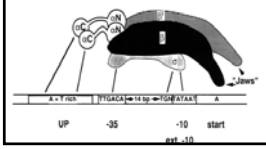
Computational Genomics

10-810/02-710, Spring 2009

Motif Detection

Eric Xing



Lecture 8, February 4-9, 2009

Reading: Durbin Chap 9, class assignment

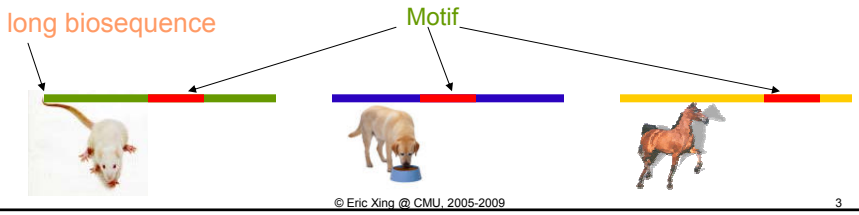
Motifs - Sites - Signals - Domains

- 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, ...

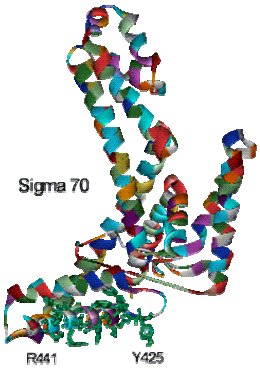
Motif

- Set of similar substrings,
  - within a single long sequence

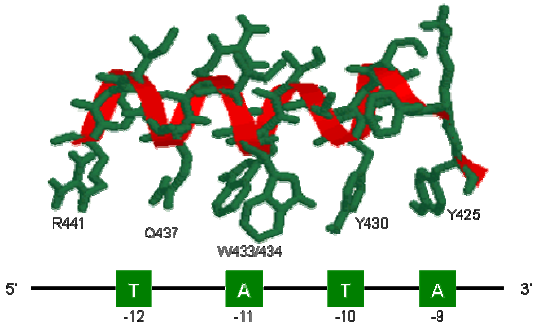
  - or a family of diverged sequences



Protein Motif: Activity Sites



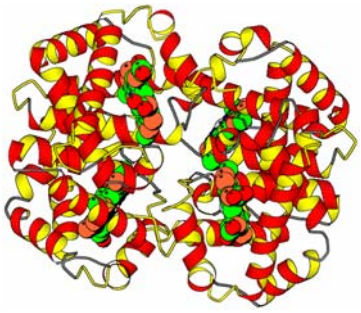
..YKFSTYATWWIRQAITR..



