Motif Detection

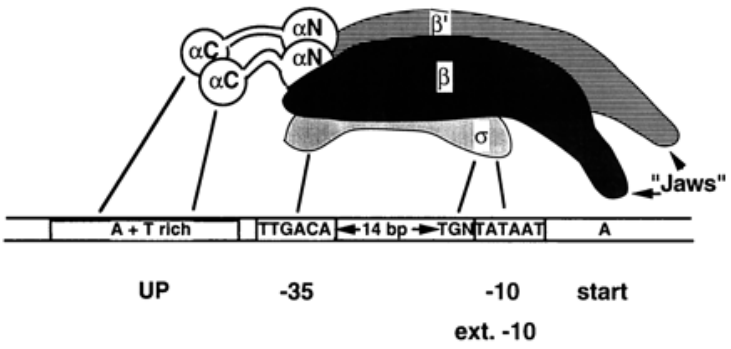
10-810, CMB lecture 6---Eric Xing

Motifs - Sites - Signals - Domains

- For this lecture, I'll use these terms interchangeably to describe **recurring elements** of interest to us.
- In **PROTEINS** we have: transmembrane domains, coiled-coil domains, EGF-like domains, signal peptides, phosphorylation sites, antigenic determinants, ...
- In **DNA / RNA** we have: enhancers, promoters, terminators, splicing signals, translation initiation sites, centromeres, ...

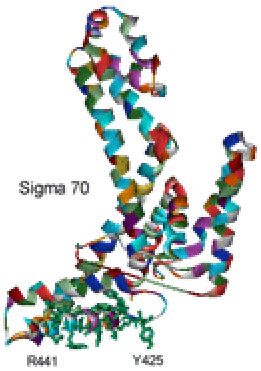
Transcription initiation in *E. coli*

RNA polymerase-promotor interactions

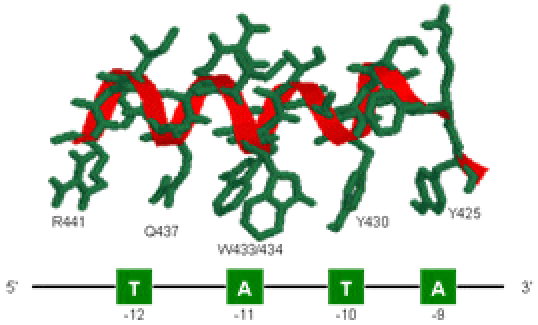


In *E. coli* transcription is initiated at the ***promotor*** , whose sequence is recognised by the ***Sigma factor*** of RNA polymerase.

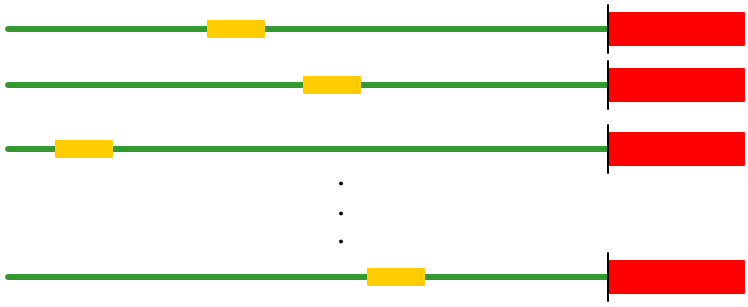
Protein motif: activity sites



..YKFSTYATWWIRQAITR..

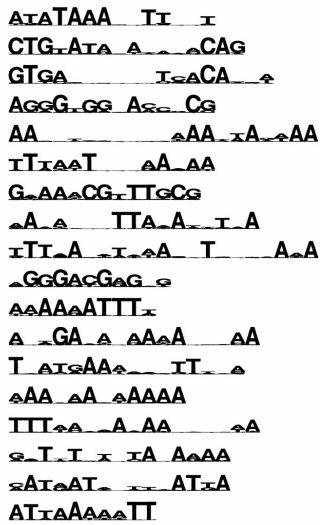


DNA motif: regulatory signals



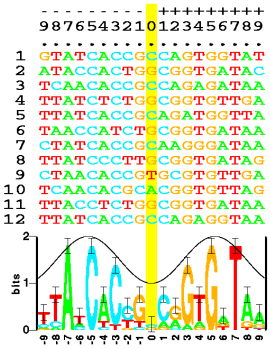
Given a collection of genes with common expression,
Can we find the TF-binding motif in common?

