Exercises

1) Fill in the A and E matrix given this training data

E	A	G	C	T
I	0	3/7	1/7	3/7
O	1	0	0	0

A	I	O
I	4/6	2/6
O	3/7	4/7

```
I to I: 4/6
I to O: 2/6
O to I: 3/7
O to O: 4/7
```

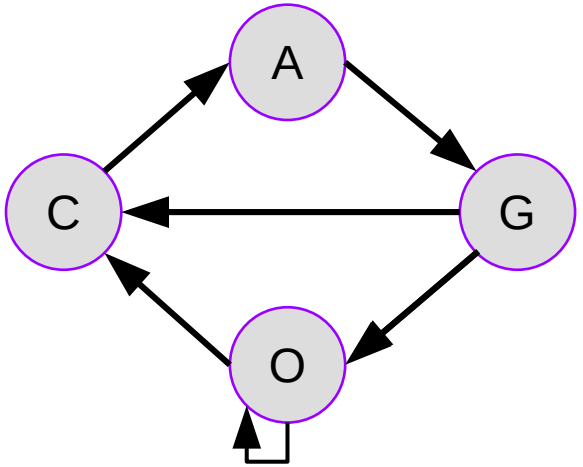
```
Sequence:    AAAAATCGGGATAT

Labels:      OOOOOIIIIIOIOI
```

2) Huntington's disease is caused by $(CAG)_n$ repeats. Propose an HMM that would identify such repeats.

3) Give one possible sequence and labels for the following HMM, assuming orange is possible, white is not possible, IA emits A, etc.

CAG repeat



A	IA	IG	IC	IT	OA	OG	OC	OT
IA								
IG								
IC								
IT								
OA								
OG								
OC								
OT								

Give one possible sequence and labels for the following HMM, assuming orange is possible, white is not possible, IA emits A, etc.

A	IA	IG	IC	IT	OA	OG	OC	OT
IA								
IG								
IC								
IT								
OA								
OG								
OC								
OT								

```
OO/OO/OO/OO
AA/AG/GG/GA not poss

IO/IO/IO/IO
AC/AT/GC/GT only

I can't be C/T

ATCGACTGACTTCG
OOOIIOOIIOOOOO

AGAGAGTGGGGGGG
IIIIIIOOIIIIII
or
IIIIIIOIIIIIII
```

Exercises

1) What are pseudocounts used to avoid in the EM algorithm? Zero probabilities & log(0)

2) Why does this model not work well in identifying CpG islands: CpG are not depleted,

   can be C or G rich alone,

   no dependence on prior state

A	I	O
I	0.8	0.2
O	0.2	0.8

E	A	C	G	T
I	0.1	0.4	0.4	0.1
O	0.25	0.25	0.25	0.25

3) Why do profile HMMs work better for distant homology searches? Profile HMMs have profile at each position, like PSSM

4) Propose an HMM model (A and E matrix) for the following:

```
Sequence:    ATCGAAAATCGGGATATATATGACTTAATTCTCGTA
Labels:      OOOOOOOOOOOOIIIIIIIIOOOOOOOOOOOOOO
```

HMM for AT repeat

First attempt

E	A	G	C	T
I	1	0	0	1
O	.25	.25	.25	.25

A	I	O
I	.8	
O		.8

```
CAGACTCAATATAAATTTA
OOOOOOOIIIIIIIIIII
```

But no dependency on prior state, what if we only want AT_n repeat.

HMM for AT repeat

Second attempt

A	IA	IG	IC	IT	OA	OG	OC	OT
IA								
IG								
IC								
IT								
OA				?				
OG				?				
OC				?				
OT				?				

E	A	G	C	T
IA				
IG				
IC				
IT				
OA				
OG				
OC				
OT				

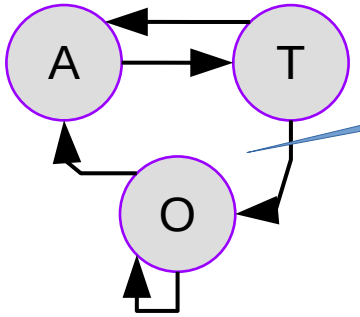
```
CAGACTCTATATATAT
OOOOOOOIIIIIIII
or
OOOOOOOIIIIIIIII
```

HMM for AT repeat

Second attempt

A	IA	IG	IC	IT	OA	OG	OC	OT
IA	0							0
IG								
IC								
IT				0	0			
OA								
OG								
OC								
OT								

E	A	G	C	T
IA				
IG				
IC				
IT				
OA				
OG				
OC				
OT				



Collapse?

```
Sequence:    ATCGAAAATCGGGATATATATGACTTAATTCTCGTA
Labels:      OOOOOOOOOOOOIIIIIIIIOOOOOOOOOOOOOO
```

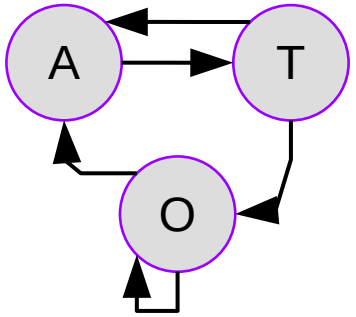
HMM for AT repeat

Second attempt alternative

A	A	T	O
A	0		0
T		0	
O		0	

E	A	G	C	T
A		0	0	0
T	0	0	0	
O				

```
Sequence:   ATCGAAAATCGGGATATATATACTTAAT
Labels:     OOOOOOOOOOOOIIIIIIIIOOOOOOO
```



Today's objectives

- Motif biology and examples
- Position Weight Matrices
- Scoring motifs
- Motif Logos
- Motif finding

Motifs in Biology

A motif is a nucleotide or amino-acid sequence pattern that is widespread and thought to have a biological function

DNA motifs
- transcription factor binding sites
- splice sites
- micro RNA binding sites

Protein motifs
- phosphorylation sites
- localization sequence (nucleus/mito)
- protein binding/interaction sites

Motifs are caused by physical interactions (binding energy)

Types of interactions: varying specificity

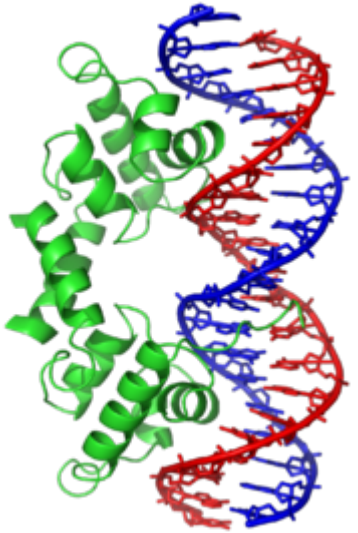
DNA - Protein

- restriction enzymes (high specificity)
- transcription factors (medium specificity)
- nucleosomes (low specificity)