Example: Globin Motifs

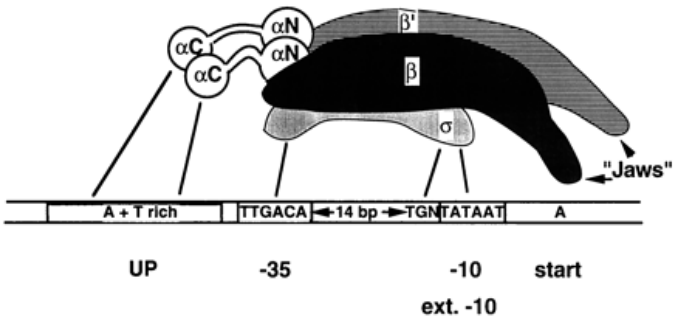
```
          xxxxxxxxxx.xxxxxxxx.xxxxx.........xxxxxx.xxxxxx.xxxxxxxxx.xxxxxxxx
HAHU    V.LSPADKTN..VKAAWGKVG.AHAGE.........YGAEAL.ERMFLSF..PTTKTYFPH.FDLS.HGSA
HAOR    M.LTDAEKKE..VTALWGKAA.GHGEE.........YGAEAL.ERLFQAF..PTTKTYFSH.FDLS.HGSA
HADK    V.LSAADKTN..VKGVFSKIG.GHAEE.........YGAETL.ERMFIAY..PQTKTYFPH.FDLS.HGSA
HBHU    VHLTPEEKSA..VTALWGKVN.VDEVG..........G.EAL.GRLLVVY..PWTQRFFES.FGDL.STPD
HBOR    VHLSGGEKSA..VTNLWGKVN.INELG..........G.EAL.GRLLVVY..PWTQRFFEA.FGDL.SSAG
HBDK    VHWTAEEKQL..ITGLWGKVNvAD.CG..........A.EAL.ARLLIVY..PWTQRFFAS.FGNL.SSPT
MYHU    G.LSDGEWQL..VLNVWGKVE.ADIPG.........HGQEVL.IRLFKGH..PETLEKFDK.FKHL.KSED
MYOR    G.LSDGEWQL..VLKVWGKVE.GDLPG.........HGQEVL.IRLFKTH..PETLEKFDK.FKGL.KTED
IGLOB   M.KFFAVLALCiVGAIASPLT.ADEASlvqsswkavsHNEVEIlAAVFAAY.PDIQNKFSQFaGKDLASIKD
GPUGNI  A.LTEKQEAL..LKQSWEVLK.QNIPA.........HS.LRL.FALIIEA.APESKYVFSF.LKDSNEIPE
GPYL    GVLTDVQVAL..VKSSFEEFN.ANIPK..........N.THR.FFTLVLEiAPGAKDLFSF.LKGSSEVPQ
GGZLB   M.L.DQQTIN..IIKATVPVLkEHGVT..........ITTTF.YKNLFAK.HPEVRPLFDM.GRQ..ESLE

          xxxxx.xxxxxxxxxxxx..xxxxxxxxxxxxxxx..xxxxxxx.xxxxxxx...xxxxxxxxxxxxxxx
HAHU    QVKGH.GKKVADA.LTN.....AVA.HVDDMPNA...LSALS.D.LHAHKL....RVDPVNF.KLLSHCLL
HAOR    QIKAH.GKKVADA.L.S.....TAAGHFDDMDSA...LSALS.D.LHAHKL....RVDPVNF.KLLAHCIL
HADK    QIKAH.GKKVAAA.LVE.....AVN.HVDDIAGA...LSKLS.D.LHAQKL....RVDPVNF.KFLGHCFL
HBHU    AVMGNpKVKAHGK.KVLGA..FSDGLAHLDNLKGT...FATLS.E.LHCDKL....HVDPENF.RL.LGNVL
HBOR    AVMGNpKVKAHGA.KVLTS..FGDALKNLDDLKGT...FAKLS.E.LHCDKL....HVDPENFNRL..GNVL
HBDK    AILGNpMVRAHGK.KVLTS..FGDAVKNLDNIKNT...FAQLS.E.LHCDKL....HVDPENF.RL.LGDIL
MYHU    EMKASeDLKKHGA.TVL.....TALGGILKKKGHH..EAEIKPL.AQSHATK...HKIPVKYLEFISECII
MYOR    EMKASaDLKKHGG.TVL.....TALGNILKKKGQH..EAELKPL.AQSHATK...HKISIKFLEYISEAII
IGLOB   T.GA...FATHATRIVSFLseVIALSGNTSNAAAV...NSLVSKL.GDDHKA....R.GVSAA.QF..GEFR
GPUGNI  NNPK...LKAHAAVIFKTI...CESATELRQKGHAVwdNNTLKRL.GSIHLK....N.KITDP.HF.EVMKG
GPYL    NNPD...LQAHAG.KVFKL..TYEAAIQLEVNGAVAs.DATLKSL.GSVHVS....K.GVVDA.HF.PVVKE
GGZLB   Q......PKALAM.TVL.....AAAQNIENLPAIL..PAVKKIAVKHCQAGVaaaH.YPIVGQELLGAIK

          xxxxxxxxx.xxxxxxxxx.xxxxxxxxxxxxxxxxxxxxxxx..x
HAHU    VT.LAA.H..LPAEFTPA..VHASLDKFLASV.STVLTS..KY..R
HAOR    VV.LAR.H..CPGEFTPS..AHAAMDKFLSKV.ATVLTS..KY..R
HADK    VV.VAI.H..HPAALTPE..VHASLDKFMCAV.GAVLTA..KY..R
HBHU    VCVLAH.H..FGKEFTPP..VQAAYQKVVAGV.ANALAH..KY..H
HBOR    IVVLAR.H..FSKDFSPE..VQAAWQKLVSGV.AHALGH..KY..H
HBDK    IIVLAA.H..FTKDFTPE..CQAAWQKLVRVV.AHALAR..KY..H
MYHU    QV.LQSKHPgDFGADAQGA.MNKALELFRKDM.ASNYKELGFQ..G
MYOR    HV.LQSKHSaDFGADAQAA.MGKALELFRNDM.AAKYKEFGFQ..G
IGLOB   TA.LVA.V...QANVSWGDnVAAAWNKA.LDN.TFAIVV..PR..L
GPUGNI  ALLSTIKEA.IKENWSDE..MGQAWTEAYNQLVATIKAE..MK..E
GPYL    AILKTIKEV.VGDKWSEE..LNTAWTIAYDELAIIIKKE..MKdaA
GGZLB   EVLGDAAT..DDILDAKGK.AYGVIADVFIQVEADLYAQ..AV..E
```

