Genome 559

Lecture 12a, 2/11/10
Larry Ruzzo

A little more about motif models

Your Feedback

- Most seemed happy
- Plurality think pace is about right (but significant spread of opinions)
- More and more complex examples?
- Memory efficiency? General strategies?

Motifs II – Outline

Quick review of motifs and WMM/PSSM

Statistical justification for log ratios

Statistical justification for frequency counts

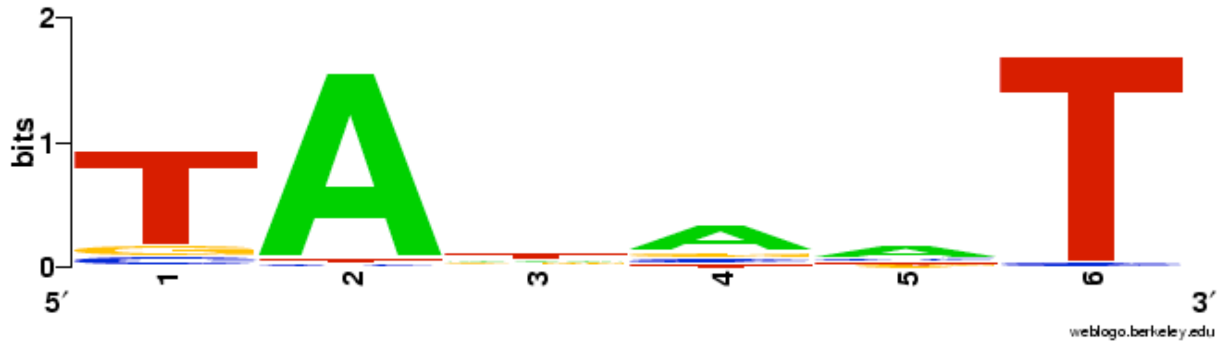
Another example

TATA Box Frequencies

pos / base	1	2	3	4	5	6
A	2	94	26	59	50	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

Sequence Logo

http://weblogo.berkeley.edu



Frequencies

base \ pos	1	2	3	4	5	6
A	2	94	26	59	50	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

Frequency ⇒ Scores:

$\log_2$ (freq/background)

Scores

base \ pos	1	2	3	4	5	6
A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

(For convenience, scores multiplied by 10, then rounded)