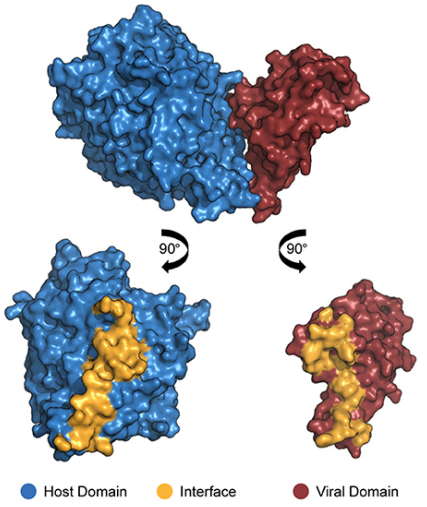
Protein - Protein

- post-translational modification sites, e.g. phosphorylation sites
- localization sequences (e.g. nucleus/mito)
- protein binding/interaction sites

Lambda repression-DNA



Herpes glycop D domain



Motif representation

Consensus:
ACTGCGCGG

Degenerate sequence:
ACTGC[GC][GC][TG][AG]

Motif logo: Position frequency matrix

Higher specificity

Lower specificity

	Sequences	binding energy or ΔG
bound	ACTGCGCGG	-4
	ACTGCGCGA	-3
	ACTGCGGGG	-3
	ACTGCCCTG	-2
	ACTGCCGGA	-1
unbound	.	.
	.	.
	.	.
	GGGGGGGGG	4

Why do we need motifs? (i) Different sequences can have the same binding energy, (ii) Not all functional (bound) sequences have the same binding energy

Finding Transcription Factor Binding Sites (TFBS)

Experimental approaches (one by one)

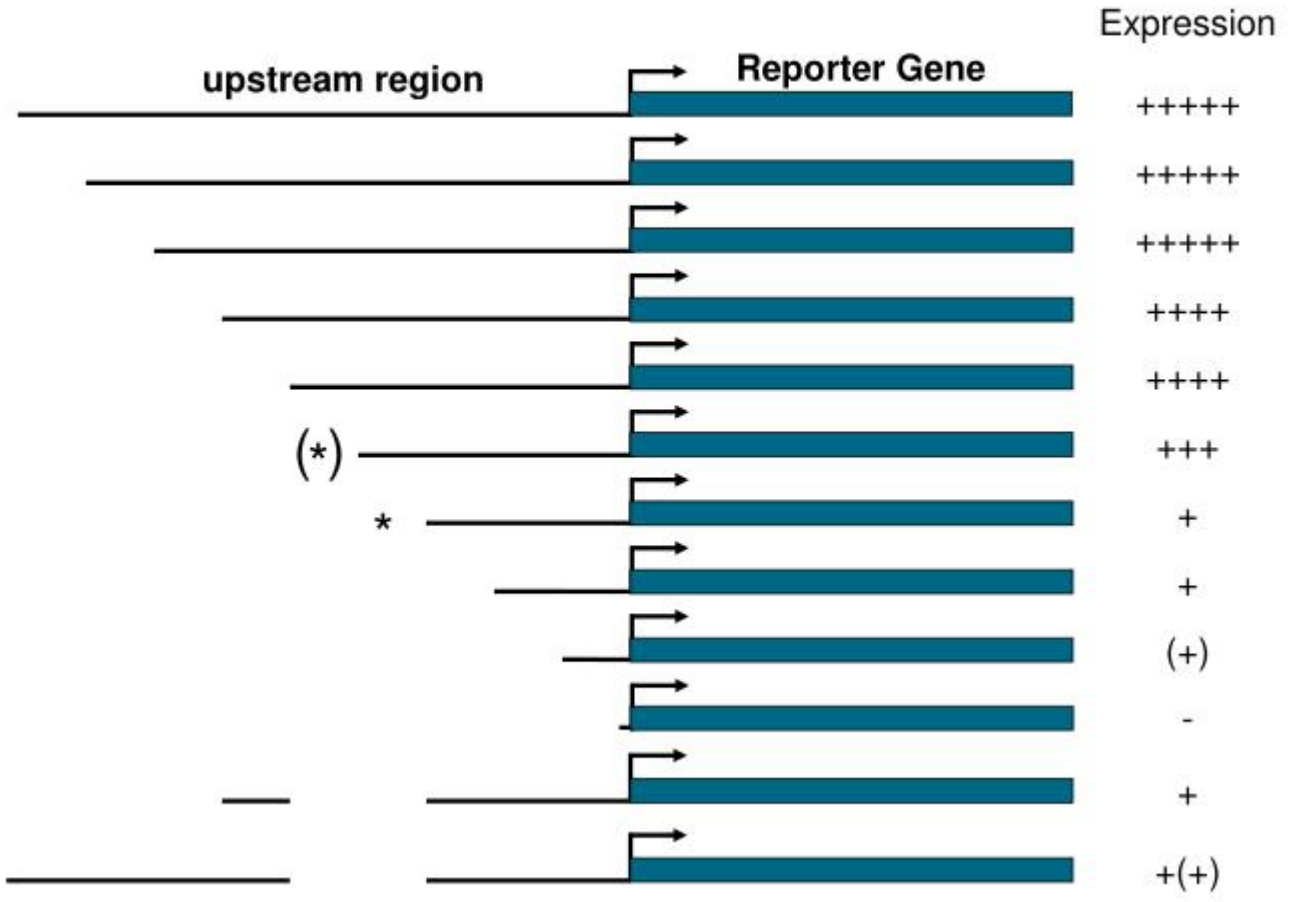
- promoter bashing*
- gel shift*
- footprinting*

*now high-throughput

Computational approaches (motif finding)

- words
- motifs
  - Expectation maximization
  - Gibbs Sampling
  - Phylogenetic footprinting