Hemoglobin alpha subunit



DNA Motif

- **RNA polymerase-promotor interactions**
  - Transcription Initiation in *E. coli*



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

Example: Gcn4

Regulatory Signals

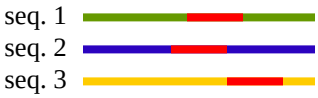
```
5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …HIS7
5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …ARO4
5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …ILV6
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …THR4
5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …ARO1
5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …HOM2
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …PRO3
```

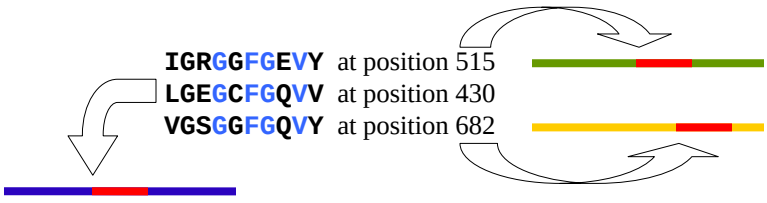
Given a collection of genes with common expression,
Can we find the TF-binding site in common?

Motif discovery problem

- Given sequences

- Find motif
  - the number of motifs
  - the width of each motif
  - the locations of motif occurrences
  - and a "model" for evaluating a motif candidate …



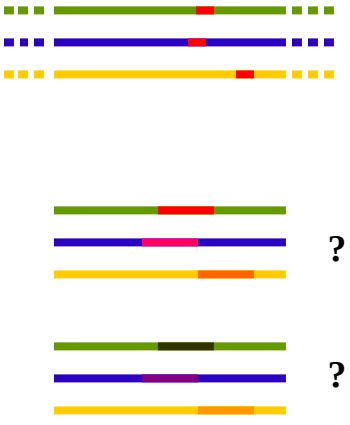


Why finding motifs?

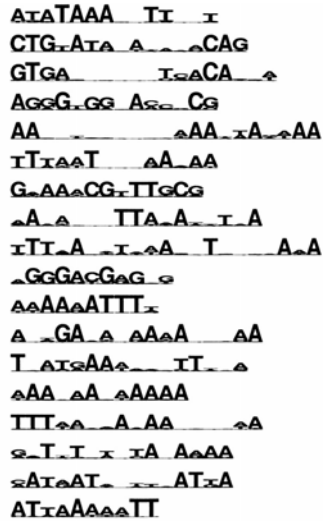
- In proteins—may be a critical component
  - Find similarities to known proteins
  - Find important areas of new protein family
  - Prediction of protein function
- In DNA—may be a *binding site*
  - Discover how the gene expression is regulated
  - Prediction of temporal/spatial gene activity
  - Inference of regulatory network

Why is this hard?

- Input sequences are long (thousands or millions of residues)
- Motif may be *subtle*
  - Instances are short.
  - Instances may be only slightly similar.



Characteristics of Regulatory Motifs

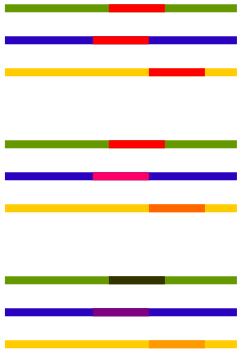


- Tiny
- Highly Variable
- ~Constant Size
  - Because a constant-size transcription factor binds
- Often repeated
- Low-complexity-ish

Motif Representation

Measuring similarity

- What counts as a similarity?
- How can such a pattern be searched for?
- Need a *concrete measure* of how good a *motif* is, and how well-matched an *instance* is.



