

Computational Genomics

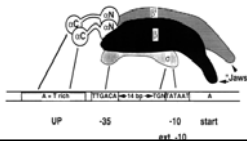
10-810/02-710, Spring 2009

Motif Detection

Eric Xing

Lecture 8, February 4-9, 2009

Reading: Durbin Chap 9,
class assignment



© Eric Xing @ CMU, 2005-2009

1

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, ...

© Eric Xing @ CMU, 2005-2009

2

-
- Figure 1 consists of three panels. Panel A is a line graph showing the relative levels of various transcription factors (Tor, Hb, Kni, Hb, Tl) across the embryo. The x-axis represents the position from anterior to posterior. Tor is high in the anterior, Hb is high in the posterior, Kni is high in the middle, and Tl is high in the posterior. Panel B is a schematic of the embryo showing the location of the *eve* st. 3 and *eve* st. 7 enhancers. Panel C is a schematic of the *Tor* gene structure, showing exons (K) and introns (H). The gene is organized into three main regions: a 5' region with exons K1, K2, K3, K4, K5, K6, K7, K8, K9, K10, K11, K12, K13, K14, K15, K16, K17, K18, K19, K20, K21, K22, K23, K24, K25, K26, K27, K28, K29, K30, K31, K32, K33, K34, K35, K36, K37, K38, K39, K40, K41, K42, K43, K44, K45, K46, K47, K48, K49, K50, K51, K52, K53, K54, K55, K56, K57, K58, K59, K60, K61, K62, K63, K64, K65, K66, K67, K68, K69, K70, K71, K72, K73, K74, K75, K76, K77, K78, K79, K80, K81, K82, K83, K84, K85, K86, K87, K88, K89, K90, K91, K92, K93, K94, K95, K96, K97, K98, K99, K100, K101, K102, K103, K104, K105, K106, K107, K108, K109, K110, K111, K112, K113, K114, K115, K116, K117, K118, K119, K120, K121, K122, K123, K124, K125, K126, K127, K128, K129, K130, K131, K132, K133, K134, K135, K136, K137, K138, K139, K140, K141, K142, K143, K144, K145, K146, K147, K148, K149, K150, K151, K152, K153, K154, K155, K156, K157, K158, K159, K160, K161, K162, K163, K164, K165, K166, K167, K168, K169, K170, K171, K172, K173, K174, K175, K176, K177, K178, K179, K180, K181, K182, K183, K184, K185, K186, K187, K188, K189, K190, K191, K192, K193, K194, K195, K196, K197, K198, K199, K200, K201, K202, K203, K204, K205, K206, K207, K208, K209, K210, K211, K212, K213, K214, K215, K216, K217, K218, K219, K220, K221, K222, K223, K224, K225, K226, K227, K228, K229, K230, K231, K232, K233, K234, K235, K236, K237, K238, K239, K240, K241, K242, K243, K244, K245, K246, K247, K248, K249, K250, K251, K252, K253, K254, K255, K256, K257, K258, K259, K260, K261, K262, K263, K264, K265, K266, K267, K268, K269, K270, K271, K272, K273, K274, K275, K276, K277, K278, K279, K280, K281, K282, K283, K284, K285, K286, K287, K288, K289, K290, K291, K292, K293, K294, K295, K296, K297, K298, K299, K300, K301, K302, K303, K304, K305, K306, K307, K308, K309, K310, K311, K312, K313, K314, K315, K316, K317, K318, K319, K320, K321, K322, K323, K324, K325, K326, K327, K328, K329, K330, K331, K332, K333, K334, K335, K336, K337, K338, K339, K340, K341, K342, K343, K344, K345, K346, K347, K348, K349, K350, K351, K352, K353, K354, K355, K356, K357, K358, K359, K360, K361, K362, K363, K364, K365, K366, K367, K368, K369, K370, K371, K372, K373, K374, K375, K376, K377, K378, K379, K380, K381, K382, K383, K384, K385, K386, K387, K388, K389, K390, K391, K392, K393, K394, K395, K396, K397, K398, K399, K400, K401, K402, K403, K404, K405, K406, K407, K408, K409, K410, K411, K412, K413, K414, K415, K416, K417, K418, K419, K420, K421, K422, K423, K424, K425, K426, K427, K428, K429, K430, K431, K432, K433, K434, K435, K436, K437, K438, K439, K440, K441, K442, K443, K444, K445, K446, K447, K448, K449, K450, K451, K452, K453, K454, K455, K456, K457, K458, K459, K460, K461, K462, K463, K464, K465, K466, K467, K468, K469, K470, K471, K472, K473, K474, K475, K476, K477, K478, K479, K480, K481, K482, K483, K484, K485, K486, K487, K488, K489, K490, K491, K492, K493, K494, K495, K496, K497, K498, K499, K500, K501, K502, K503, K504, K505, K506, K507, K508, K509, K510, K511, K512, K513, K514, K515, K516, K517, K518, K519, K520, K521, K522, K523, K524, K525, K526, K527, K528, K529, K530, K531, K532, K533, K534, K535, K536, K537, K538, K539, K540, K541, K542, K543, K544, K545, K546, K547, K548, K549, K550, K551, K552, K553, K554, K555, K556, K557, K558, K559, K560, K561, K562, K563, K564, K565, K566, K567, K568, K569, K570, K571, K572, K573, K574, K575, K576, K577, K578, K579, K580, K581, K582, K583, K584, K585, K586, K587, K588, K589, K590, K591, K592, K593, K594, K595, K596, K597, K598, K599, K600, K601, K602, K603, K604, K605, K606, K607, K608, K609, K610, K611, K612, K613, K614, K615, K616, K617, K618, K619, K620, K621, K622, K623, K624, K625, K626, K627, K628, K629, K630, K631, K632, K633, K634, K635, K636, K637, K638, K639, K640, K641, K642, K643, K644, K645, K646, K647, K648, K649, K650, K651, K652, K653, K654, K655, K656, K657, K658, K659, K660, K661, K662, K663, K664, K665, K666, K667, K668, K669, K670, K671, K672, K673, K674, K675, K676, K677, K678, K679, K680, K681, K682, K683, K684, K685, K686, K687, K688, K689, K690, K691, K692, K693, K694, K695, K696, K697, K698, K699, K700, K701, K702, K703, K704, K705, K706, K707, K708, K709, K710, K711, K712, K713, K714, K715, K716, K717, K718, K719, K720, K721, K722, K723, K724, K725, K726, K727, K728, K729, K730, K731, K732, K733, K734, K735, K736, K737, K738, K739, K740, K741, K742, K743, K744, K745, K746, K747, K748, K749, K750, K751, K752, K753, K754, K755, K756, K757, K758, K759, K760, K761, K762, K763, K764, K765, K766, K767, K768, K769, K770, K771, K772, K773, K774, K775, K776, K777, K778, K779, K780, K781, K782, K783, K784, K785, K786, K787, K788, K789