Experimental Data: promoter bashing

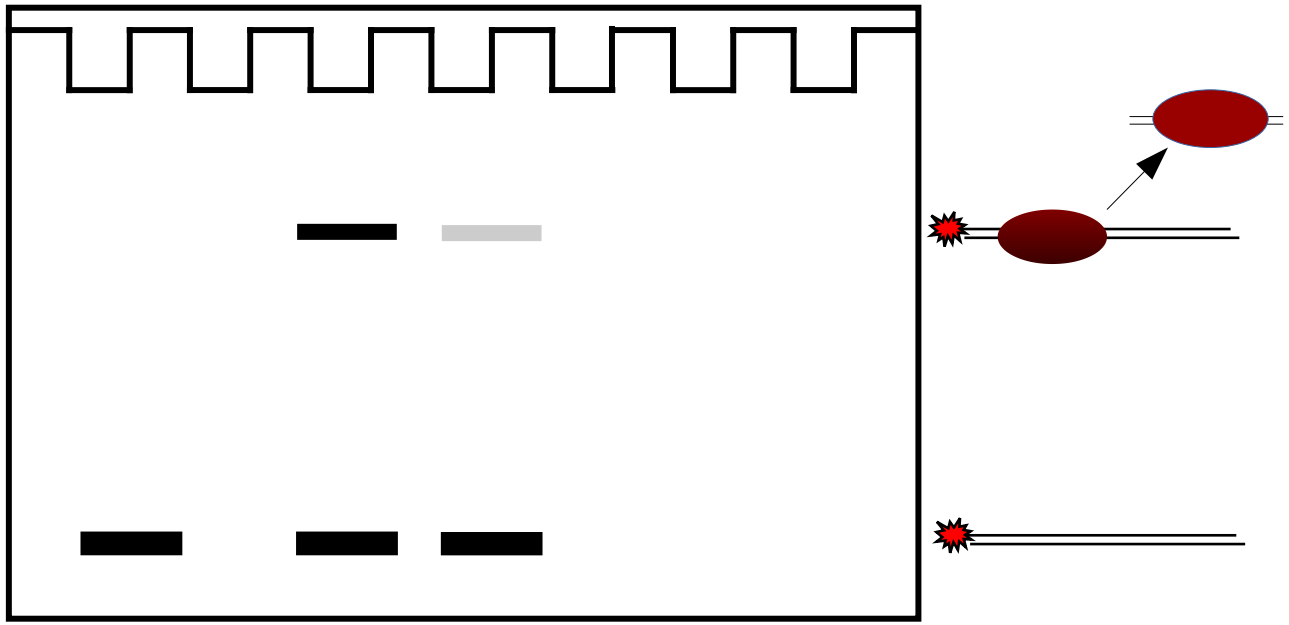


Where is the cis-regulatory element (transcription factor binding site)?

EMSA: electrophoretic mobility shift assay

Cold competitor			+
Protein		+	+
Labeled DNA	+	+	+

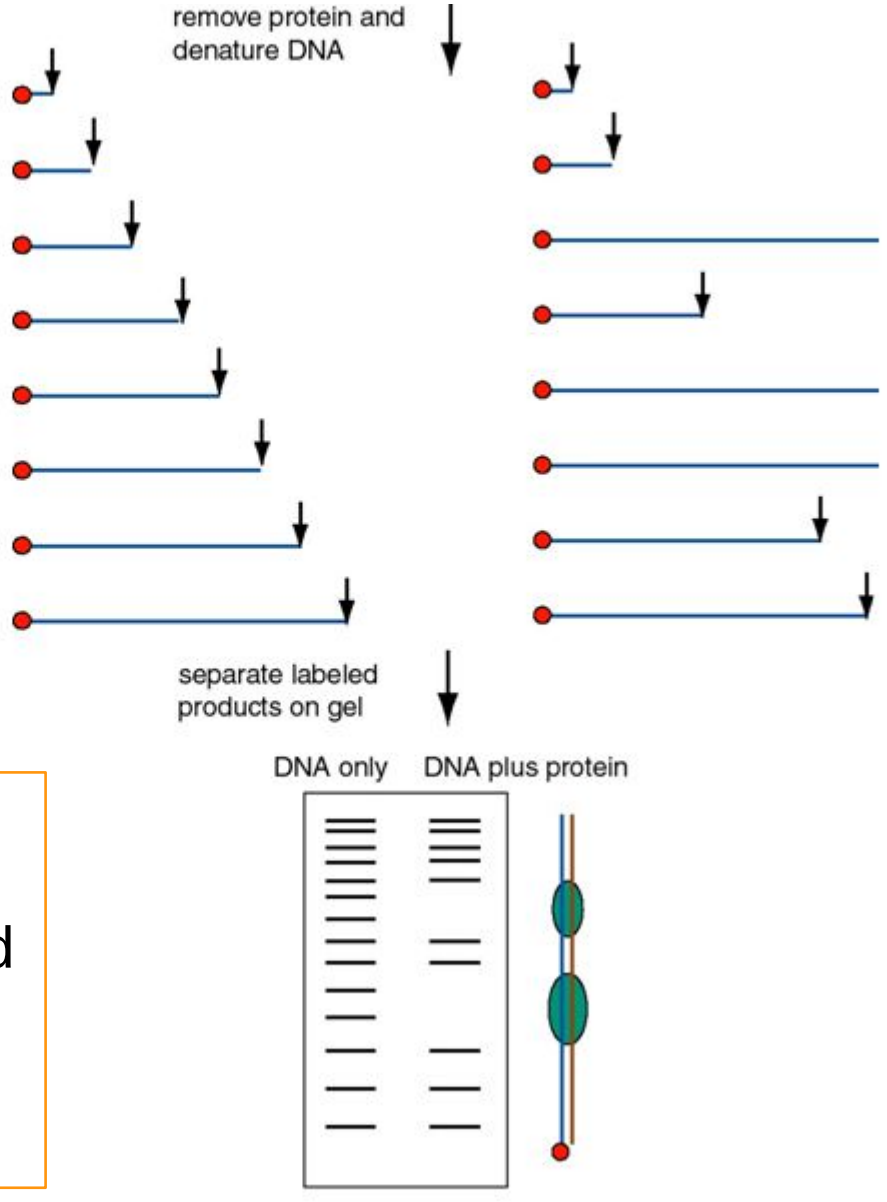
- Competitor = non-specific
- Competitor = mutant
- What if cold competitor has no effect?



Does a protein bind DNA, is there a physical interaction?

DNase footprinting assay

Specific locations of protected segments show the binding site(s) for the protein.