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***

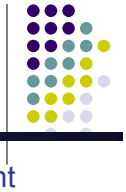
  C-X(2,4)-C-X(3)-[LIVMFYWC]-X(8)-H-X(3,5)-H

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

  1SP1:

  KKFACPECPKRFMRSDHLSKHIKTHQNKK

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.

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	.04	.88	.26	.59	.49	.03
C	.09	.03	.11	.13	.21	.05
G	.07	.01	.12	.16	.12	.02
T	.80	.08	.51	.13	.18	.89

Relative frequencies: θ_{li}

less interesting

more interesting

- Weight matrices are also known as
  - Position-specific scoring matrices
  - Position-specific probability matrices
  - Position-specific weight matrices
- A motif is *interesting* if it is very different from the background distribution

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

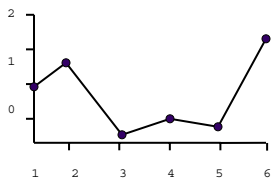
Counts from 242 known σ^{70} sites

A	.04	.88	.26	.59	.49	.03
C	.09	.03	.11	.13	.21	.05
G	.07	.01	.12	.16	.12	.02
T	.80	.08	.51	.13	.18	.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

$10 \log_2 \theta_{li} / \theta_{0i}$



Informativeness: $2\text{-}\Sigma_i \theta_{li} \log_2 \theta_{li} / \theta_{0i}$

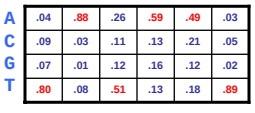
Relative entropy

- A motif is *interesting* if it is very different from the background distribution
- Use relative entropy:

$$\sum_{\text{position } l}\left(\sum_{\text{letter } i}\theta_{li}\log_2\frac{\theta_{li}}{\theta_{0i}}\right)$$

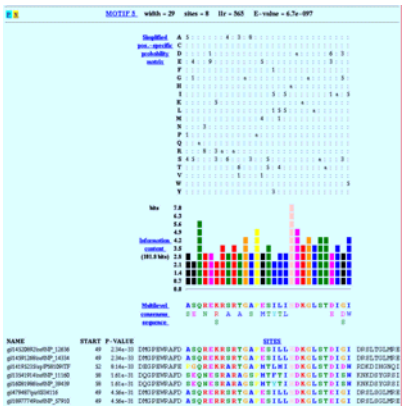
θ_{li} = probability of i in matrix position l
θ_{0i} = background frequency (in non-motif sequence)

  - Relative entropy is sometimes called information content

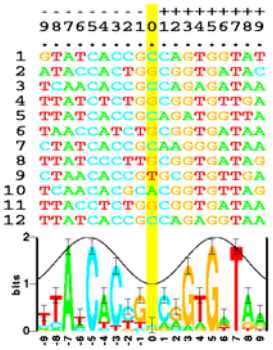


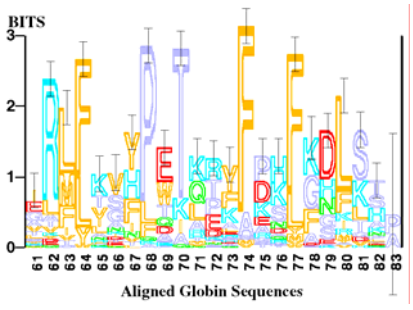
A	.04	.88	.26	.59	.49	.03
C	.09	.03	.11	.13	.21	.05
G	.07	.01	.12	.16	.12	.02
T	.80	.08	.51	.13	.18	.89

Informativeness: $2\text{-}\Sigma_i\theta_{li}\log_2\theta_{li}/\theta_{0i}$



Sequence Logo

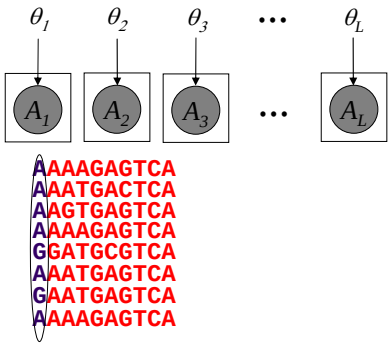




- Information at pos'n l, $H(l) = -\Sigma_{\{\text{letter } i\}}\,\theta_{li}\log_2\theta_{li}$
- Height of x at pos'n l, $L(l,i) = \theta_{li}(2 - H(l))$
  - Examples:
    - $\theta_{lA} = 1$; H(l) = 0; L(l, A) = 2
    - A: ½; C: ¼; G: ¼; H(l) = 1.5; L(l, A) = ¼; L(l, not T) = ¼