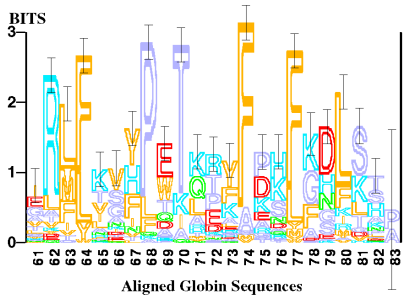
Characteristics of regulatory motifs



- Tiny
- Highly Variable
- ~Constant Size
  - Because a constant-size transcription factor binds
- Often repeated
- Low-complexity-ish

Sequence logo





- Information at pos'n I, $H(i) = -\Sigma_{\{\text{letter } x\}} \text{freq}(x, i) \log_2 \text{freq}(x, i)$
- Height of x at pos'n i, $L(x, i) = \text{freq}(x, i)\,(2 - H(i))$
  - Examples:
    - freq(A, i) = 1; H(i) = 0; L(A, i) = 2
    - A: ½; C: ¼; G: ¼; H(i) = 1.5; L(A, i) = ¼; L(not T, i) = ¼

Problem definition

Given a collection of promoter sequences $s_1, \ldots, s_N$ of genes with common expression

Combinatorial

Motif M: substring $m_1 \ldots m_W$

Some of the m_i's blank

- Find M that occurs in all s_i with ≤ k differences
- Or, Find M with smallest total hamming dist

Probabilistic

Motif M: $\{q_{ij}\}$; $1 \le i \le W$

$j = A,C,G,T$

q_{ij} = Prob[letter j, pos i]

Find best M, and positions $p_1, \ldots, p_N$ in sequences

Algorithms and models

- Combinatorial
  CONSENSUS, TEIRESIAS, SP-STAR, others
- Probabilistic

1. Expectation Maximization:
   e.g., MEME
2. Gibbs Sampling:
   e.g., AlignACE, BioProspector
3. Advanced models:
   Bayesian network, Bayesian Markovian models

Determinism 1: consensus sequences

σ Factor	Promotor ***consensus*** sequence	
	-35	-10
σ^{70}	TTGACA	TATAAT
σ^{28}	CTAAA	CCGATAT

Similarly for σ^{32}, σ^{38} and σ^{54}.

Consensus sequences have the obvious limitation: there is usually some deviation from them.

Determinism 2: regular expressions

The characteristic motif of a Cys-Cys-His-His zinc finger DNA binding domain has ***regular expression***

C-X(2,4)-C-X(3)-[LIVMFYWC]-X(8)-H-X(3,5)-H

Here, as in algebra, **X** is unknown. The 29 a.a. sequence of our example domain 1SP1 is as follows, clearly fitting the model.

1SP1: **KKFACPECPKRFMRSDHLSKHIKTHQNKK**

The (w,s) score ...

Regular expressions can be limiting

- The regular expression syntax is still too rigid to represent many highly divergent protein motifs.
- Also, short patterns are sometimes insufficient with today's large databases. Even requiring perfect matches you might find many false positives. On the other hand some real sites might not be perfect matches.
- We need to go beyond apparently equally likely alternatives, and ranges for gaps. We deal with the former first, having a distribution at each position.

Motif alignment

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …HIS7

5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …ARO4

5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …ILV6

5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …THR4

5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …ARO1

5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …HOM2

5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …PRO3

A=

```
1: AAAAGAGTCA
2: AAATGACTCA
.  AAGTGAGTCA
.  AAAAGAGTCA
.  GGATGAGTCA
.  AAATGAGTCA
.  GAATGAGTCA
M: AAAAGAGTCA
```

Weight matrix model (WMM)

Weight matrix model (WMM) = Stochastic consensus sequence

A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

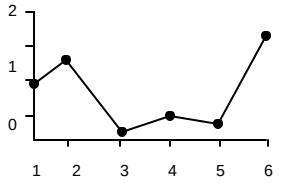
Counts from 242 known σ^{70} sites

A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.21	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

Relative frequencies: f_{bl}

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-40
T	17	-32	8	-9	-6	19

$10 \log_2 f_{bl}/p_b$

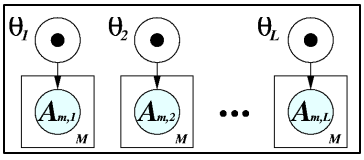


Informativeness: $2 - \sum_b p_{bl} \log_2 p_{bl}$

The product multinomial (PM) model

- Positional specific multinomial distribution (PSMD): $q_l = [q_{lA}, \ldots, q_{lC}]^T$
- Position weight matrix (PWM):

	1	2	3	4	5	6	7	8	9	10
A	.750	.875	.875	.375	0	1	0	0	0	1
C	0	0	0	0	0	0	.125	0	1	0
G	.250	.125	.125	0	1	0	.875	0	0	0
T	0	0	0	.635	0	0	0	1	0	0



The nucleotide distributions at different sites are independent !