[illegible]

- long biosequence

3



R441

Y425

R441

042

1

Y4

Y42

25

5'

T

4

T

31

15

15

2

4

Example: Globin Motifs

```

XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX
HAHU V.LSPADKTN..VKAMGKRV..AHAGE.....YGAEL..ERMFLSF..PTTKTYFPH..FDLS..HGSA
HAOR M.LTDAEKKE..VTALMGKAA..GHGEE.....YGAEL..ERLFGAF..PTTKTYFPH..FDLS..HGSA
HADK V.LSAAKDN..VKGVFSKIG..GHAGE.....YGAETL..ERMFIAY..PQTQTYFPH..FDLS..STPD
HBHU VHLPTEKSA..VTALMGKVN..VDEVG.....G..EAL..GRLLVYV..PWTQRFPEA..FGDL..STPD
HBOR VHLSGGEKSA..VTNLMGKVN..INELG.....G..EAL..GRLLVYV..PWTQRFPEA..FGDL..SSAG
HBDK VMHTAEKQL..ITLTMGKVN..ADIPG.....G..EAL..ARLLVYV..PWTQRFPEA..FGDL..SSPT
MYHU G.LSDGEQL..VLKVMGKVE..ADIPG.....HQEVL..IRLFKTH..PETLEKFDK..FKGL..KSED
MYOR G.LSDGEQL..VLKVMGKVE..GDLPG.....HQEVL..IRLFKTH..PETLEKFDK..FKGL..KTED
IGLOB M.KFFAVLALCIVGAZASPLT..ADEASVqsswkvshNEVEILAAVFAAY..PDQNKFSQFagkDLASTKD
GPUGNI A.LTEKQEL..LKQSMELVK..QNIIPA.....HS..LRL..FALTIE..APESKYVFSF..LKDSNEIPE
GPYL GVLTDVQAL..VKSSFEEN..ANIPK.....N..THR..FTLVLEIAPGAKDLFSF..LKGSSEVPQ
GGZLB M.L.DQQTIN..IHKATVPVLENGVT.....ITTF..YKMLFAK..HPEVRPLFDW..GRQ..ESLE

