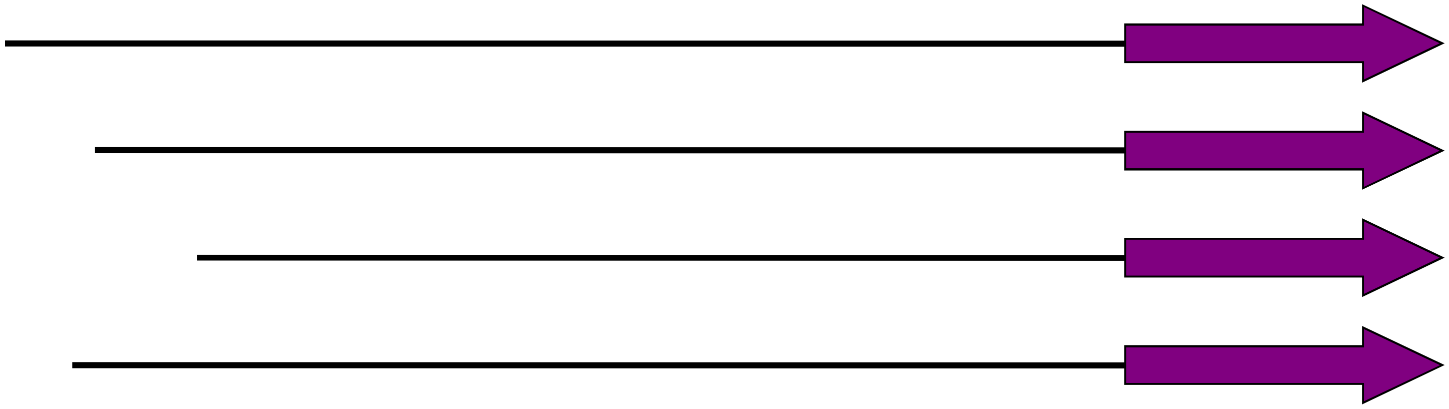
1. Start with random motif model
2. E Step: estimate probability of motif positions for each sequence
3. M Step: use estimate to update motif model
4. Iterate (to convergence)

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

Demo: Initialization

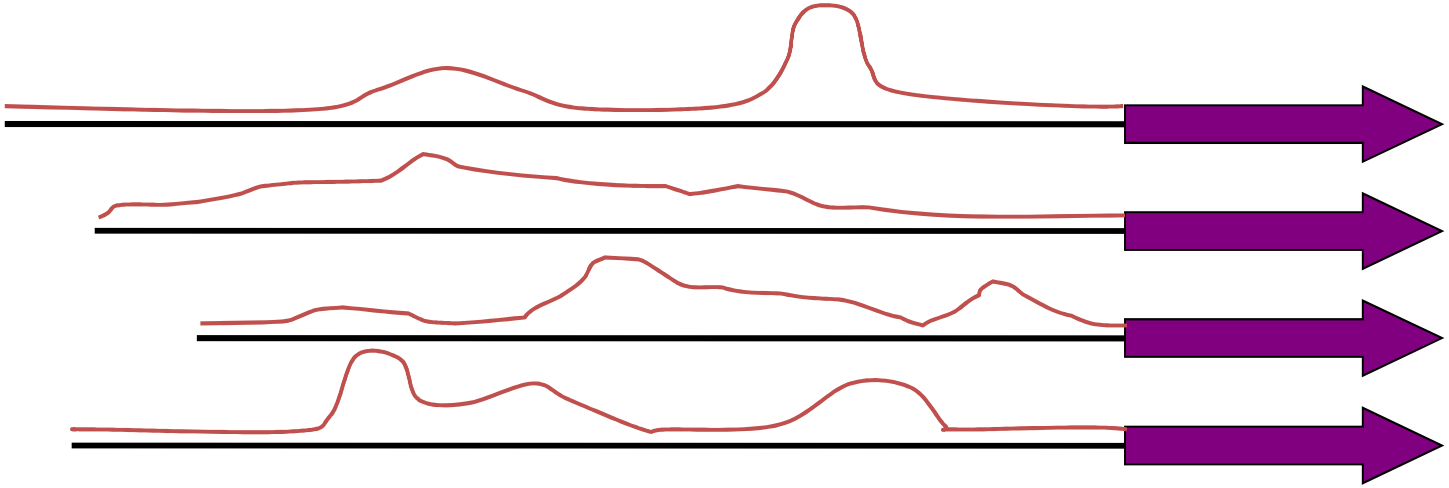
- Given a random motif model

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:

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

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

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

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

$$Pr(X_i, Z_{i3} = 1|P)$$

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

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

The E-step: Estimating *Z*

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

$$\begin{aligned} Z_{ij}^{(t)} &= Pr(Z_{ij}=1|X_i,P^{(t)}) \\ &= \frac{Pr(X_i,Z_{ij}=1|P^{(t)})}{\sum_{k=1}^{L-w+1} Pr(X_i,Z_{ik}=1|P^{(t)})} \end{aligned}$$

Example: Estimating *Z*

$X_i =$ **G C T G T A G**

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

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

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

$$\vdots$$

- 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

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

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

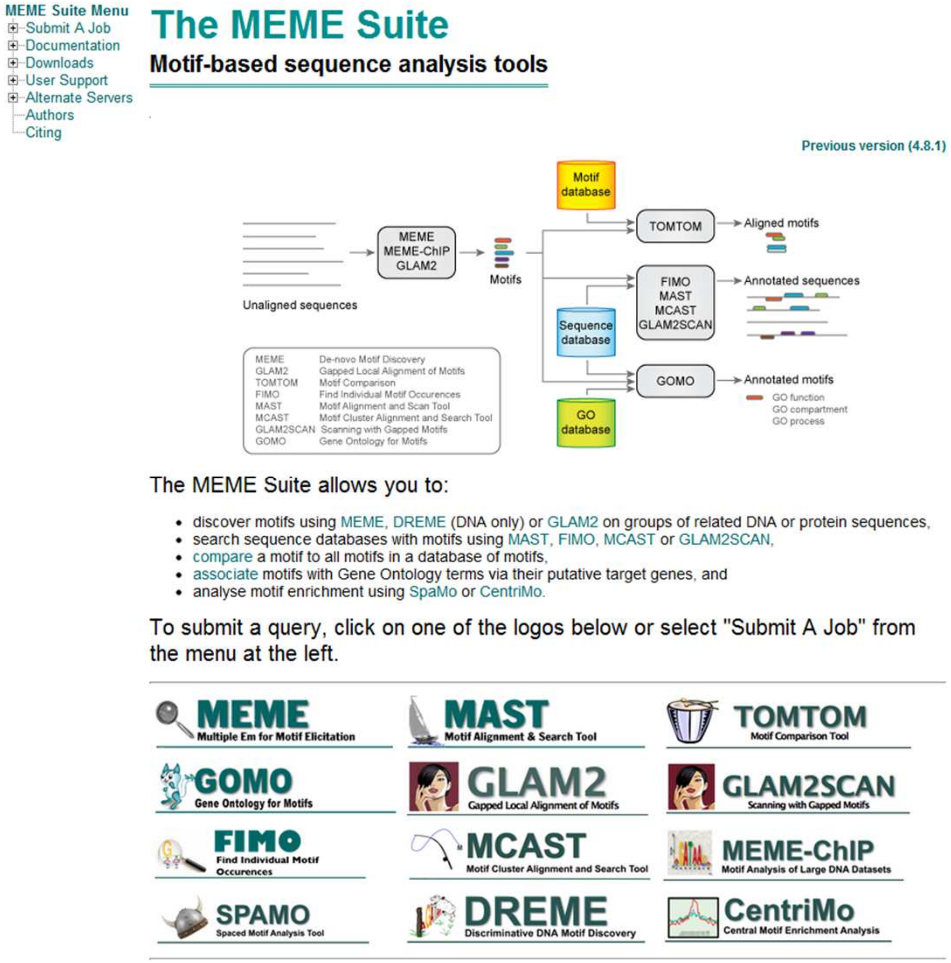
$$Z_{3,1}=0.2, Z_{3,2}=0.6, Z_{3,3}=0.1, Z_{3,4}=0.1$$

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

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.