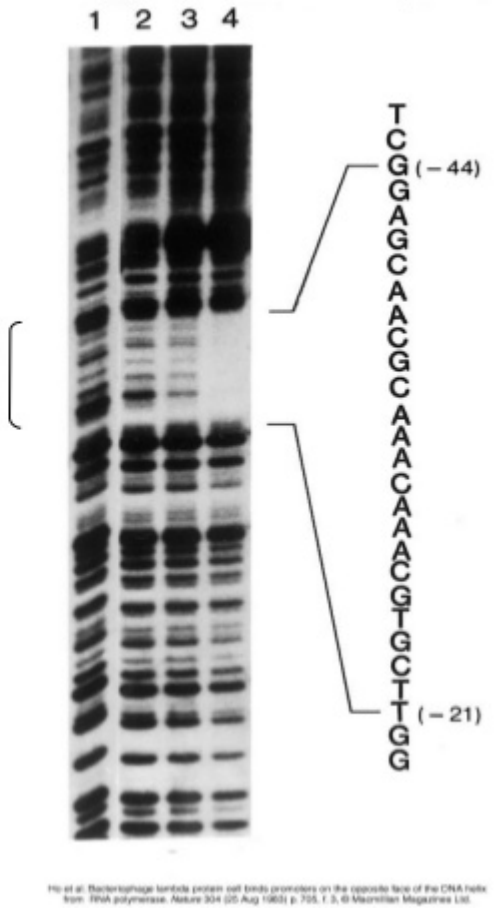
- deoxyribonuclease (DNase) is used to cut end labeled DNA followed by gel electrophoresis
- DNA bound by protein isn't cut

Where are proteins bound to DNA?

Footprinting assay

Sample of a DNase I footprinting gel.

Fig. 5.37b

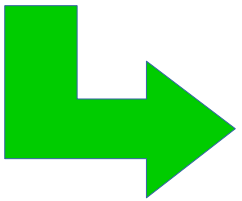


Samples in lanes 2-4 had increasing amounts of the DNA-binding protein (lambda protein cII); lane 1 had none.

Upstream Activating Sequences (UAS)

Table 2 Similar elements in the promoters of genes expressed in gluconeogenesis. Sequences for which no reference is given were taken from the EMBL data base

Gene	Sequence	Position	Reference
ICL1	**CGG ATG AAT GGA**	–388, –399	Schöler and Schüller 1994
PCK1	**CGG GTG AAT GGA**	–562, –551	Mercado and Gancedo 1992
ACR1	**CGG TTG AAT GGA**	–618, –607	Fernández et al. 1994
ACR1	**CGG TTT AAT GGA**	–679, –668	Fernández et al. 1994
CIT2	**CGG ATC AAT GGA**	–854, –865	–
PCK1	**CGG ATG AAA GGA**	–471, –482	Mercado and Gancedo 1992
FBP1	**CGG ACG GAT GGA**	–508, –493	Rogers et al. 1988
ACS1	**CGG ACG AAC GGC**	–426, –415	Kratzer and Schüller 1995
MLS1	**CGG CCC AAT GGA**	–490, –501	Caspary et al. 1997
MLS1	**CGG CTC AAT GGA**	–531, –520	Caspary et al. 1997
MDH2	**CGG CCG AAT GGG**	–229, –240	–
FBP1	**CGG ACA CCC GGA**	–432, –421	Rogers et al. 1988



Transcription factor binding sites to Position Weight Matrix

Motif models

A Position Weight Matrix (PWM) is a probabilistic representation of a biological motif.

Aligned sequences

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

Consensus: TAGGTAAGT

Why PWM?
Background [A,C,G,T]

Creating a PWM

- position frequency matrix (PFM) – counts
- position probability matrix (PPM) – probabilities
- position weight matrix (PWM) – $\log_2$(P(motif/background))

$$\text{PFM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}.$$

$$\text{PPM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

$$\text{PWM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}.$$

Pseudocounts

Is the probability of a [A,C,T] really zero at position 4?

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

A pseudocount is an amount (not generally an integer, despite its name) added to the number of observed cases in order to change the expected probability in a model of those data, when not known to be zero.

$$\theta_i = \frac{x_i + \alpha}{N + \alpha d} (i = 1, \ldots, d)$$

Theta = frequency of base i
x = counts of base *i*
N = total observations
alpha = pseudocount
d = alphabet size (DNA = 4)

Probability under motif model

Both PPMs and PWMs assume statistical independence between positions in the motif, as the probabilities for each position are calculated independently of other positions.

Given this assumption, the probability of a motif given a sequence is the product of the PPM.

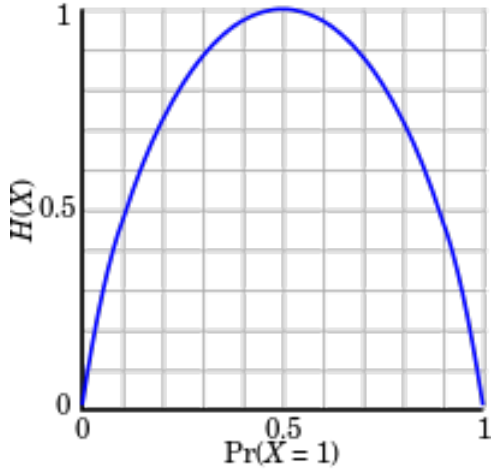
Sequence: ATCATGAT

```
ATC = .2 x .2 x .1 = 0.004
 TCA = .2 x .2 x .1 = 0.004
  CAT = .3 x .5 x .7 = 0.105
```

$$PPM = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.2 & .5 & .1 \\ 0.3 & .1 & .1 \\ 0.3 & .2 & .1 \\ 0.2 & .2 & .7 \end{bmatrix}$$

Entropy

Entropy is a measure of the unpredictability of a state.



H = entropy

$$H = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

x = [.25, .25, .25, .25]
H = 2 (high entropy = unpredictable)
x = [.997, .001, .001, .001]
H = 0.034 (low entropy = predictable)

Information Content

How much information is conveyed by a motif model?