```

```

XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX
HAHU QVKGH..GKKVADA..LTN.....AVA..HYDDMPNA..LSALS..D..LHAHKL.....RVDVPNF..KLLSHC..L
HAOR QVKAH..GKKVADA..L..S.....TAAGHFDMDGA..LSALS..D..LHAHKL.....RVDVPNF..KLLSHC..L
HADK QVKAH..GKKVAAA..LYE.....AVN..HYDDZAGA..LSKLS..D..LHAQKL.....RVDVPNF..KFLGHC..L
HBHU AVMGnpKVKAHGK..KVLGA..FGDALKNLDLKG...FATLS..E..LHCDKL.....HVDPENF..RL..LGNV..L
HBOR AVMGnpKVKAHGA..KVLTS..FGDALKNLDLKG...FAKLS..E..LHCDKL.....HVDPENF..RL..LGNV..L
HBDK AILGnpMVAHGK..KVLTS..FGDAVKNLNDIKNT...FAQLS..E..LHCDKL.....HVDPENF..RL..LGDIL..L
MYHU EMKASDILKKHGA..TVL.....TALGGLKKKH...EAEKPL..AQSHATK...HKIPVKYLEFTSEC..I
MYOR EMKASDILKKHGG..TVL.....TALGGLKKKH...EAEKPL..AQSHATK...HKISIKLEYISEA..I
IGLOB T..GA...FATHATRVISFSLseVIALSGNTSNAAAV...NSLSVKL..GDHKA...R..GVSA..QF..GEFR
GPUGNI NMPK...LKAHAFVFKTI...CESATELRQKHGAVdNNTLKRLL..GSIHLK...N..KITDP..HF..EVMKG
GPYL NMPD...LQAHAH..KVFKL...TYEAAIQLEVNGAVAS..DATLKRLL..GSHVRS...K..GVDA..HF..PVVKE
GGZLB Q.....PKALAM..TVL.....AAQNIENLPAIL...PAVKKIAVKKHQAGvaaaH..YPIVQGE...LGAIK

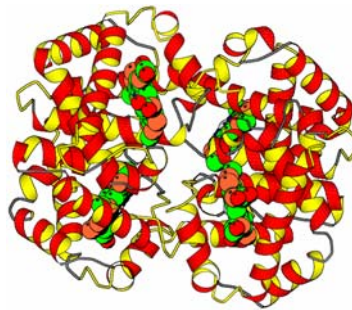
```

```

XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX,XXXXXXXXXX
HAHU VT..AA..H..LPAEFTPA..VHASIDKFLASV..STVLTS..KY..R
HAOR VV..AR..H..CPGEFTPS..AHASIDKFLSKV..ATVLTS..KY..R
HADK VV..AI..H..HFAALTP..VHASIDKFLCAV..GAVLTA..KY..R
HBHU VCV..AH..H..FSKEFTPP..VQAVQKVAVGV..AHALAR..KY..H
HBOR IIV..AR..H..FSKDFSP..VQAVQKLVGV..AHALGR..KY..H
HBDK IIV..AA..H..FTKDFTP..CQAVQKLVVRV..AHALAR..KY..H
MYHU QV..LQSKHPDGFADAQA..WKALELFRDM..ASNYKELGFQ..G
MYOR QV..LQSKHSDGFADAQA..WKALELFRDM..AAKYNEFGFQ..G
IGLOB TA..LV..Y..QANVSGDVAH..HNKA..LDN..TFATVY..PR..L
GPUGNI ALLGTIKEA..I..KENMSDE..MGDANTEAYNQVATIKKE..MK..E
GPYL ATLGTIKEV..V..KDNMSDE..LNTANTAYDELATIKKE..MKdaa
GGZLB EVLGDAAT..DDI..AMGK..AYGVIAVFIQVADLYAQ..AV..E