MEME Suite Menu
Submit A Job
Documentation
Downloads
User Support
Alternate Servers
Authors
Citing
Previous version (4.8.1)
Unaligned sequences
MEME
MEME-ChIP
GLAM2
Motifs
Motif database
TOMTOM
Aligned motifs
FIMO
MAST
MCAST
GLAM2SCAN
Annotated sequences
Sequence database
GOMO
Annotated motifs
GO function
GO compartment
GO process
GO database
MEME Multiple Em for Motif Elicitation
MAST Motif Alignment & Search Tool
TOMTOM Motif Comparison Tool
GOMO Gene Ontology for Motifs
GLAM2 Gapped Local Alignment of Motifs
GLAM2SCAN Scanning with Gapped Motifs
FIMO Find Individual Motif Occurences
MCAST Motif Cluster Alignment and Search Tool
MEME-ChIP Motif Analysis of Large DNA Datasets
SPAMO Spaced Motif Analysis Tool
DREME Discriminative DNA Motif Discovery
CentriMo Central Motif Enrichment Analysis

The MEME Suite

Motif-based sequence analysis tools

MEME	De-novo Motif Discovery
GLAM2	Gapped Local Alignment of Motifs
TOMTOM	Motif Comparison
FIMO	Find Individual Motif Occurences
MAST	Motif Alignment and Scan Tool
MCAST	Motif Cluster Alignment and Search Tool
GLAM2SCAN	Scanning with Gapped Motifs
GOMO	Gene Ontology for Motifs

The MEME Suite allows you to:

- discover motifs using MEME, DREME (DNA only) or GLAM2 on groups of related DNA or protein sequences,
- search sequence databases with motifs using MAST, FIMO, MCAST or GLAM2SCAN,
- compare a motif to all motifs in a database of motifs,
- associate motifs with Gene Ontology terms via their putative target genes, and
- analyse motif enrichment using SpaMo or CentriMo.

To submit a query, click on one of the logos below or select "Submit A Job" from the menu at the left.