The maximum information encoded in a coin flip is:

$\log_2(2/1) = 1$ bit

The max information encoded in a single base pair is:

$\log_2(4/1) = 2$ bits

The information content of a motif is:

$$IC_i = \log_2(4) - (H_i)$$

$$H_i = -\sum_{b=A,C,G,T} PPM_{i,b} \log_2(PPM_{i,b})$$

H = entropy

$$H = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

IC is inverse of entropy and conveys how much information a site provides relative to background

IC = average score of sequences of a motif

Information Content

AAA motif
IC(AAA) = 2+2+2 = 6 (max)
Degenerate motif NNN
IC(NNN) = 0+0+0 = 0 (min)

High information content = low entropy

Low information content = high entropy

The information content of a site is:

$$IC_i = \log_2(4) - (H_i)$$

$$H_i = -\sum_{b=A,C,G,T} PPM_{i,b} \log_2(PPM_{i,b})$$

H = entropy

$$H = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

IC is inverse of entropy and conveys how much information a site provides relative to background

IC = average score of sequences of a motif

Information Content

AAA motif
IC(AAA) = 2+2+2 = 6 (max)
Degenerate motif NNN
IC(NNN) = 0+0+0 = 0 (min)

High information content = low entropy

Low information content = high entropy

The information content of a site is:

$$IC_i = \log_2(4) - (H_i)$$

$$H_i = -\sum_{b=A,C,G,T} PPM_{i,b} \log_2(PPM_{i,b})$$

Including background, the information content of a motif is:

$$IC = \sum_{i=1}^{L} \sum_{b=A,C,G,T} f_{b,i} log_2(f_{b,i}/p_b)$$

$$IC = \sum_{i=1}^{L} \sum_{b=A,C,G,T} PPM(b,i) PWM(b,i)$$

IC is inverse of entropy and conveys how much information a site provides relative to background

IC = average score of sequences of a motif

Kullback-Leibler information or relative entropy ~Log-likelihood ratio

Information content

- Lower information content
  - fewer informative sites
  - shorter = lower IC
- More sites in a genome expected by chance
- Lower specificity (targets vs random)

- Higher information content
  - more informative sites
  - longer = higher IC
- Fewer sites in a genome expected by chance
- Higher specificity (targets vs random)

Summary of statistical motif model:

- Experimentally identify TFBS (the more the better)
- Generate PWM using these sequences

But how does PWM related to binding energy?

Binding energies ~ P(bound)

$$TF + S_i \underset{K_{off}}{\overset{K_{on}}{\Leftrightarrow}} TF - S_i$$

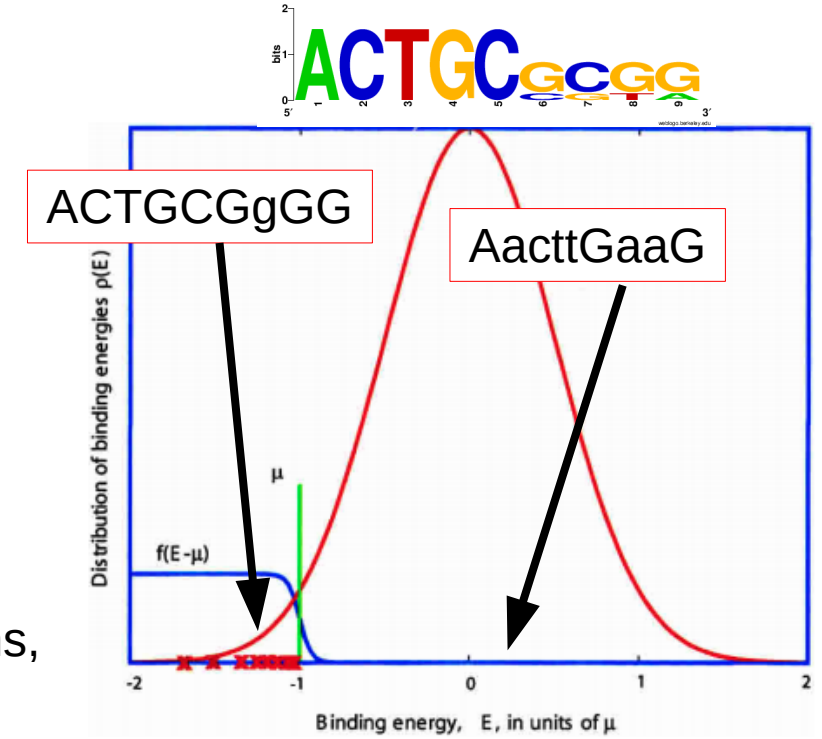
TF = transcription factor
S_i = Sequence i
K_{on} = binding rate
K_{off} = dissociation rate

$$K_i = \frac{K_{on}}{K_{off}} = \frac{[TF + S_i]}{[TF][S_i]}$$

K_i = equilibrium constant depends on concentrations, brackets.

$$P_{bound}(S_i) = \frac{1}{1 + e^{(E(S_i) - \mu)}}$$

Probability of Sequence i being bound, depends on:
u = chemical potential set by the factor concentration and
$E(S_i)$ = binding energy or ΔG



Binding energies of all sequences, those on the left are in the bound state

Djordjevic M, Sengupta AM, Shraiman BI. A biophysical approach to transcription factor binding site discovery. Genome research. 2003 Nov 1;13(11):2381-90.

Binding Energy ~ PWM

Binding energy (also called free energy) is the minimum energy required to separate DNA-protein complex.

Binding energy can be defined by PWM:

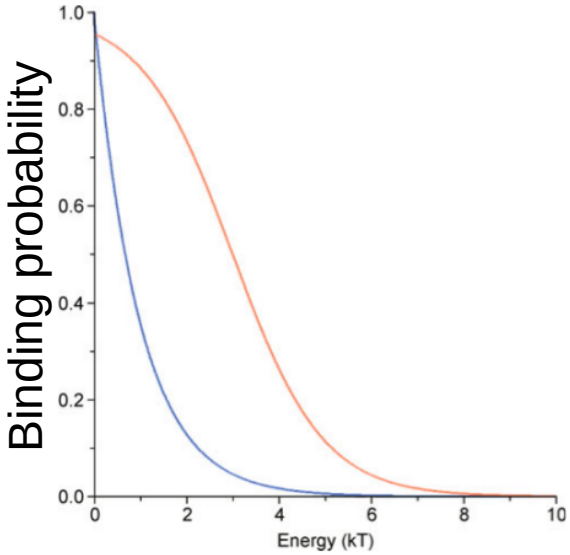
$$E_j = W \cdot S_j$$

E = energy
W = energy matrix
S = sequence
F = PPM
u ~ [TF]

$$W(b,i) = \log \frac{\max F(b,i)}{F(b,i)} \rightarrow E_j = E_{\min S_j} - W_{LP} \cdot S_j$$

W_{LP} = log prob matrix

$$P(\text{bound}|S_j) = \frac{e^{-E_j}}{e^{-\mu} + e^{-E_j}} = \frac{1}{1 + e^{E_j - \mu}}$$



absolute binding probability, 95% bound

relative binding probability vs consensus

PWM energy model with offset = probability model

Stormo (2013)