The Product Multinomial (PM) Model [Lawrence *et al*. Science 1993]

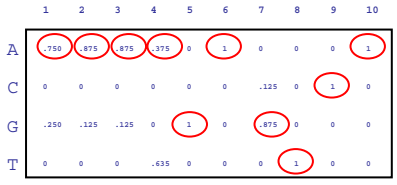
- Positional specific multinomial distribution: $\theta_l = [\theta_{lA}, \ldots, \theta_{lC}]^T$

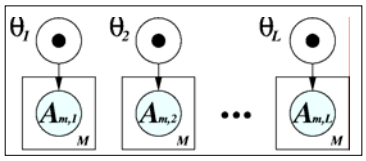


- Position weight matrix (PWM): θ
  - The nucleotide distributions at different positions are independent

More on PM Model

- The PM parameter, $\theta_l = [\theta_{lA}, \ldots, \theta_{lC}]^T$, corresponds exactly to the PWM of a motif





The nucleotide distributions at different sites are independent !

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

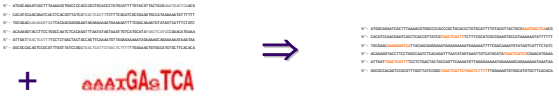
$$R = \frac{p(x=\{\text{AAAAGAGTCA}\}\,|\,\text{PWM})}{p(x=\{\text{AAAAGAGTCA}\}\,|\,\text{bk})} = \prod_{l=1}^{10} \frac{p(y_l\,|\,\text{PWM})}{p(y_l\,|\,\text{bk})} = \prod_{l=1}^{10} \frac{\theta_{l,y_l}}{\theta_{0,y_l}}$$

- Log Likelihood-Ratio: $\text{LR} = \sum_{\text{position } l} \left(\sum_{\text{letter } i} \log_2 \frac{\theta_{l,y_l}}{\theta_{0,y_l}} \right)$

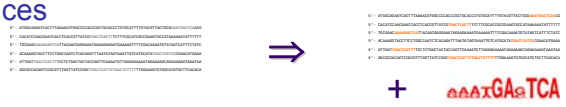
Computational problems for *in silico* motif detection

- Extract a motif model based on (experimentally) identified motifs

- Search for motif instances based on given motif model(s)



- Uncover novel motifs computationally from genomic sequences



de novo motif detection

Computational problems for *in silico* motif detection

- Extract a motif model based on (experimentally) identified motifs

Supervised learning

- Search for motif instances based on given motif model(s)

Prediction

- Uncover novel motifs computationally from genomic sequences

Unsupervised learning

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 $\leq k$ differences
- Or, Find M with smallest total hamming dist

Probabilistic

Motif M: $\{\theta_{ij}\}$; $1 \leq i \leq W$

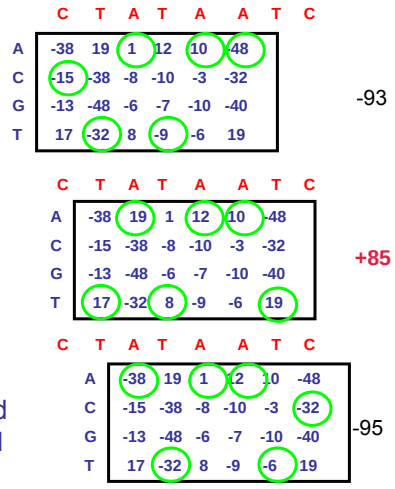
j = A,C,G,T

θ_{ij} = Prob[letter j, pos i]

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

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.

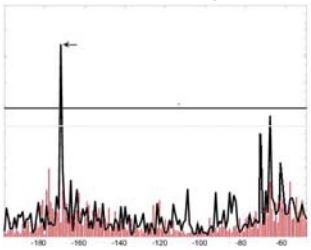


Supervised motif search

- **Supervised learning**
  - Given biologically identified alinged motifs **A**, maximal likelihood estimation:

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

- **Application:**
  - search for **known** motifs *in silico* from genomic sequences



- Need more more sophisticated search model: HMM?