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

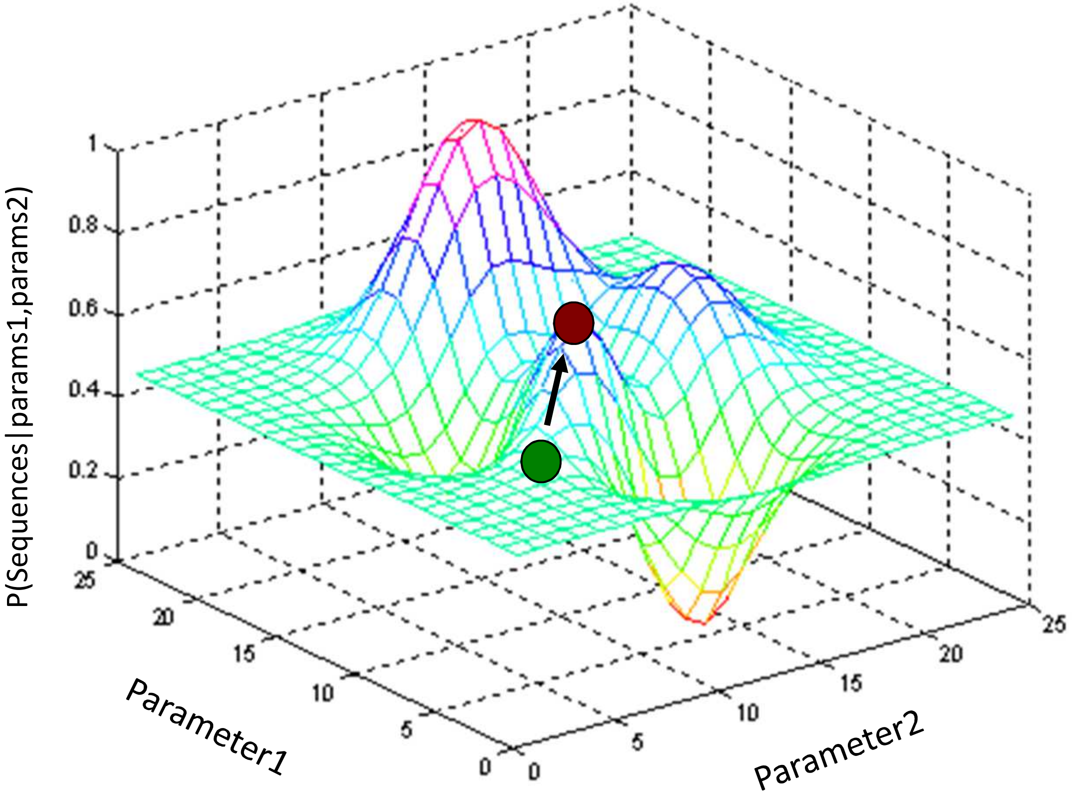
P(Seq|Model) Landscape

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

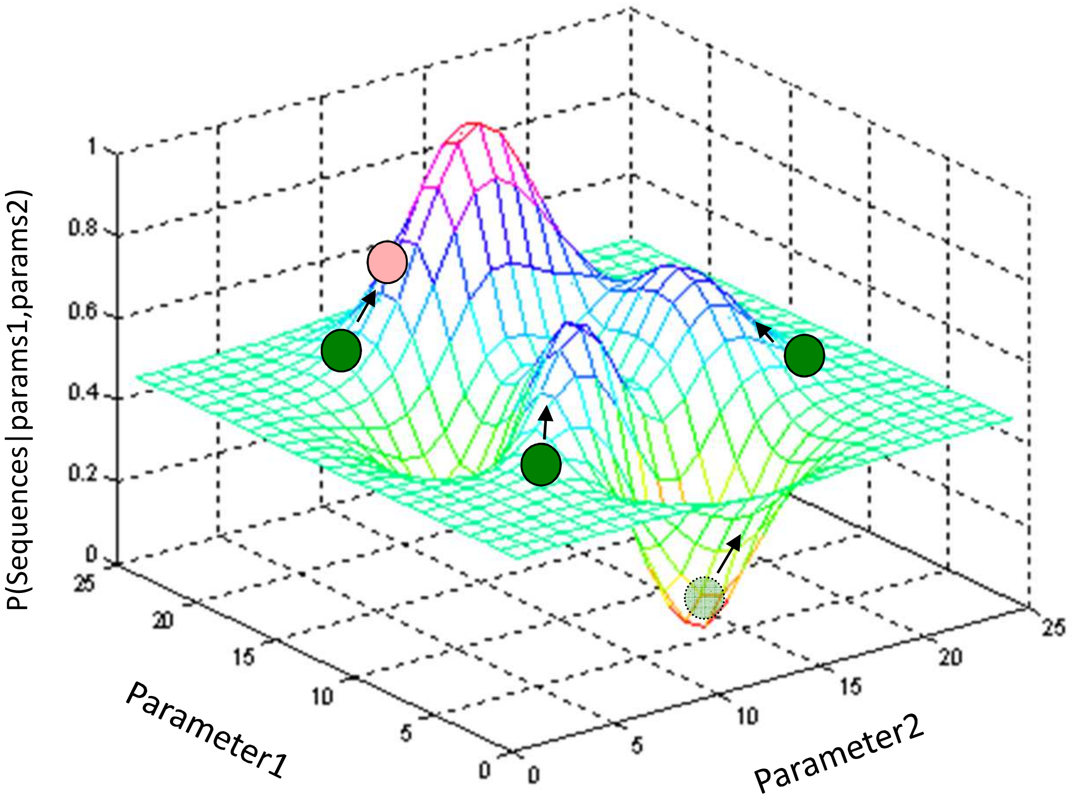
To minimize the effects of local maxima, you should search multiple times from different starting points

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

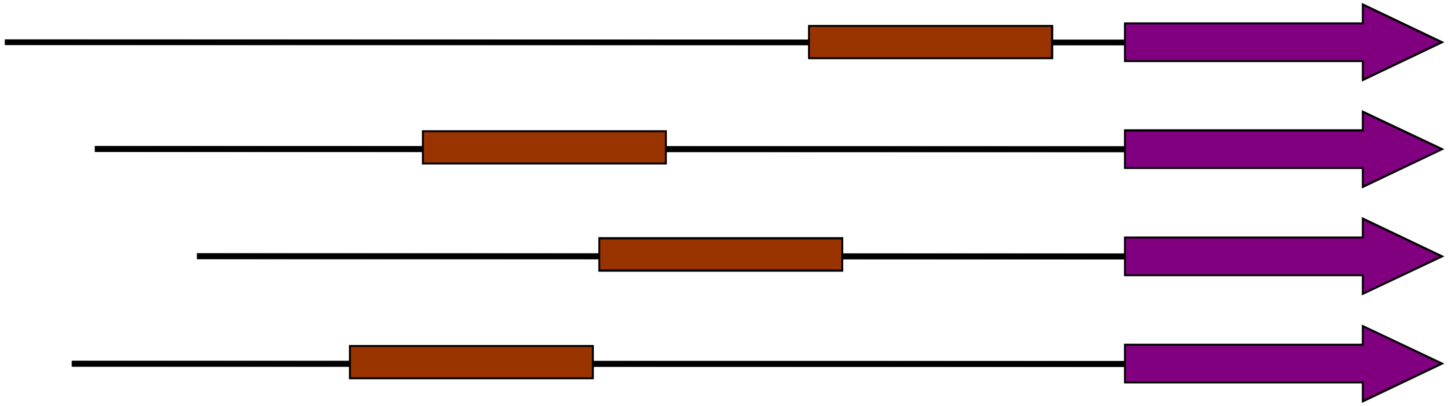
- For each possible starting position, $a_i = j$, compute a weight (likelihood ratio)