Finding Transcription Factor Binding Sites (TFBS)

Experimental approaches (one by one)

- promoter bashing*
- gel shift*
- footprinting*

*now high-throughput

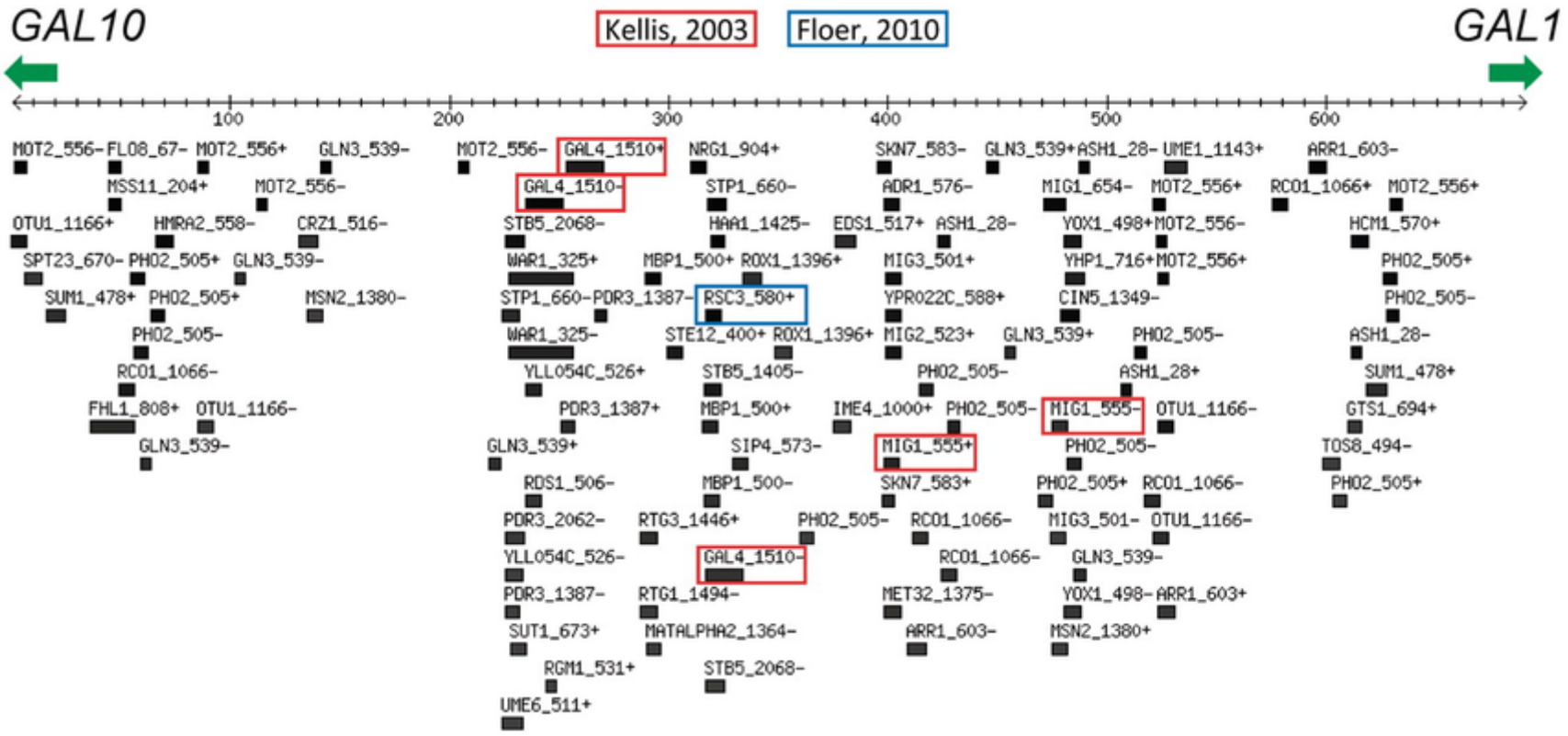
Computational approaches (motif finding)

- words
- motifs
  - Expectation maximization
  - Gibbs Sampling
  - Phylogenetic footprinting

Computational Challenge

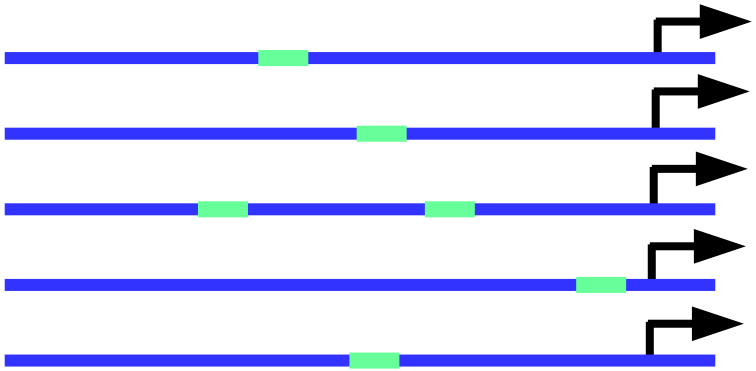
- Short 4-12 bp
- Degenerate sequences, low IC
- 1 site every 1kb
- e^{-IC} upper limit of motif frequency

IC(AAA) = 6
exp(-6) = 0.0025
0.25^3 = 0.0156



Over-represented motif

Bound or co-regulated genes



Unbound or background genes

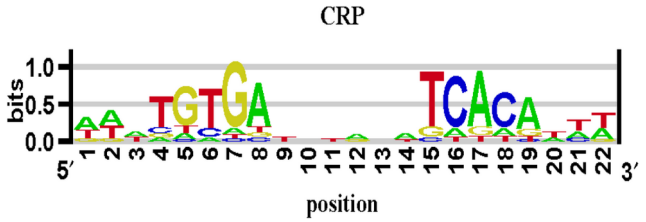
First Motif Finding

The Data Set: Sequences containing sites for cAMP receptor protein (CRP)