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**:
  - Expectation Maximization: e.g., MEME
  - Gibbs Sampling: e.g., AlignACE, BioProspector
  - Advanced models: Bayesian network, Bayesian Markovian models

de novo motif detection

- **Problem setting:**
  - Given UTR sequences: $y = \{y_1, \ldots, y_N\}$
  - Goal: the background model: $\theta_0 = \{\theta_{0,A}, \theta_{0,T}, \theta_{0,G}, \theta_{0,C}\}t$

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

    where $\theta^k = \left\{\theta^k_{i,j} : i = 1,\ldots,L_k,\ j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}\right\}$
- **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.

Expectation-maximization

- The EM idea:

EM

For each subsequence of width W
  convert subsequence to a matrix
  do {
    re-estimate motif occurrences from matrix
    re-estimate matrix model from motif occurrences
  } until (matrix model stops changing)
end
select matrix with highest score

Sample DNA sequences

```
>ce1cg
TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATA
GCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTCAC
AAAAATGGAAGTCCACAGTCTTGACAG
>ara
GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAG
AAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTG
CTATGCCATAGCATTTTTATCCATAAG
>bglr1
ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATA
TAACTTTATAAATTCCTAAAATTACACAAAGTTAATAAC
TGTGAGCATGGTCATATTTTTATCAAT
>crp
CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTAC
AGTAATACATTGATGTACTGCATGTATGCAAAGGACGTC
ACATTACCGTGCAGTACAGTTGATAGC
```


Motif occurrences

```
>ce1cg
taatgtttgtgctggttttgtggcatcgggcgagaata
gcgcgtggtgtgaaagactgtttt[TTTGATCGTTTTCAC]
aaaaatggaagtccacagtcttgacag
>ara
gacaaaaacgcgtaacaaaagtgtctataatcacggcag
aaaagtccacattgatta[TTTGCACGGCGTCAC]actttg
ctatgccatagcatttttatccataag
>bglr1
acaaatcccaataacttaattattgggatttgttatata
taactttataaattcctaaaattacacaaagttaataac
[TGTGAGCATGGTCAT]atttttatcaat
>crp
cacaaagcgaaagctatgctaaaacagtcaggatgctac
agtaatacattgatgtactgcatgta[TGCAAAGGACGTC
AC]attaccgtgcagtacagttgatagc
```

One-occurrence per sequence (oops) model

Starting point

…gactgttttTTTGATCGTTTTCACaaaaatgg…

Or

	T	T	T	G	A	T	C	G	T	T
A	0.17	0.17	0.17	0.17	0.50	...				
C	0.17	0.17	0.17	0.17	0.17					
G	0.17	0.17	0.17	0.50	0.17					
T	0.50	0.50	0.50	0.17	0.17					

This a special initialization scheme, many others scheme, including random starts, are also valid

Re-estimating motif occurrences

TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATA

	T	T	T	G	A	T	C	G	T	T
A	0.17	**0.17**	**0.17**	0.17	0.50	...				
C	0.17	0.17	0.17	0.17	0.17					
G	0.17	0.17	0.17	0.50	**0.17**					
T	**0.50**	0.50	0.50	**0.17**	0.17					

Score = 0.50 + 0.17 + 0.17 + 0.17 + 0.17 + ...

Scoring each subsequence

- Score from each sequence the subsequence with maximal score.

```
Sequence: TGTGCTGGTTTTTGTGGCATCGGGCGAGAATA

          Subsequences      Score
        TGTGCTGGTTTTTGT      2.95
         GTGCTGGTTTTTGTG     4.62
          TGCTGGTTTTTGTGG    2.31
           GCTGGTTTTTGTGGC     ...
```


Re-estimating motif matrix

- Under a one-occurrence per sequence (oops) model
- From each sequence, take the substring that has the maximal score
- Align all of them and count:

```
  Occurrences                Counts
TTTGATCGTTTTCAC        A 000132011000040
TTTGCACGGCGTCAC   ---->  C 001010300200403
TGTGAGCATGGTCAT        G 020301131130000
TGCAAAGGACGTCAC        T 423001002114001
```

Adding pseudocounts

```
      Counts
A 000132011000040
C 001010300200403
G 020301131130000
T 423001002114001
```

```
Counts + Pseudocounts
A 111243122111151
C 112121411311514
G 131412242241111
T 534112113225112
```