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

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

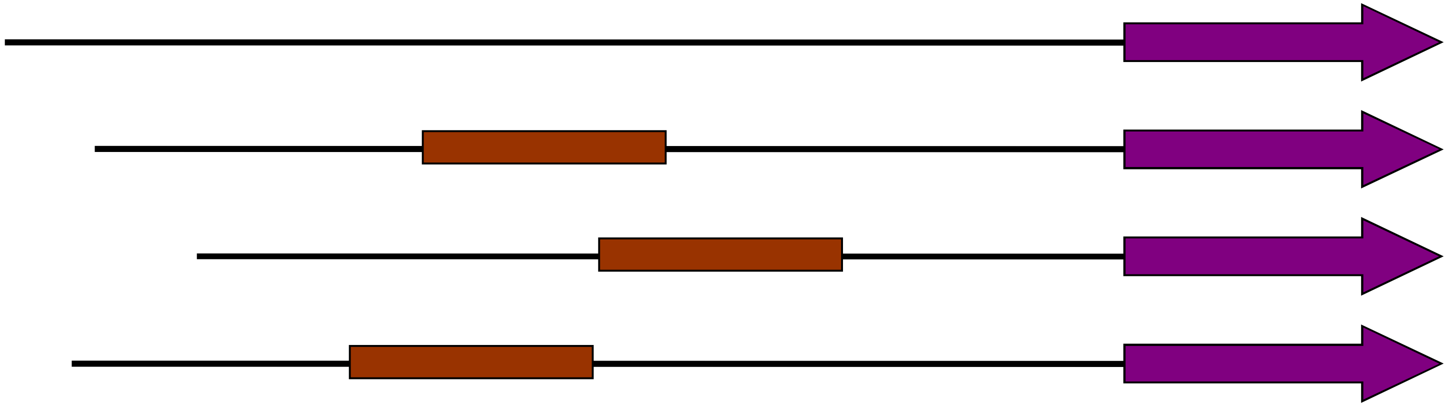
Demo: Initialization

- Random choose motif location



Demo: Step 2

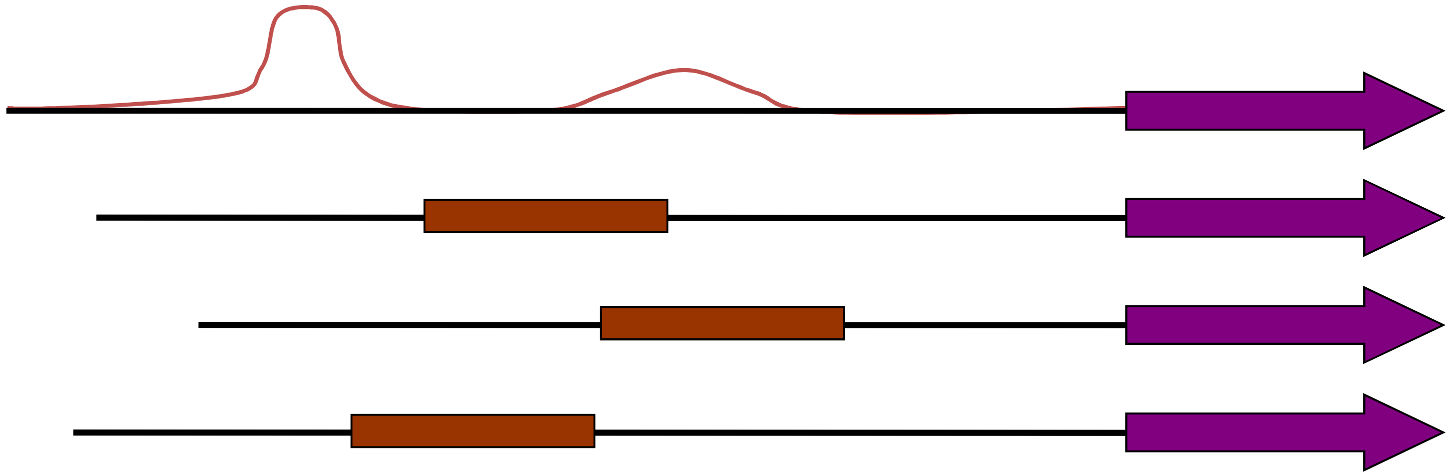
- Random Select One Sequence and Remove Its Motif. Recalculate its Temporal Model