```

Hemoglobin alpha subunit

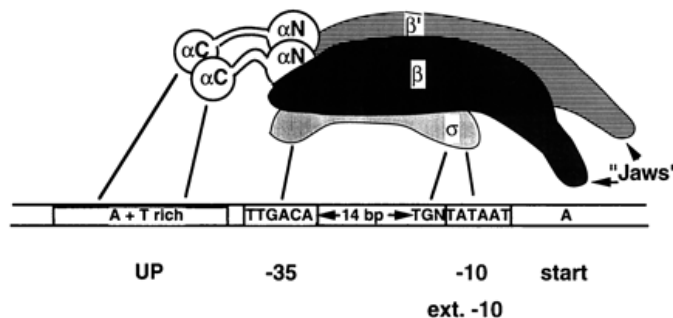


© Eric Xing @ CMU, 2005-2009

5

DNA Motif

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



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

© Eric Xing @ CMU, 2005-2009

6

Example: Gcn4



Regulatory Signals

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAAGACATCGAAACATACAT ...HIS7
 5' - ATGGCAGAATCACTTTAAACGTGGCCCCACCCGCTGCACCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...AR04
 5' - CACATCCAACGAATCACCTCACCGTTATCGTGAATCATTCTTTTCGCATCGCCGAAGTGCCATAAAAAATATTTTT ...ILV6
 5' - TGCGAACAAAAGAGTCAATTACAACGAGGAAATAGAAGAAAAATTTTCGACAAAATGTATAGTCATTCTATC ...THR4
 5' - ACAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATGTCATGCATATGACTCATCCGGAACATGAAA ...AR01
 5' - ATTGATTGACTCATTTCCTCTGACTACTACCAAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAATAAATAA ...HOM2
 5' - GGCGCCACAGTCCGCGTTTGGTTATCCGGCTGACTCCTTTTGGAAAGTGTCATGTGCTTCACACA ...PRO3

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

© Eric Xing @ CMU, 2005-2009

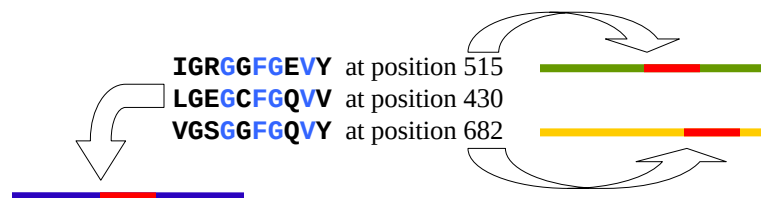
7

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 ...

seq. 1
 seq. 2
 seq. 3



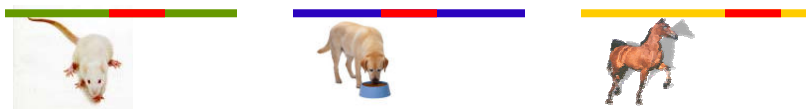
© Eric Xing @ CMU, 2005-2009

8

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



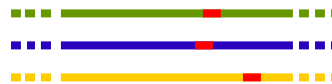
© Eric Xing @ CMU, 2005-2009

9

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.



© Eric Xing @ CMU, 2005-2009

10

Characteristics of Regulatory Motifs



```

ATATAAA TT T
CTGATA A CAG
GTGA TCA CA
AGGGGG AG CG
AA AA AA AA
TTTAA T AA AA
GAAACG TTGCG
AA TTA A T A
TTT A T A T AA
CGGACGAG
AAAAATTT
A GA A AAAA AA
T ATGAA TT
AA AA AAAAA
TTT A AA
T T T T A AAA
ATATAT AT A
ATTA AAAA TT
    
```

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

© Eric Xing @ CMU, 2005-2009

11

Motif Representation



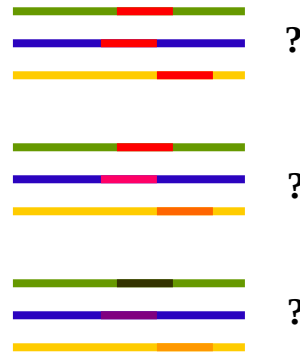
© Eric Xing @ CMU, 2005-2009

12

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.