locus	sequence
colel	taatgtttgtgctggt**TTTTGTGGCATCGGGCGAGAAT**agcgcgtggtgtgaaagactgt**TTTTTTGATCGTTTTCACAAAA**atggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaac**AAAAGTGTCTATAATCACGGCA**gaaaagtccacattga**TTATTTGCACGGCGTCACACTT**tgctatgccatagcatttttatccataag
ecobglrl	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaat**AACTGTGAGCATGGTCATATTT**ttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcat**GTATGCAAAGGACGTCACATTA**ccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagca**AGGTGTTAAATTGATCACGTTT**tagaccatttttttcgtcgtgaaactaaaaaaacc
ecodeop	agtgaa**TTATTTGAACCAGATCGCATTA**cagtgatgcaaacttgtaagtagatttccttt**AATTGTGATGTGTATCGAAGTG**tgttgcggagtagatgttagaata
ecogale	gcgcataaaaaacggctaaattcttgtgtaaacgattccac**TAATTTATTCCATGTCACACTT**ttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggttttttgttatctgcaattcagtaca**AAACGTGATCAICCCCTCAATT**ttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ecolac	aacgcaat**TAATGTGAGTTAGCTCACTCAT**taggcaccccaggctttacactttatgcttccggctcgtatgttgtgtgg**AATTGTGAGCGGATAACAATTT**cac
ecomale	acattaccgccaa**TTCTGTAACAGAGATCACACAA**agcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggc**TTCTGTGAACTAAACCGAGGTC**atgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttgg**AATTGTGACACAGTGCAAATTC**agacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagactttttttttcat**ATGCCTGACGGAGTTCACACTT**gtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaac**GATTGTGATTCGATTCACATTT**aaacaatttcaga
ecouxul	cccatgagagtgaaat**TGTTGTGATGTGGTTAACCCAA**ttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatg**CGGTGTGAAATACCGCACAGAT**gcgtaaggagaaaataccgcatcaggcgctc
trn9cat	**CTGTGACGGAAGATCACTTC**gcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcga**AAATGAGACGTTGATCGGCACG**
tdc	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagtta**ATTTGTGAGTGGTCGCACATAT**cctgtt

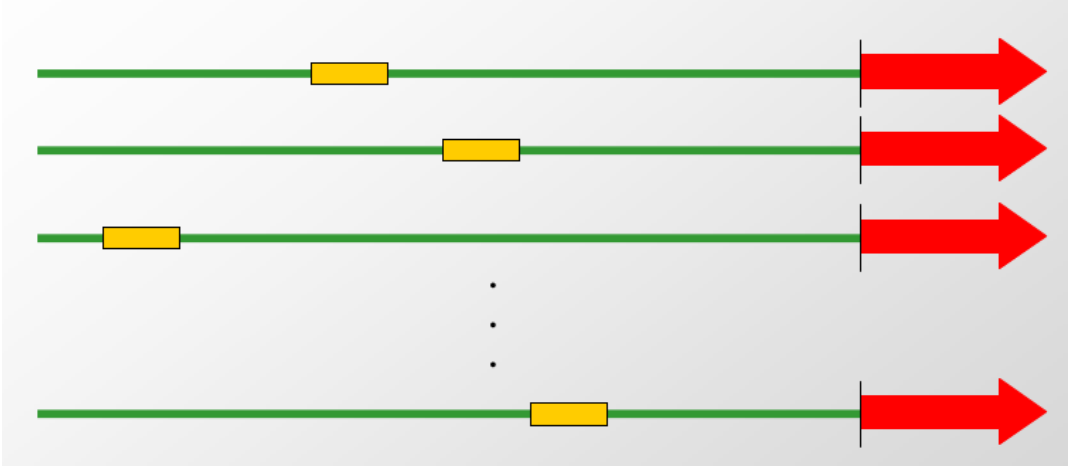
For this case, there are 18 sequences of length 105 bp and we are looking for a motif of width 20 bp. There are 86 different 20 bp subsequences per example and ~7x10^{34} alignments to check.

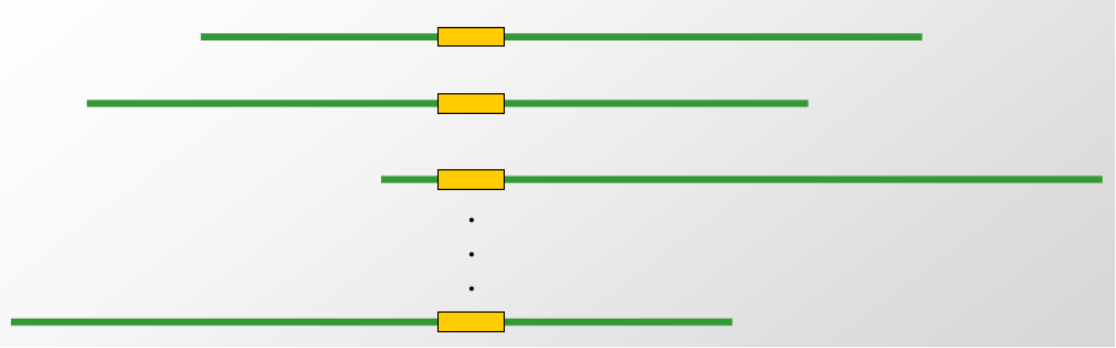
Stormo and Hartzell, Proc. Natl. Acad. Sci. (1989)



Alignment (PWM ~ sequence)

Exhaustive and prohibitive
Given a motif of width w, and k sequences of length l, there are L = (l-w+1) possible locations in each sequence, and L^k alignments to check





Algorithms for motif finding

Combinatorial

- Exhaustive search
- Greedy motif clustering

Probabilistic

- Expectation maximization
- Gibbs Sampling

Combinatorial (exhaustive)

Given a set of t DNA sequences, find a pattern that appears in all t sequences with the minimum number of mutations.

Hamming distance, $d_H(v,w)$ = number of mismatches between v and w

$d_H(AACA, ACCC) = 2$

- Given ***v*** = "acgtacgt" and **s**

v is the sequence in red, *x* is the sequence in blue

- $TotalDistance(\mathbf{v,DNA}) = 1+0+2+0+1 = 4$

Combinatorial (exhaustive)

1. Pattern-driven algorithm:

For W = AA…A to TT…T (4^K possibilities)

Find d(W, S)

Report W* = argmin(d(W, S))

Running time: O(K N 4^K)

(where N = $\Sigma_i |x^i|$)

d(W, S) is the hamming distance between a word and sequence

Hamming distance = number of mismatches

Advantage: Finds provably "best" motif W

Disadvantage: Time

Combinatorial (faster)

2. Sample-driven algorithm:

For W = any K-long word occurring in some x^i
Find d(W, S)