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

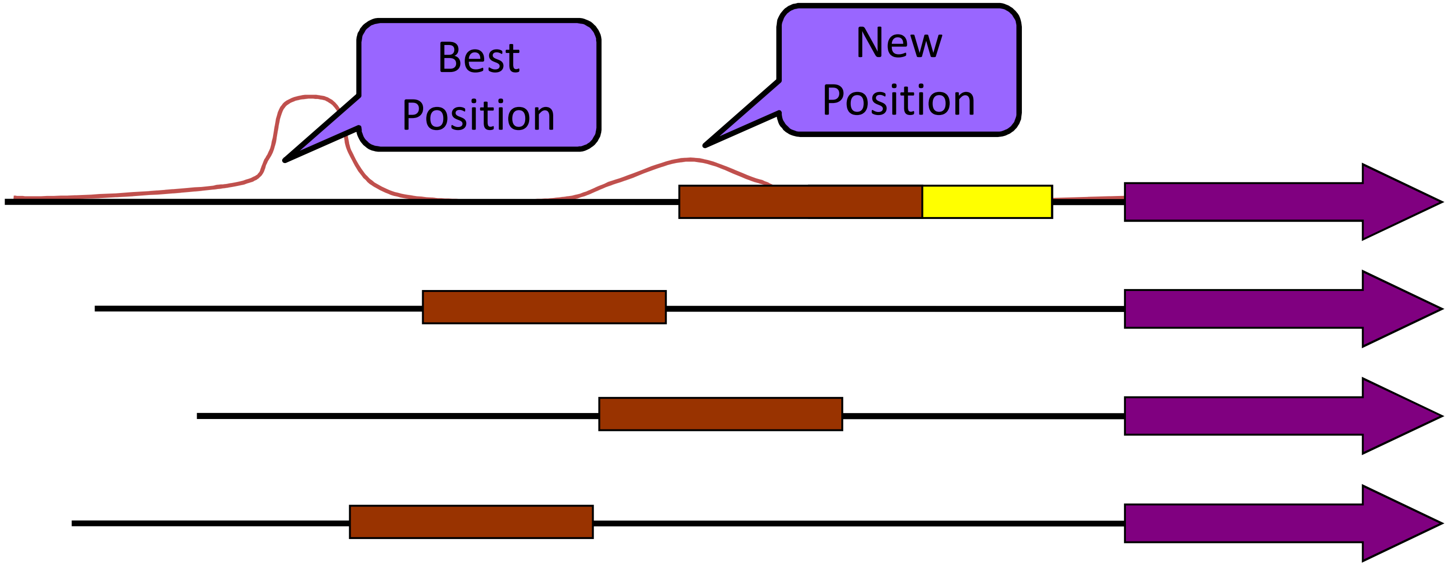
Demo: Step 3

- Calculate Probability of motif at each position on sequence



Demo: Step 4

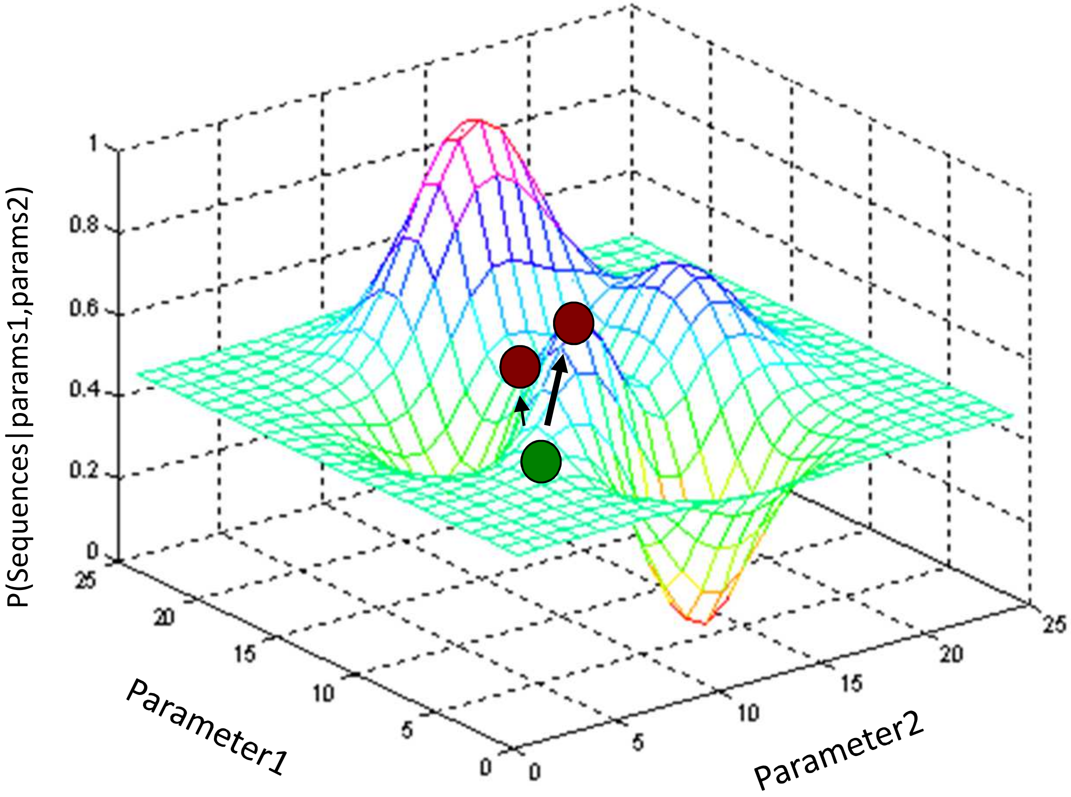
- Demo: Select New Position, Update Motif Model