© Eric Xing @ CMU, 2005-2009

13

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.

© Eric Xing @ CMU, 2005-2009

14

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

© Eric Xing @ CMU, 2005-2009

15

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**.

© Eric Xing @ CMU, 2005-2009

16

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	.8

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

© Eric Xing @ CMU, 2005-2009

17

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

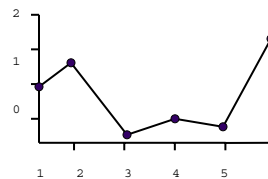
Counts from 242 known σ^{70} sites

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_{bi}

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

$10 \log_2 \theta_{li} / \theta_{oi}$



Informativeness: $2 - \sum_i \theta_{li} \log_2 \theta_{li} / \theta_{oi}$

© Eric Xing @ CMU, 2005-2009

18

Relative entropy

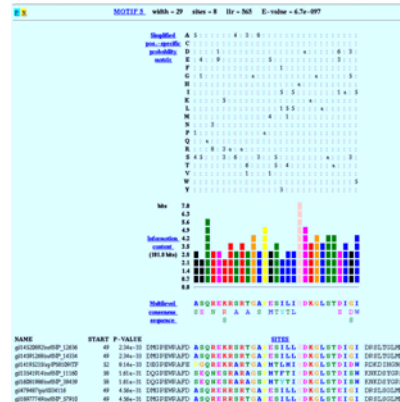
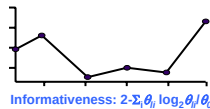
- A motif is *interesting* if it is very different from the background distribution
- Use relative entropy:

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

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

- Relative entropy is sometimes called information content

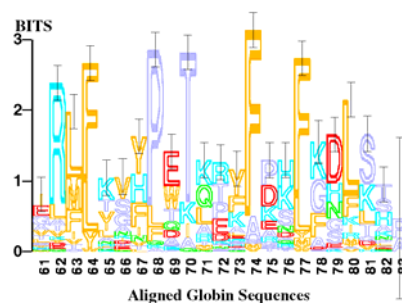
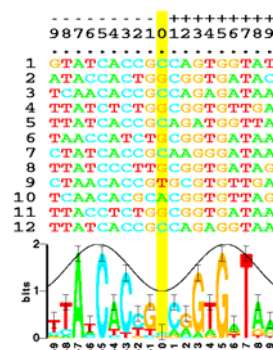
A	.04	.88	.26	.59	.49	.03
C	.09	.03	.11	.13	.21	.05
G	.07	.01	.12	.16	.12	.02
T	.80	.08	.51	.13	.18	.89



© Eric Xing @ CMU, 2005-2009

19

Sequence Logo



- Information at pos'n l , $H(l) = -\sum_{\{\text{letter } i\}} \theta_{li} \log_2 \theta_{li}$
- Height of x at pos'n l , $L(l, x) = \theta_{li} (2 - H(l))$
 - Examples:
 - $\theta_{lA} = 1$; $H(l) = 0$; $L(l, A) = 2$
 - $A: \frac{1}{2}$; $C: \frac{1}{4}$; $G: \frac{1}{4}$; $H(l) = 1.5$; $L(l, A) = \frac{1}{4}$; $L(l, \text{not T}) = \frac{1}{4}$

© Eric Xing @ CMU, 2005-2009

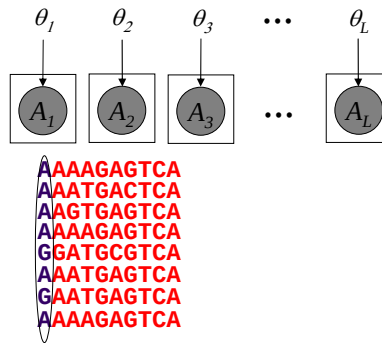
20

The Product Multinomial (PM) Model

[Lawrence *et al.* Science 1993]



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



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

© Eric Xing @ CMU, 2005-2009

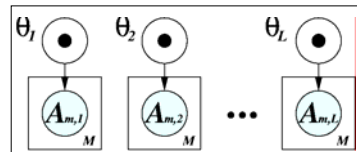
21

More on PM Model



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

	1	2	3	4	5	6	7	8	9	10
A	.750	.875	.875	.275	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	.625	0	0	0	.1	0	0



The nucleotide distributions at different sites are independent !