The **score** (likelihood-ratio) of a candidate substring: **AAAAGAGTCA**

$$R = \frac{p(x = \{\text{AAAAGAGTCA}\} \mid \text{PWM})}{p(x = \{\text{AAAAGAGTCA}\} \mid \text{bk})} = \prod_{l=1}^{10} \frac{p(y_l \mid \text{PWM})}{p(y_l \mid \text{bk})} = \prod_{l=1}^{10} \frac{q_{l,y_l}}{q_{0,y_l}}$$

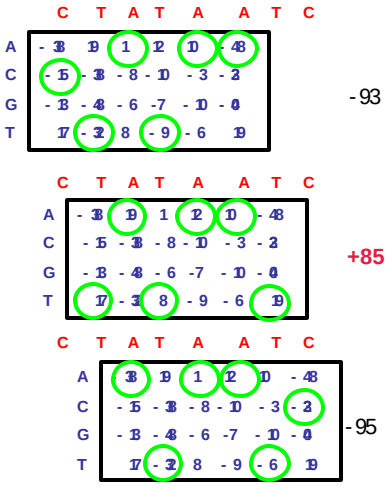
Use of the matrix to find sites

Hypothesis:
S=site (and independence)
R=random (equiprobable, independence)

Move the matrix along the sequence and score each **window.**

Peaks should occur at the true sites.

Of course in general any threshold will have some **false positive** and **false negative** rate.



Supervised motif search

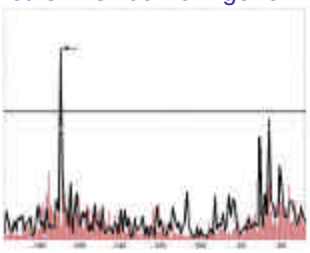
- **Supervised learning**

Given biologically identified **A**, maximal likelihood estimation:

$$\Theta_{ML} = \arg\max_{\Theta} p(\mathbf{A} \mid \Theta)$$

Application:

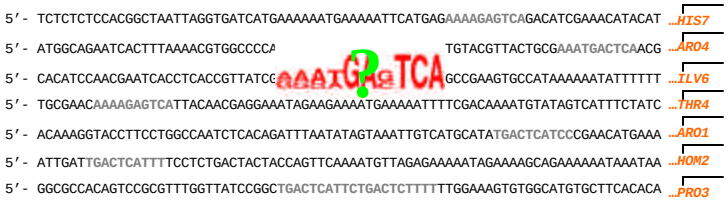
search for **known** motifs *in silico* from genomic sequences



de novo motif detection

- **Unsupervised learning**

Given no training examples, predict locations of all instances of ***novel*** motifs in given sequences, and learn motif models simultaneously.



Learning algorithms: EM, Gibbs sampling

Application:

identify **unknown** motifs *in silico* from genomic sequences

de novo motif detection

Problem setting:

Given UTR sequences: $y = \{y_1, \ldots, y_N\}$

Goal: the background model: $q_0 = \{q_{0,A}, q_{0,T}, q_{0,G}, q_{0,C}\}^t$

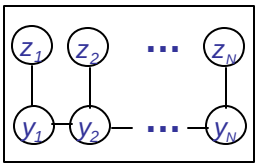
and K motif models $q^1, \ldots, q^K$ from y,

where $q^k = \{q^k_{i,j} : i = 1,\ldots,L_k,\ j \in \{A, C, G, T\}\}$

A missing value problem:

The locations of instances of motifs are unknown, thus the aligned motif sequences $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^K$ and the background sequence are not available.

A binary indicator model



$Z \in \{0, 1\}^N$

Let:

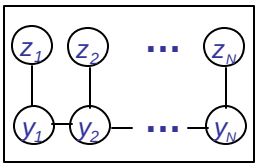
$Y_n = \{y_n, y_{n+1}, \ldots, y_{n+L-1}\}$: an L-long word starting at position n

$p(z_n = 1) = e, \quad p(z_n = 0) = 1 - e$

$$p(Y_n \mid z_n = 0) = q_{0,y_i} q_{0,y_{n+1}}, \ldots, q_{0,y_{n+L-1}} = \prod_{l=0}^{L-1} q_{0,y_{n+l}} = \prod_{l=1}^{L} \prod_{j=1}^{4} q_{0,j}^{d(y_{n+l-1}, j)} \quad \text{(background)}$$

$$p(Y_n \mid z_n = 1) = q_{1,y_n} q_{2,y_{n+1}}, \ldots, q_{L,y_{n+L-1}} = \prod_{l=1}^{L} q_{l,y_{n+l-1}} = \prod_{l=1}^{L} \prod_{j=1}^{4} q_{l,j}^{d(y_{n+l-1}, j)} \quad \text{(motif seq.)}$$

A binary indicator model



$Z \in \{0, 1\}^N$

Complete log-likelihood :

