

More Motifs

WMM, log odds scores, Neyman-Pearson, background;
Greedy & EM for motif discovery

Talks this week

- Tue, 3:30 EE-105, Me
“*The Search for Non-Coding RNA*”
- Wed, 1:30 K-069, Zasha Weinberg
“something similar...”

Neyman-Pearson

- Given a sample x_1, x_2, \dots, x_n , from a distribution $f(\Theta; \dots)$ with parameter Θ , want to test hypothesis $\Theta = \theta_1$ vs $\Theta = \theta_2$.
- Might as well look at *likelihood ratio*
 $f(\theta_1; x_1, x_2, \dots, x_n) / f(\theta_2; x_1, x_2, \dots, x_n) > \tau$

What's best WMM?

- Given 20 sequences s_1, s_2, \dots of length 8, assumed to be generated at random according to a WMM defined by $8 \times (4-1)$ parameters θ , what's the best θ ?
- E.g., what MLE for θ given data s_1, s_2, \dots ?
- Answer: count frequencies per position.

Weight Matrix Example

8 Sequences

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

Profile	1	2	3
A	.625	0	0
C	0	0	0
G	.250	0	1
T	.125	1	0

Log Likelihood Ratio	1	2	3
A	1.32	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2
T	-1	2	-∞

$$\log_2 \frac{f_{xi} \cdot i}{f_{xi}} \quad , f_{xi} = \frac{1}{4}$$

Non Uniform Background

E. coli - DNA approximately $\frac{1}{4}$ A, C, G, T

M. jannaschi - 68% A-T, 32% GC

LLR from Previous Example w $f_A = f_T = \frac{3}{8}$

$$f_C = f_G = \frac{1}{8}$$

	1	2	3
A	.937	-∞	-∞
C	-∞	-∞	-∞
G	1.00	-∞	3
T	-1.58	1.42	-∞

E.g. "G" in
position 3
is $2^3 = 8 \times$
more likely
than background

How "Informative" is a WMM?

Recall Relative Entropy

$$H(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

If x are sequences (fixed length)

$P(x)$ = prob. of x according to WMM (or other model)

$Q(x)$ = Background model

$H(P||Q)$ is expected log likelihood score

of a randomly chosen site (random according to site model)

For WMM, can show

$$H(P||Q) = \sum_{i=1}^n H(P_i||Q_i)$$

where P_i, Q_i are distributions of i^{th} position

[Follows from assumption of independence]

Example (cont.)

	1	2	3
A	.625	0	0
C	0	0	0
G	.250	0	1
T	.125	1	0

	1	2	3		1	2	3
D	1.32	-∞	-∞	A	.737	-∞	-∞
C	-∞	-∞	-∞	C	-∞	-∞	-∞
G	0	-∞	2	G	1.00	-∞	3.00
T	-1	2	-∞	T	-1.58	1.42	-∞
Rel. Ext.	.701	2	2	rel. Ext.	.512	1.42	3.0

uniform

non-uniform

Pseudo counts

Are the $- \infty$'s a problem?

- if you are certain a given residue never occurs in a given position, then $- \infty$ is just right
- if not, then it's probably an artifact of small sample

Typical fix:

add a small constant (eg .5, 1, 2)
to all observed counts - a pseudocount

Questions

- Given aligned instances of motifs,
How do you build model?

Frequency counts, as above

- Given model, how do you find
(probable) instances?

Scanning

- Given unaligned strings thought
to contain a motif, how do you find it?
Eg. upstream regions from μ array cluster

Motif Discovery Three Approaches

- ① Greedy Search
- ② Expectation Maximization
- ③ Gibbs Sampler

P.S. Finding a site of max relative entropy in a set of unaligned sequences is NP-hard (Akutsu)

GREEDY Algorithm [Hertz & Stormo]

Inputs:

Sequences $S_1 \dots S_K$, motif length l , "breadth" d ,
& Background

Algorithm

1. Create a singleton set with each length l subsequence of each of $S_1 \dots S_K$
2. For each set returned add each possible length l subseq not already present
3. Compute relative entropy of each
Return d best.
4. Repeat until each set has K strings.

NB: usual Greedy problems

Expectation Maximization

MEME [Bailey & Elkan]

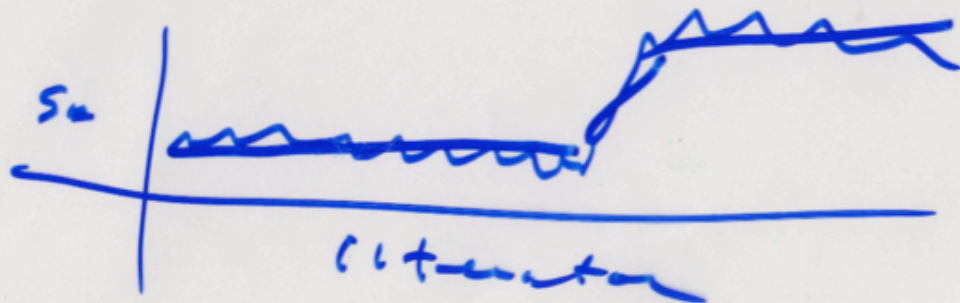
Input: again sequences $s_1 \dots s_k$ & motif length l
& background model

Again assume 1 instance per sequence
(variants possible)

Observed data: the sequences

Parameters Θ : the WMM

Hidden Data: where's motif? $Y_{ij} = \begin{cases} 1 & \text{if start} \\ & \text{at pos } j \\ & \text{of seq } i \\ 0 & \text{otherwise} \end{cases}$



Expectation Step

$$\hat{Y}_{ik} = E(Y_{ik} | S_i, \Theta)$$
$$= P(Y_{ik}=1 | S_i, \Theta)$$

$$\} E = 1 \cdot P(1) + 0 \cdot P(0)$$

$$= P(S_i | Y_{ik}=1, \Theta) \frac{P(Y_{ik}=1 | \Theta)}{P(S_i | \Theta)}$$

Bayes
again

$$= c \cdot P(S_i | Y_{ik}=1, \Theta)$$

$$= c' \prod_{j=1}^k P(S_{i,j+k-j-1} | \Theta)$$

$$\text{Fix } c' \text{ so } \sum_k \hat{Y}_{ik} = 1$$

Maximization Step

given parameter Θ^t @ t^{th} iteration

Find Θ maximizing Expected value

$$Q(\Theta | \Theta^t) = E_{Y \sim \Theta^t} [\log P(S, Y | \Theta)]$$

$$= E \left[\log \prod_{i=1}^K P(s_i, Y_i | \Theta) \right]$$

\vdots

$$= \sum_i \sum_j E(Y_{ij}) \log P(s_i | \Theta, Y_{ij}=1)$$

Θ maximized by "counting" frequencies in alignment, where counts are \hat{Y}_{ij} .

Initialization

1. Buy a super computer; call it SDSC
2. Try every motif-length substring
as initial ~~Q~~ & use WMM with, say, 80% of mass
on that sequence, rest uniform
3. Run a couple of iterations of each;
4. Run best few to convergence