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

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

DeepBind

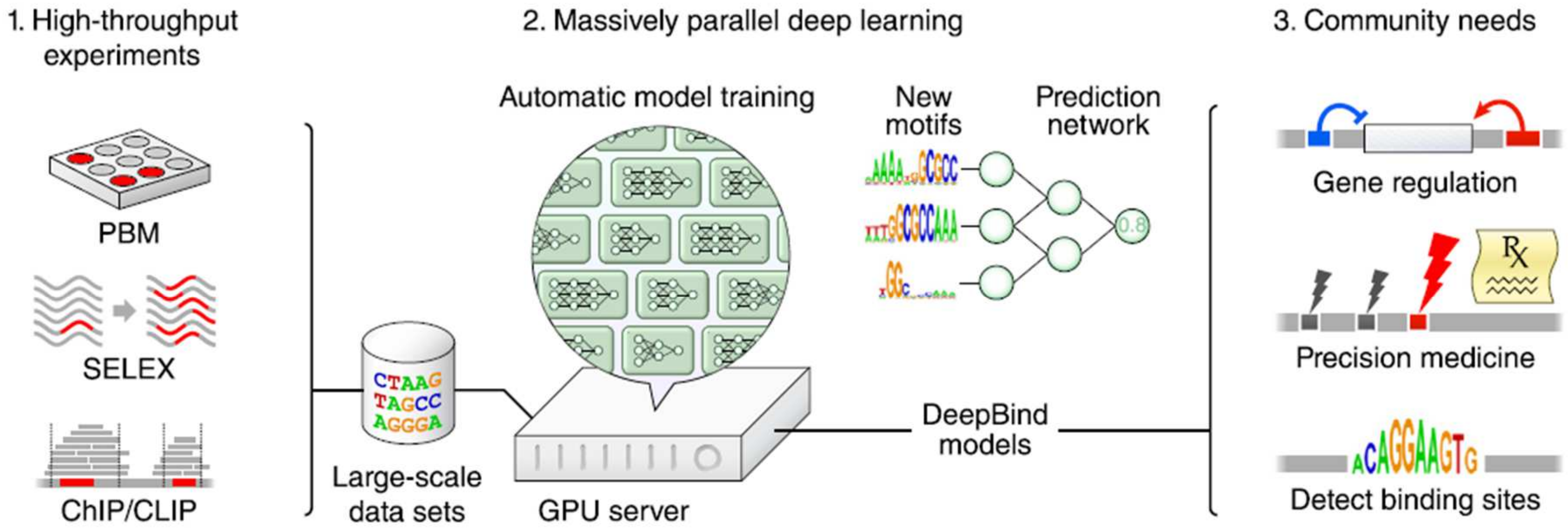


Figure 1 DeepBind's input data, training procedure and applications. 1. The sequence specificities of DNA- and RNA-binding proteins can now be measured by several types of high-throughput assay, including PBM, SELEX, and ChIP- and CLIP-seq techniques. 2. DeepBind captures these binding specificities from raw sequence data by jointly discovering new sequence motifs along with rules for combining them into a predictive binding score. Graphics processing units (GPUs) are used to automatically train high-quality models, with expert tuning allowed but not required. 3. The resulting DeepBind models can then be used to identify binding sites in test sequences and to score the effects of novel mutations.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotech. 33(8), 831-839, 2015, doi:10.1038/nbt.3300