Converting to frequencies

```
Counts + Pseudocounts
A 111243122111151
C 112121411311514
G 131412242241111
T 534112113225112
```

```
      T    T    T    G    A    T  C  G  T  T
  A 0.13 0.13 0.13 0.25 0.50 ...
  C 0.13 0.13 0.25 0.13 0.25
  G 0.13 0.38 0.13 0.50 0.13
  T 0.63 0.38 0.50 0.13 0.13
```

Expectation-maximization

- **Problem of the previous procedure:**
  - This procedure doesn't allow the motifs to move around very much. Taking the max is too brittle.

EM

```
For each subsequence of width W
   convert subsequence to a matrix
   do {
      re-estimate motif occurrences from matrix
      re-estimate matrix model from motif occurrences
   } until (matrix model stops changing)
end
select matrix with highest score
```

- **Solution:**
  - Associate with each start site a probability of motif occurrence.

Converting to probabilities

Sequence: TGTGCTGGTTTTTGTGGCATCGGGCGAGAATA

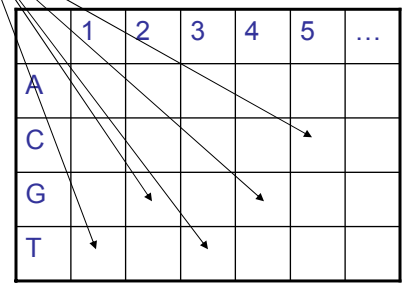
Occurrences	Score	Prob
TGTGCTGGTTTTTGT	2.95	0.023
GTGCTGGTTTTTGTG	4.62	0.037
TGCTGGTTTTTGTGG	2.31	0.018
GCTGGTTTTTGTGGC	...	...
Total	128.2	1.000

$$p(L-\text{mer}|\theta)$$

Computing weighted counts

Occurrences	Prob
TGTGCTGGTTTTTGT	0.023
GTGCTGGTTTTTGTG	0.037
TGCTGGTTTTTGTGG	0.018
GCTGGTTTTTGTGGC	...

Include counts from all subsequences, weighted by the prob to which they match the motif model.



A comparison of supervised and de nove motif finding

5′- TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT …HIS7
5′- ATGGCAGAATCACTTTAAAACGTGGCCCCACCCGCTGCACCCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG …ARO4
5′- CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT …ILV6
5′- TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC …THR4
5′- ACAAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCCGAACATGAAA …ARO1
5′- ATTGATTGACTCATTTTCCTCTGACTACTACCAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAAATAAATAA …HOM2
5′- GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGGAAAGTGTGGCATGTGCTTCACACA …PRO3

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

⇓

Q. and A.

- **Problem**: How do we estimate counts accurately when we have only a few examples?
  - **Solution**: Use Dirichlet mixture priors.
- **Problem**: Too many possible starting points.
  - **Solution**: Save time by running only one iteration of EM.
- **Problem**: Too many possible widths.
  - **Solution**: Consider widths that vary by $\sqrt{2}$ and adjust motifs afterwards.
- **Problem**: Algorithm assumes exactly one motif occurrence per sequence.
  - **Solution**: Normalize motif occurrence probabilities across all sequences, using a user-specified parameter.

Q. and A.

- **Problem**: The EM algorithm finds only one motif.
  - **Solution**: Probabilistically erase the motif from the data set, and repeat.
- **Problem**: The motif model is too simplistic.
  - **Solution**: Use a two-component mixture model that captures the background distribution. Allow the background model to be more complex.
- **Problem**: The EM algorithm does not tell you how many motifs there are.
  - **Solution**: Compute statistical significance of motifs and stop when they are no longer significant.

MEME algorithm

```
do
    for (width = min; width *= √2; width < max)
        foreach possible starting point
            run 1 iteration of EM
        select candidate starting points
        foreach candidate
            run EM to convergence
    select best motif
    erase motif occurrences
until (E-value of found motif > threshold)
```

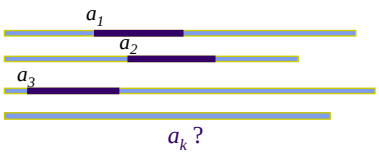
Is this a true EM algorithm?