Scanning for TATA

A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

A C T A T A A T C G

= -90

A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

A C T A T A A T C G

= 85

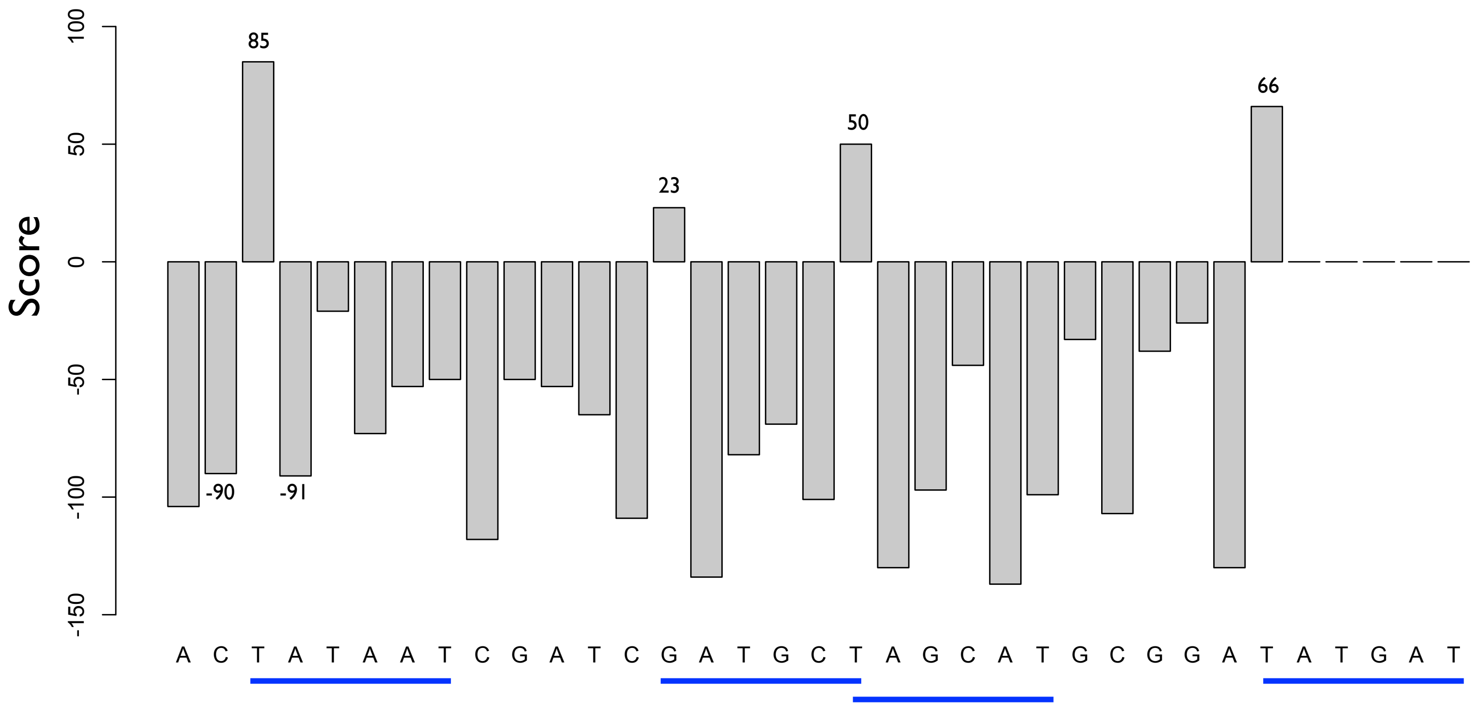
A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

A C T A T A A T C G

= -91

Stormo, Ann. Rev. Biophys. Biophys Chem, 17, 1988, 241-263

Scanning for TATA



Weight Matrices: Thermodynamics

Experiments show ~80% correlation of (log likelihood) weight matrix scores to measured binding energy of RNA polymerase to variations on TATAAT consensus [Stormo & Fields]

Justification?

Kinda sensible, kinda works

Is there a less *ad hoc* view?

One such framework:

Statistical Hypothesis Testing:

Is this sequence more like my "TATA" model
or more like my "everything else" model

Hypothesis Testing: A Very Simple Example

Given: A coin, either fair (p(H)=1/2) or biased (p(H)=2/3)

Decide: which

How? Flip it 5 times. Suppose outcome D = HHHTH

Null Model/Null Hypothesis M_0: p(H)=1/2

Alternative Model/Alt Hypothesis M_1: p(H)=2/3

Likelihoods:

$P(D \mid M_0) = (1/2)\ (1/2)\ (1/2)\ (1/2)\ (1/2) = 1/32$

$P(D \mid M_1) = (2/3)\ (2/3)\ (2/3)\ (1/3)\ (2/3) = 16/243$

Likelihood Ratio: $$\frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{16/243}{1/32} = \frac{512}{243} \approx 2.1$$

I.e., alt model is ≈ 2.1x more likely than null model, given data

Hypothesis Testing, II

Log of likelihood ratio is equivalent, often more convenient

add logs instead of multiplying…

"Likelihood Ratio Tests": reject null if LLR > threshold

LLR > 0 disfavors null, but higher threshold gives stronger evidence against, i.e., shifts false positive/false negative rates

Neyman-Pearson Theorem: For a given error rate, LRT is as good a test as any (subject to some fine print).