(suppose all sequences are concatenated into one big sequence of $y=y_1y_2\ldots y_N$, with appropriate constraints preventing overlapping and boundaries limits)

$$p(y_n, z_n) = p(y_n \mid z_n)\,p(z_n) = p(y_i \mid q_0)^{1-z_n}\, p(y_n \mid q)^{z_n} \times (1-e)^{1-z_n} e^{z_n}$$

$$l_c(\Theta) = \sum_{n=1}^{N} z_n \left(\sum_{l=1}^{L} \sum_{j=1}^{4} d(y_{n+l-1}, j) \log q_{l,j} \right) + \sum_{n=1}^{N} (1 - z_n) \left(\sum_{j=1}^{4} d(y_{n+l-1}, j) \log q_{0,j} \right)$$
$$+ |z| \log e + (N - |z|) \log(1-e)$$

Expectation Maximization

Maximize expected likelihood, in iteration of two steps:

Expectation:

Find expected value of complete log likelihood:

$$E[\log P(X_1 \ldots X_n, Z \mid q, q_0, e)]$$

Maximization:

Maximize expected value over q, q_0, e

Expectation Maximization: E-step

Expectation:

Find expected value of log likelihood:

$$\langle l_c(\Theta)\rangle = \sum_{n=1}^{N}\langle z_n\rangle\left(\sum_{l=1}^{L}\sum_{j=1}^{4} d(y_{n+l-1}, j)\log q_{l,j}\right) + \sum_{n=1}^{N}(1-\langle z_n\rangle)\left(\sum_{j=1}^{4} d(y_{n+l-1}, j)\log q_{0,j}\right)$$

$$+\sum_{n=1}^{N}\langle z_n\rangle\log e + \left(N-\sum_{n=1}^{N}\langle z_n\rangle\right)\log(1-e)$$

where expected values of Z can be computed as follows:

$$\langle z_i\rangle = p(z_i = 1\,|\,Y) = \frac{ep(X_i\,|\,q)}{ep(X_i\,|\,q)+(1-e)\,p(X_i\,|\,q_0)}$$

Expectation Maximization: M-step

Maximization:

Maximize expected value over θ and λ independently

For e, this is easy:

$$e^{NEW} = \arg\max_{e}\sum_{n=1}^{N}\langle z_n\rangle\log e + \left(N-\sum_{n=i}^{N}\langle z_i\rangle\right)\log(1-e) = \sum_{n=1}^{N}\frac{\langle z_n\rangle}{N}$$

Expectation Maximization: M-step

For $Q = (q, q_0)$, define

$c_{l,j}$ = E[# times letter *j* appears in motif position *l*]
$c_{0,j}$ = E[# times letter *j* appears in background]

- $c_{l,j}$ values are calculated easily from E[Z] values

It easily follows:

$$q_{l,j}^{NEW} = \frac{c_{l,j}}{\sum_{j=1}^{4} c_{l,j}} \qquad q_j^{NEW} = \frac{c_{0,j}}{\sum_{j=1}^{4} c_{0,j}}$$

to not allow any 0's, add pseudocounts

Initial Parameters Matter!

Consider the following "artificial" example:

$x^1, \ldots, x^N$ contain:

- 2^{12} patterns on {A, T}: **A…AA, A…AT, ……, T…TT**
- 2^{12} patterns on {C, G}: **C…CC, C…CG, ……, G…GG**
- D << 2^{12} occurrences of 12-mer ACTGACTGACTG

Some local maxima:

$e \approx$ ½; B = ½C, ½G; M_i = ½A, ½T, i = 1,…, 12

$e \approx D/2^{L+1}$; B = ¼A,¼C,¼G,¼T;
M_1 = 100% A, M_2= 100% C, M_3 = 100% T, etc.

Overview of EM Algorithm

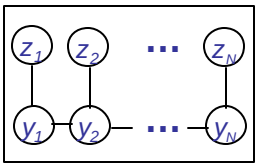
1. Initialize parameters $Q = (q, q_0)$, :
   – Try different values of e from $N^{-1/2}$ up to 1/(2L)
2. Repeat:
   a. Expectation
   b. Maximization
3. Until change in $Q = (q, q_0)$, falls below d
4. Report results for several "good" e

Overview of EM Algorithm

- One iteration running time: O(NL)
  – Usually need < N iterations for convergence, and < N starting points.
  – Overall complexity: unclear – typically $O(N^2L)$ - $O(N^3L)$
- EM is a local optimization method
- Initial parameters matter

MEME: Bailey and Elkan, ISMB 1994.

Any problem with the model?



$Z \in \{0, 1\}^N$

- How about multiple types of motifs?
- Can motif overlap?
- Are motif sites independent?