The Gibbs Motif Sampler

- Initialized by choosing random starting positions

$$a_1^{(0)}, a_2^{(0)}, \ldots\ldots, a_K^{(0)}$$

- Iterate the following steps many times:
  - Randomly or systematically choose a sequence, say, ***sequence k***, to exclude.
  - Carry out the ***predictive-updating*** step to update a_k



(no need to sample $\theta^{(t)}$ at each t, we can compute it in close-form, see next lecture

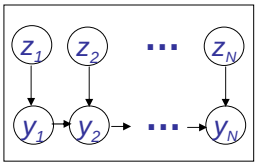
  - Notations:

$$A_{[-k]} \equiv \{a_1, \cdots, a_{k-1}, a_{k+1}, \cdots, a_K\} \Rightarrow c_{l,j,[-k]},\ l = 1 \cdots, W,\ j = \mathrm{A,T,G,C}$$
$$a_k \Rightarrow c_{l,j}(a_k)$$

- Stop when not much change observed, or some criterion met.

What is underlying the EM algorithm? – the statistical foundation

- **A binary indicator model** (you will find this model in AlignACE)



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

Let:

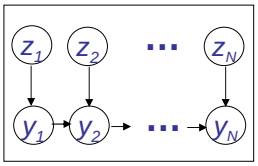
$Y_{n...n+L-1}=\{y_n, y_{n+1}, \ldots, y_{n+L-1}\}$: an L-long word starting at position n

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

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

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

A binary indicator model



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

- Complete log-likelihood :
  - suppose all words are concatenated into one big sequence of $y=y_1y_2\ldots y_N$, with appropriate constraints preventing overlapping and boundaries limits

$$p(y_n, z_n) = p(y_n \mid z_n)p(z_n) = p(y_i \mid \theta_0)^{1-z_n} p(y_n \mid \theta)^{z_n} \times (1-\varepsilon)^{1-z_n} \varepsilon^{z_n}$$

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

The Maximal likelihood approach

- Maximize expected likelihood, in iteration of two steps:

Expectation:

Find expected value of complete log likelihood:

$$\mathrm{E}[\log P(Y_1 \ldots Y_n, Z \mid \theta, \theta_0, \varepsilon)]$$

Maximization:

Maximize the expected complete likelihood over θ, θ_0, ε

Expectation Maximization: E-step

- **Expectation:**
  Find expected value of log likelihood:

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

- where the expected value of Z can be computed as follows:

$$\langle z_i \rangle = p(z_i = 1 \mid Y) = \frac{\varepsilon p(y_i \mid \theta)}{\varepsilon p(y_i \mid \theta) + (1 - \varepsilon) p(y_i \mid \theta_0)}$$

  - recall the weights for each substring in the MEME algorithm

Expectation Maximization: M-step

- **Maximization:**

  Maximize expected value over θ and ε independently

  For ε, this is easy:

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

Expectation Maximization: M-step

- For $\Theta = (\theta, \theta_0)$, 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:

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

  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:

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

  - very bad !

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

  - the correct solution !

Overview of EM Algorithm

1. Initialize parameters $\Theta = (\theta, \theta_0)$, :
   - Try different values of ε, say, from $N^{-1/2}$ up to 1/(2L)
2. Repeat:
   a. Expectation
   b. Maximization
3. Until change in $\Theta = (\theta, \theta_0)$, falls below δ
4. Report results for several "good" ε

Overview of EM Algorithm

- One iteration running time: O(NL)
  - Usually need < N iterations for convergence, and < N starting points.
  - Overall complexity: unclear
- EM is a local optimization method
- Initial parameters matter
- MEME: Bailey and Elkan, ISMB 1994.

Overview of Sampling Algorithm

- One iteration running time: O(NL)
- More flexible in designing complex models (see next lectures)
- Gibbs is a asymptotically converging to global optimum
- Usually need multiple random restarts, and test of convergence
- GMS: Lawrence et al. (1993), Science, 262:208-214

Discussion

- Model versus algorithm:
  - Model: e.g., oops, zoops, HMM,
  - Algorithms: EM, Gibbs, heuristics …
- Different algorithms can be used for solving the same model!
  - Need to be clear whether improvement/loss is due to model or algorithm (many paper/author got confused with these two aspects, and discuss results in convolved way!)
  - Fix one, and analyze the other, one at a time!!