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

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

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

© Eric Xing @ CMU, 2005-2009

22

Computational problems for *in silico* motif detection



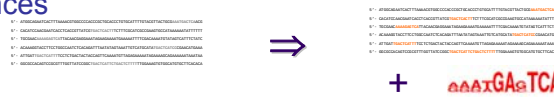
- Extract a motif model based on (experimentally) identified motifs



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



- Uncover novel motifs computationally from genomic sequences



de novo motif detection

23

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

© Eric Xing @ CMU, 2005-2009

24

Problem definition



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

Combinatorial

Motif M: substring $m_1 \dots m_W$

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

$j = A, C, G, T$

$\theta_{ij} = \text{Prob}[\text{letter } j, \text{ pos } i]$

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

© Eric Xing @ CMU, 2005-2009

25

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.

	C	T	A	T	A	A	T	C	
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			-93

	C	T	A	T	A	A	T	C	
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			+85

	C	T	A	T	A	A	T	C	
A	-38	19	1	2	10	48			
C	-15	-38	-8	-10	-3	-32			
G	-13	-48	-6	-7	-10	-40			
T	17	-32	8	-9	-6	19			-95

© Eric Xing @ CMU, 2005-2009

26

Supervised motif search

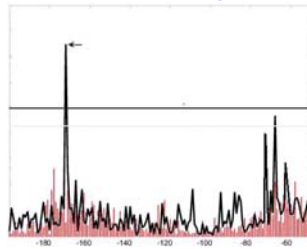
- Supervised learning

- Given biologically identified aligned motifs **A**, maximal likelihood estimation:

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

- Application:

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



- Need more more sophisticated search model: HMM?

© Eric Xing @ CMU, 2005-2009

27

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.

```

5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...HIS7
5' - ATGGCAGAATCACTTTAAACGTGGCCCC? TGTACGTTACTGCGAAATGACTCAACG ...AR04
5' - CACATCCAACGAATCACCTCACCGTTATCC AAAATG?TCA GCCGAAGTGCCATAAAAAATATTTTTT ...ILV6
5' - TCGGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTCTATC ...THR4
5' - ACAAGGTACCTTCCTGGCCAATCTCACAGATTTAATATAGTAAATTGTCATGCATATGACTCATCCGAACATGAAA ...AR01
5' - ATTGATTGACTCATTTTCTCTGACTACTACCAAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAATAAATAA ...HOM2
5' - GGCGCCACAGTCCGCGTTGGTTATCCGGCTGACTCATTCTGACTCTTTTTTGAAAAGTGTGGCATGTGCTTCACACA ...PRO3
    
```

- Learning algorithms:

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

© Eric Xing @ CMU, 2005-2009

28

de novo motif detection



- **Problem setting:**

- Given UTR sequences: $y = \{y_1, \dots, y_N\}$
- Goal: the background model: $\theta_0 = \{\theta_{0,A}, \theta_{0,T}, \theta_{0,G}, \theta_{0,C}\}^t$
and K motif models $\theta^1, \dots, \theta^K$ from y ,
where $\theta^k = \{\theta_{i,j}^k : i = 1, \dots, L_k, j \in \{A, C, G, T\}\}$

- **A missing value problem:**

- The locations of instances of motifs are unknown, thus the aligned motif sequences $A_1, \dots, A_K$ and the background sequence are not available.

© Eric Xing @ CMU, 2005-2009

29

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
```

© Eric Xing @ CMU, 2005-2009

30

Sample DNA sequences



```
>ce1cg
TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATA
GCGCGTGGTGTGAAAGACTGTTTTTTGATCGTTTTCAC
AAAAATGGAAGTCCACAGTCTTGACAG

>ara
GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAG
AAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTG
CTATGCCATAGCATTTTTATCCATAAG

>bglr1
ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATA
TAACTTTATAAATTCCTAAAATTACACAAAGTTAATAAC
TGTGAGCATGGTCATTTTTATCAAT

>crp
CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTAC
AGTAATACATTGATGTACTGCATGTATGCAAAGGACGTC
ACATTACCGTGCAGTACAGTTGATAGC
```

© Eric Xing @ CMU, 2005-2009

31

Motif occurrences