Weight Matrices: Statistics

Assume:

$f_{b,i}$ = frequency of base *b* in position *i in TATA*

f_b = frequency of base *b* in all sequences

Log likelihood ratio, given $S = B_1 B_2 \ldots B_6$:

$$\log\left(\frac{P(S|\text{"tata"})}{P(S|\text{"non-tata"})}\right) = \log \frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}} = \sum_{i=1}^{6} \log \frac{f_{B_i,i}}{f_{B_i}}$$

freq $\longrightarrow$ score

Assumes *independence*

Interpretation of Scores

A probabilistic interpretation of WMM scores: if

$$score = 10 \log_2 (ratio)$$

then

$$ratio = 2^{score/10}$$

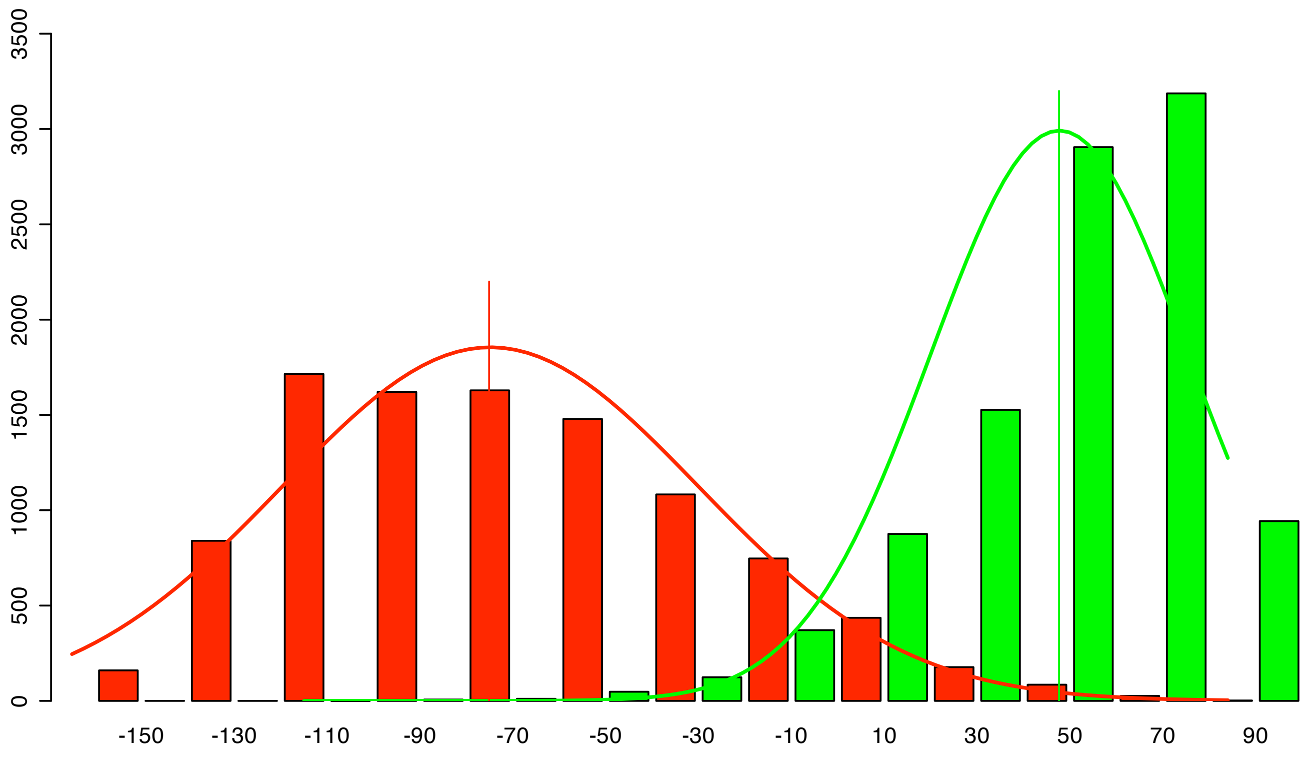
E.g., score +30 $\Rightarrow 2^{30/10} = 2^3 = 8$ times more likely under the WMM model than under the null model.

E.g., -40 $\Rightarrow 2^{-4} = 16$x more likely under the null.

But treat this cautiously; model is approximate

Score Distribution

(Simulated)



What's best WMM?

Given, say, 168 sequences $s_1, s_2, ..., s_k$ of length 6, assumed to be generated at random according to a WMM defined by 6 x (4-1) parameters θ, what's the best θ?