DeepBind

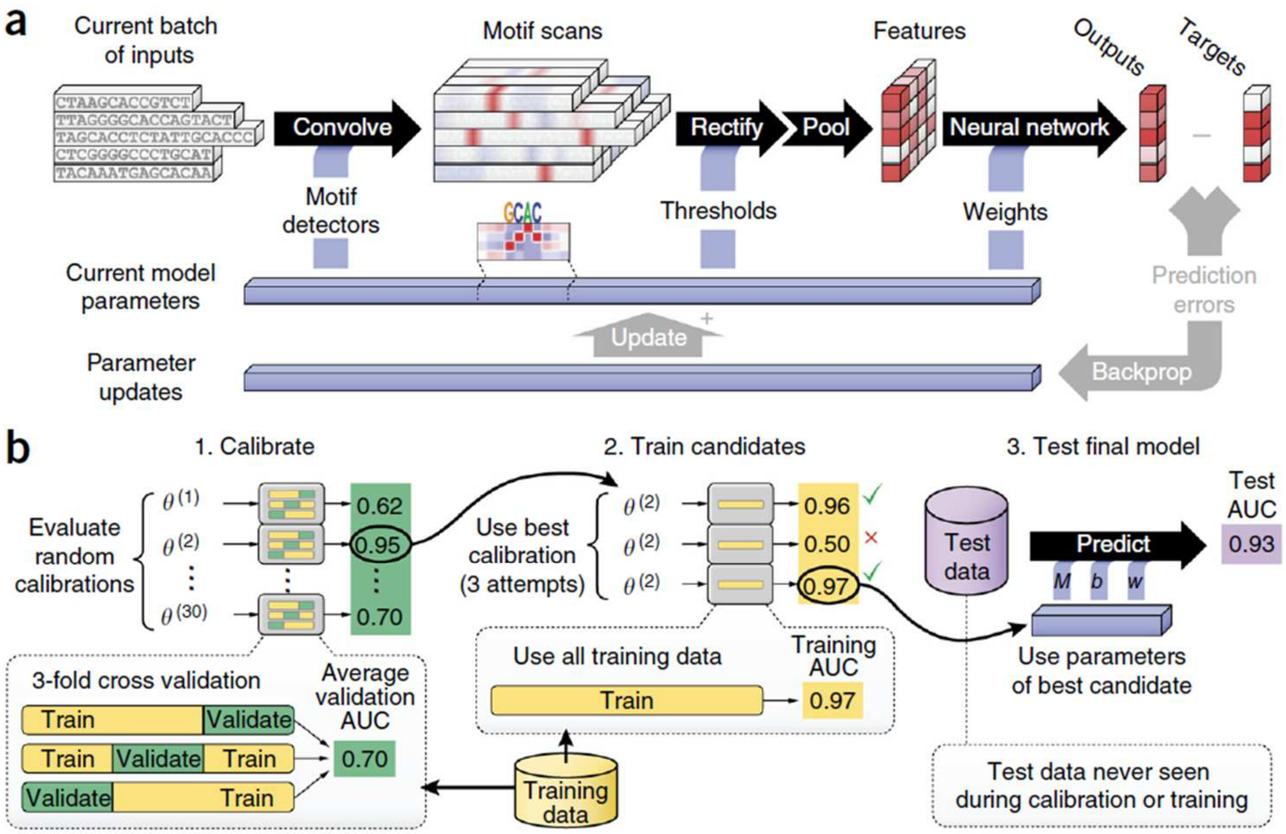


Figure 2 Details of inner workings of DeepBind and its training procedure. (**a**) Five independent sequences being processed in parallel by a single DeepBind model. The convolve, rectify, pool and neural network stages predict a separate score for each sequence using the current model parameters (**Supplementary Notes**, sec. 1). During the training phase, the backprop and update stages simultaneously update all motifs, thresholds and network weights of the model to improve prediction accuracy. (**b**) The calibration, training and testing procedure used throughout (**Supplementary Notes**, sec. 2).

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.
- **Babak Alipanahi, Andrew Delong, Matthew T Weirauch, Brendan J Frey.** Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotech. 33(8), 831-839, 2015, doi:10.1038/nbt.3300.