Report W^* = argmin(d(W, S))
or, Report a local improvement of W^*

Running time: $O(K N^2)$

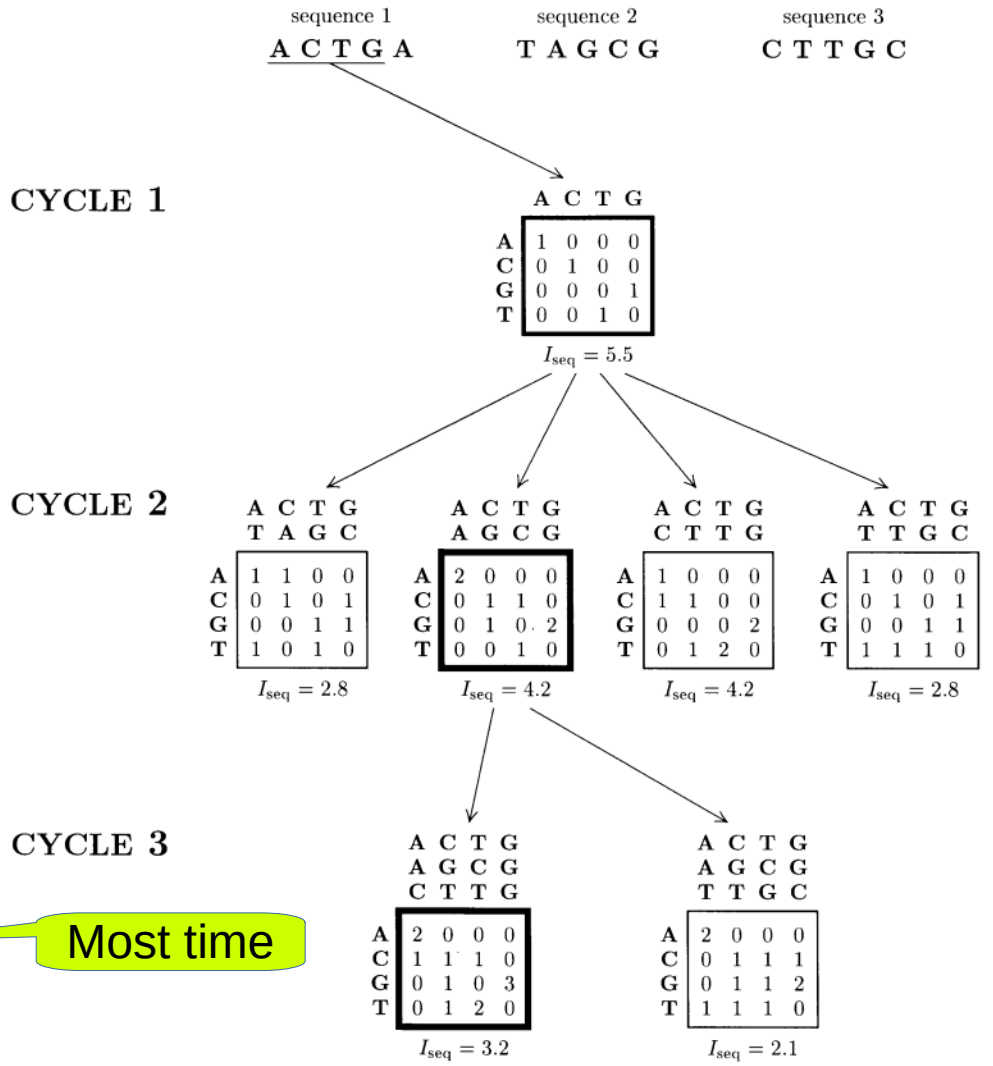
Advantage: Time

Disadvantage: If the true motif is weak and does not occur in data

then a random motif may score better than any instance of true motif

CONSENSUS (greedy) algorithm

```
GREEDYMOTIFSEARCH (DNA, t, n, l)
bestMotif := (1,…,1)
s := (1,…,1)
for s1=1 to n-l+1
    for s2 = 1 to n-l+1
        if (Score(s,2,DNA) > Score(bestMotif,2,DNA)
            bestMotif1 := s1
            bestMotif2 := s2
s1 := bestMotif1; s2 := bestMotif2
for i = 3 to t
    for si = 1 to n-l+1
        if (Score(s,i,DNA) > Score(bestMotif,i,DNA)
            bestMotifi := si
    si := bestMotifi
Return bestMotif
```



- Cycle 1: create word list from first two sequences
- Cycle 2: highest X scoring two-seq matrices are saved (X=6 saved)
- Cycle 3 to t: add words to alignment from 2 and save the highest scores.

One motif per sequence

Greedy Algorithms (Consensus)

Algorithm:

Cycle 1:

For each word W in S (of fixed length!)

    For each word W' in S

        Create alignment (gap free) of W, W'

Keep the C_1 best alignments, A_1, …, A_{C1}

ACGGTTG , CGAACTT , GGGCTCT …

ACGCCTG , AGAACTA , GGGGTGT …

Consensus (greedy) Algorithms

Algorithm:

Cycle t:
For each word W in S
    For each alignment A_j from cycle t-1
        Create alignment (gap free) of W, A_j

Keep the C_l best alignments A_1, …, A_{Ct}

ACGGTTG , CGAACTT , GGGCTCT ...
ACGCCTG , AGAACTA , GGGGTGT ...

Running time:

$O(N^2) + O(N\ C_1) + O(N\ C_2) + \ldots + O(N\ C_n)$

$= O(\ N^2 + NC_{total})$

Where $C_{total} = \Sigma_i\ C_i$, typically O(nC), where C is a big constant

1) A motif can be generate from any collection of aligned DNA or protein sequences (T/F)?

2) What is the maximum information content of the following degenerate motif, T[AG]A? As information content for a motif increases, the number of hits (above some cutoff) decreases (T/F)?

3) Footprinting tells you where a protein binds DNA (T/F)?

4) Gel shift can tell you whether a protein binds: a) probe DNA, b) competitor ‘cold’ DNA, c) both

5) Which motif has higher information content?

6) What is the probability of TGA with the following motif model?

$$PPM = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.2 & .5 & .1 \\ 0.3 & .1 & .1 \\ 0.3 & .2 & .1 \\ 0.2 & .2 & .7 \end{bmatrix}$$

7) Would you expect more or fewer matches to a high compared to a low information content motif in a genome?

8) Both the binding energy model and log-likelihood ratio model of a motif can be derived from the position probability matrix [T/F].

9) What is the advantage and disadvantgae of combinatorial motif search over greedy (consensus) search?