Answer: count frequencies per position.

> Analogously, if you saw 900 Heads in1000 coin flips, you'd perhaps estimate P(Heads) = 900/1000

Why is this sensible?

Parameter Estimation

Assuming sample $x_1, x_2, ..., x_n$ is from a parametric distribution $f(x|\theta)$, estimate θ.

E.g.:

$x_1, x_2, ..., x_5$ is HHHTH, estimate θ = prob(H)

Likelihood

$P(x \mid \theta)$: Probability of event x given model θ

Viewed as a function of x (fixed θ), it's a *probability*

E.g., $\Sigma_x P(x \mid \theta) = 1$

Viewed as a function of θ (fixed x), it's a *likelihood*

E.g., $\Sigma_\theta P(x \mid \theta)$ can be anything; *relative* values of interest.

E.g., if θ = prob of heads in a sequence of coin flips then

$P(HHHTH \mid .6) > P(HHHTH \mid .5)$,

I.e., event HHHTH is *more likely* when $\theta = .6$ than $\theta = .5$

And what θ make HHHTH *most* likely?

Maximum Likelihood Parameter Estimation

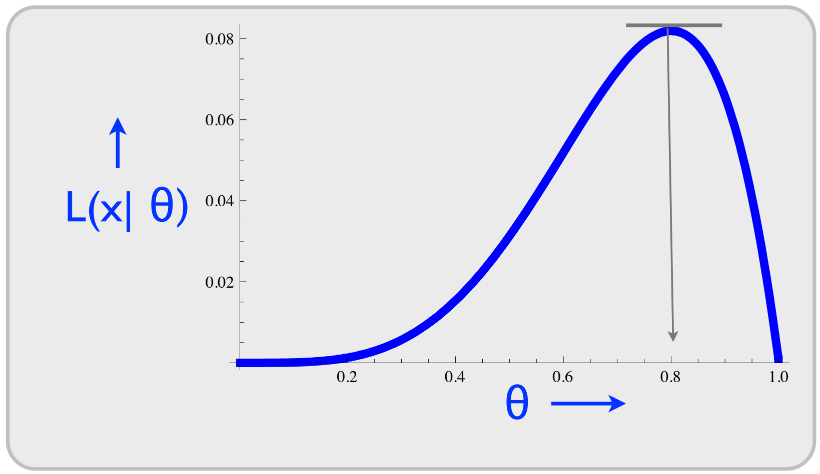
One (of many) approaches to param. est.

Likelihood of (indp) observations $x_1, x_2, \ldots, x_n$

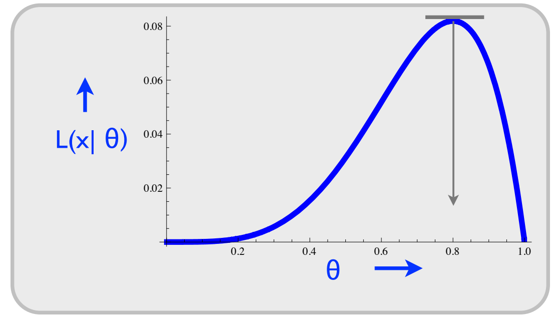
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

As a function of θ, what θ maximizes the likelihood of the data actually observed. Typical approaches:

- Numerical
- MCMC
- Analytical – $\frac{\partial}{\partial \theta} L(\vec{x} \mid \theta) = 0$
- etc.