```
>ce1cg
taatgtttgtgctggTTTTTgtggcatcgggcgagaata
gcgcgtggtgtgaaagactgTTTTTTTGATCGTTTTCAC
aaaaatggaagtccacagtcttgacag

>ara
gacaaaaacgcgtaacaaaagtgtctataatcacggcag
aaaagtccacattgattaTTTGACGGCGTCACactttg
ctatgccatagcatTTTTatccataag

>bglr1
acaaatccaataacttaattattgggatttgttatata
taactttataaattcctaaaattacacaaagttaataac
TGTGAGCATGGTCATatTTTTatcaat

>crp
cacaaagcgaaagctatgctaaaacagtcaggatgctac
agtaatacattgatgtactgcataTGCAAAGGACGTC
ACattaccgtgcagtacagttgatagc
```

One-occurrence per sequence (oops) model

© Eric Xing @ CMU, 2005-2009

32

Starting point



...gactgtttt**TTTGATCGTTTTCAC**aaaaatgg...

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 is a special initialization scheme, many other schemes, including random starts, are also valid

© Eric Xing @ CMU, 2005-2009

33

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 + ...**

© Eric Xing @ CMU, 2005-2009

34

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	...

© Eric Xing @ CMU, 2005-2009

35

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
TTTGATCGTTTTTAC	→	A 000132011000040
TTTGACGCGTCAC		C 001010300200403
TGTGAGCATGGTCAT		G 020301131130000
TGCAAAGGACGTCAC		T 423001002114001

© Eric Xing @ CMU, 2005-2009

36

Adding pseudocounts



Counts		Counts + Pseudocounts
A 000132011000040	→	A 111243122111151
C 001010300200403		C 112121411311514
G 020301131130000		G 131412242241111
T 423001002114001		T 534112113225112

© Eric Xing @ CMU, 2005-2009

37

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						

© Eric Xing @ CMU, 2005-2009

38

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.

© Eric Xing @ CMU, 2005-2009

39

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

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

© Eric Xing @ CMU, 2005-2009

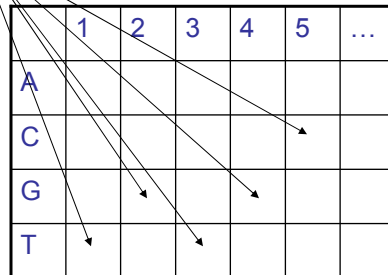
40

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 novo motif finding



5' - TCTCTCTCCACGGCTAATTAGGTGATCATGAAAAATGAAAAATTCATGAGAAAAGAGTCAGACATCGAAACATACAT ...*HIS7*
 5' - ATGGCAGAAATCACTTTAAACGTGGCCCAACCGCTGCACCTGTGCATTTTGTACGTTACTGCGAAATGACTCAACG ...*AR04*
 5' - CACATCCAACGAATCACCTCACCGTTATCGTGACTCACTTTCTTCGCATCGCCGAAGTGCCATAAAAAATATTTTTT ...*ILV6*
 5' - TGCGAACAAAAGAGTCATTACAACGAGGAAATAGAAGAAAATGAAAAATTTTCGACAAAATGTATAGTCATTTCTATC ...*THR4*
 5' - ACAAAGGTACCTTCCTGGCCAATCTCACAGATTAAATATAGTAAATTGTCATGCATATGACTCATCCGAACATGAAA ...*AR01*
 5' - ATTGATTGACTCATTTTCTCTGACTACTACGAGTTCAAAATGTTAGAGAAAAATAGAAAAGCAGAAAAATAATAA ...*HOM2*
 5' - GCGGCCACAGTCCGCGTTTGTTATCCGGCTGACTCATTCTGACTCTTTTTTGAAAAGTGCGCATGTGCTTCACACA ...*PRO3*

1: AAAAGAGTCA
 2: AAATGAGTCA
 . AAGTGAATCA
 . AAAGAGTCA
 . GGATGAGTCA
 . AAATGAGTCA
 . GAATGAGTCA
 N: AAAAGAGTCA



AAATGAGTCA

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.

© Eric Xing @ CMU, 2005-2009

43

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.

© Eric Xing @ CMU, 2005-2009

44

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?

© Eric Xing @ CMU, 2005-2009

45

The Gibbs Motif Sampler



- Initialized by choosing random starting positions

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

- 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

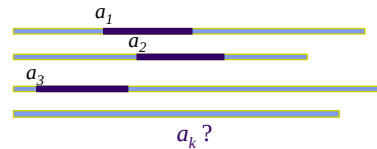
(no need to sample $\theta^{(t)}$ at each t, we can compute it in close-form, see next lecture)

- Notations:

$$A_{[-k]} \equiv \{a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_K\} \Rightarrow c_{i,j,[-k]}, \quad i=1 \dots W, j=A,T,G,C$$