Example I



n coin flips, $x_1, x_2, ..., x_n$; n_0 tails, n_1 heads, $n_0 + n_1 = n$;

θ = probability of heads

$$
\begin{aligned}
L(x_1, x_2, \ldots, x_n \mid \theta) &= (1-\theta)^{n_0}\theta^{n_1} \\
\log L(x_1, x_2, \ldots, x_n \mid \theta) &= n_0 \log(1-\theta) + n_1 \log\theta \\
\frac{\partial}{\partial\theta}\log L(x_1, x_2, \ldots, x_n \mid \theta) &= \frac{-n_0}{1-\theta} + \frac{n_1}{\theta}
\end{aligned}
$$

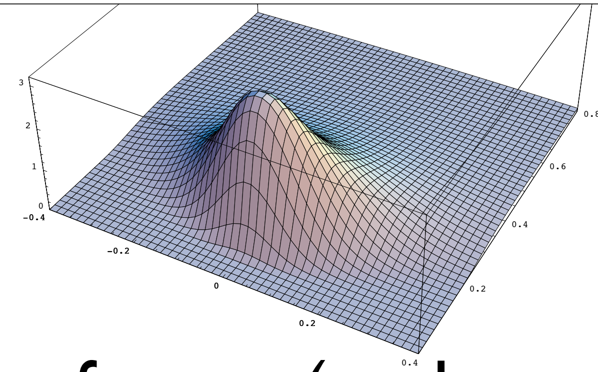
Setting to zero and solving:

$$\boxed{\hat{\theta} = \frac{n_1}{n}}$$

Observed fraction of successes in sample is MLE of success probability in population

(Also verify it's max, not min, & not better on boundary)

Example 11



n letters, $x_1, x_2, \ldots, x_n$ drawn at random from a (perhaps biased) pool of A, C, G, T, $\quad n_A + n_C + n_G + n_T = n$;

$\theta = (\theta_A, \theta_C, \theta_G, \theta_T)$ proportion of each nucleotide.

Math is a bit messier, but result is similar to coins

$$\boxed{\hat{\theta} = (n_A/n, n_C/n, n_G/n, n_T/n)}$$

Observed fraction of nucleotides in sample is MLE of nucleotide probabilities in population

Pseudocounts

Freq/count of 0 $\Rightarrow -\infty$ score; a problem?

- Certain that a given residue *never* occurs in a given position? Then $-\infty$ just right.
- Else, it may be a small-sample artifact

Typical fix: add a *pseudocount* to each observed count—small constant (e.g., .5, 1)

Sounds *ad hoc*; there is a Bayesian justification

Influence fades with more data

What's best WMM?

Given, say, 168 sequences $s_1, s_2, ..., s_k$ of length 6, assumed to be generated at random according to a WMM defined by 6 x (4-1) parameters θ, what's the best θ?

E.g., what's MLE for θ given data $s_1, s_2, ..., s_k$?

Answer: count frequencies per position.

Another WMM example

8 Sequences:

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

Log-Likelihood Ratio:

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

Freq.	Col 1	Col 2	Col 3
A	0.625	0	0
C	0	0	0
G	0.250	0	1
T	0.125	1	0

LLR	Col 1	Col 2	Col 3
A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2.00
T	-1.00	2.00	$-\infty$

Non-uniform Background

- *E. coli* - DNA approximately 25% A, C, G, T
- *M. jannaschi* - 68% A-T, 32% G-C

LLR from previous example, assuming

$$f_A = f_T = 3/8$$
$$f_C = f_G = 1/8$$

LLR	Col 1	Col 2	Col 3
A	0.74	-∞	-∞
C	-∞	-∞	-∞
G	1.00	-∞	3.00
T	-1.58	1.42	-∞

e.g., G in col 3 is 8 x more likely via WMM than background, so ($\log_2$) score = 3 (bits).

WMM Example, cont.

Freq.	Col 1	Col 2	Col 3
A	0.625	0	0
C	0	0	0
G	0.250	0	1
T	0.125	1	0

Uniform

LLR	Col 1	Col 2	Col 3
A	1.32	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2.00
T	-1.00	2.00	$-\infty$

Non-uniform

LLR	Col 1	Col 2	Col 3
A	0.74	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	1.00	$-\infty$	3.00
T	-1.58	1.42	$-\infty$

Summary

Motif description/recognition fits a simple statistical framework

- Frequency counts give MLE parameters
- Scoring is log likelihood ratio hypothesis testing
- Scores are interpretable

Log likelihood scoring naturally accounts for background (which is important):

- log(foreground freq/background freq)

These approaches broadly useful