$$a_k \Rightarrow c_{i,j}(a_k)$$

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



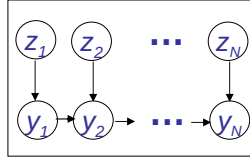
© Eric Xing @ CMU, 2005-2009

46

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:

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

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

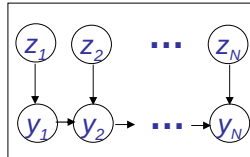
$$p(Y_{n \dots n+L-1} | z_n = 0) = \theta_{0,y_n} \theta_{0,y_{n+1}} \dots \theta_{0,y_{n+L-1}} = \prod_{l=0}^{L-1} \theta_{0,y_{n+l}} = \prod_{l=1}^L \prod_{j=1}^4 \theta_{0,j}^{\delta(y_{n+l-1}, j)} \quad (\text{background})$$

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

© Eric Xing @ CMU, 2005-2009

47

A binary indicator model



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

- Complete log-likelihood :
 - suppose all words are concatenated into one big sequence of $y = y_1 y_2 \dots y_N$, with appropriate constraints preventing overlapping and boundaries limits

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

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

© Eric Xing @ CMU, 2005-2009

48

The Maximal likelihood approach



- Maximize expected likelihood, in iteration of two steps:

Expectation:

Find expected value of complete log likelihood:

$$E[\log \mathcal{P}(\mathcal{Y}_1 \dots \mathcal{Y}_n, Z \mid \theta, \theta_0, \varepsilon)]$$

Maximization:

Maximize the expected complete likelihood over $\theta, \theta_0, \varepsilon$

Expectation Maximization: E-step



- Expectation:

Find expected value of log likelihood:

$$\begin{aligned} \langle l_c(\Theta) \rangle = & \sum_{n=1}^N \langle z_n \rangle \left(\sum_{l=1}^L \sum_{j=1}^4 \delta(\mathcal{Y}_{n+l-1}, j) \log \theta_{l,j} \right) + \sum_{n=1}^N (1 - \langle z_n \rangle) \left(\sum_{j=1}^4 \delta(\mathcal{Y}_{n+l-1}, j) \log \theta_{0,j} \right) \\ & + \sum_{n=1}^N \langle z_n \rangle \log \varepsilon + \left(N - \sum_{n=1}^N \langle z_n \rangle \right) \log(1 - \varepsilon) \end{aligned}$$

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

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

- 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:

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

Expectation Maximization: M-step



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

$c_{i,j} = E[\text{\# times letter } j \text{ appears in motif position } i]$

$c_{0,j} = E[\text{\# times letter } j \text{ appears in background}]$

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

It easily follows:

$$\theta_{i,j}^{NEW} = \frac{c_{i,j}}{\sum_{j=1}^A c_{i,j}} \quad \theta_{0,j}^{NEW} = \frac{c_{0,j}}{\sum_{j=1}^A c_{0,j}}$$

to not allow any 0's, add pseudocounts

Initial Parameters Matter!



- Consider the following “artificial” example:

$x^1, \dots, x^N$ 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 $\ll 2^{12}$ occurrences of 12-mer ACTGACTGACTG

- Some local maxima:

$$\varepsilon \approx 1/2; \quad B = 1/2C, 1/2G; \quad M_i = 1/2A, 1/2T, i = 1, \dots, 12$$

- very bad !

$$\varepsilon \approx D/2^{L+1}; \quad B = 1/4A, 1/4C, 1/4G, 1/4T; \\ M_1 = 100\% A, M_2 = 100\% C, M_3 = 100\% T, \text{ etc.}$$

- the correct solution !

© Eric Xing @ CMU, 2005-2009

53

Overview of EM Algorithm



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

© Eric Xing @ CMU, 2005-2009

54

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 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!!