Genome 559

Lecture 13a, 2/16/10
Larry Ruzzo

A little more about motif models

Motifs III – Outline

Statistical justification for frequency counts

Relative Entropy

Another example

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

Frequency ⇒ Scores:
$\log_2$ (freq/background)

Scores

pos base	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)

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 $x_1, x_2, ..., x_n$

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

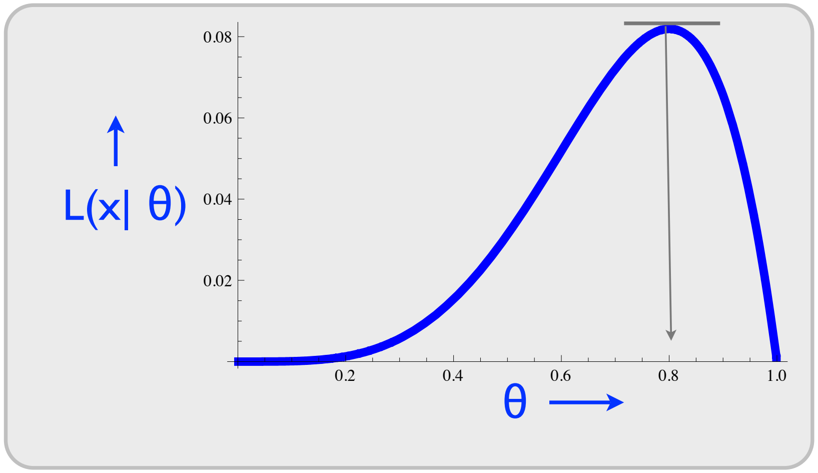
As a function of θ, what θ maximizes the likelihood of the data actually observed. Typical approaches:

Numerical

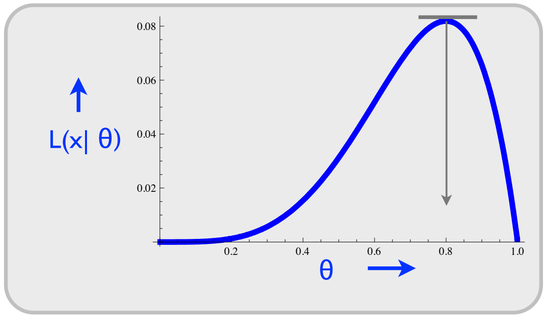
MCMC

Analytical – $\frac{\partial}{\partial\theta} L(\vec{x} \mid \theta) = 0$

EM, etc.



Example 1



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

$n_0 + n_1 = n$; θ = probability of heads

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

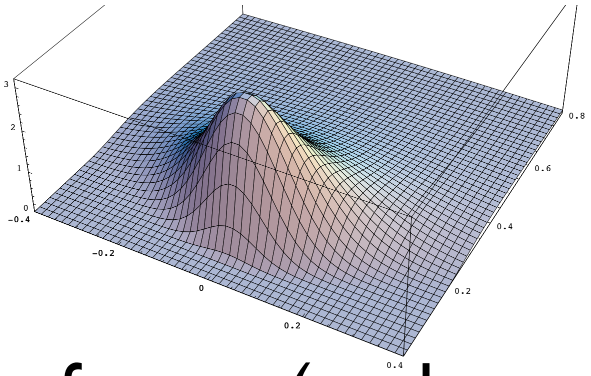
Setting to zero and solving:

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

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 II



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

$\theta = (\theta_A, \theta_C, \theta_G, \theta_T)$ 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

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:

MLE = position specific frequencies

Reminder

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

“Similarity” of Distributions: Relative Entropy

AKA Kullback-Liebler Distance/Divergence,
AKA Information Content

Given distributions P, Q

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log \frac{P(x)}{Q(x)} \quad \geq 0$$

Notes:

$$\text{Let } P(x) \log \frac{P(x)}{Q(x)} = 0 \text{ if } P(x) = 0 \text{ [since } \lim_{y \to 0} y \log y = 0]$$

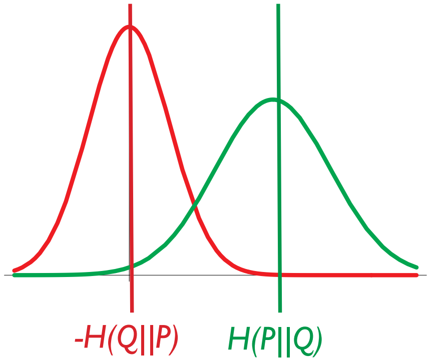
$$\text{Undefined if } 0 = Q(x) < P(x)$$

WMM: How "Informative"? Mean score of site vs bkg?

For any fixed length sequence *x*, let
P(x) = Prob. of *x* according to WMM
Q(x) = Prob. of *x* according to background

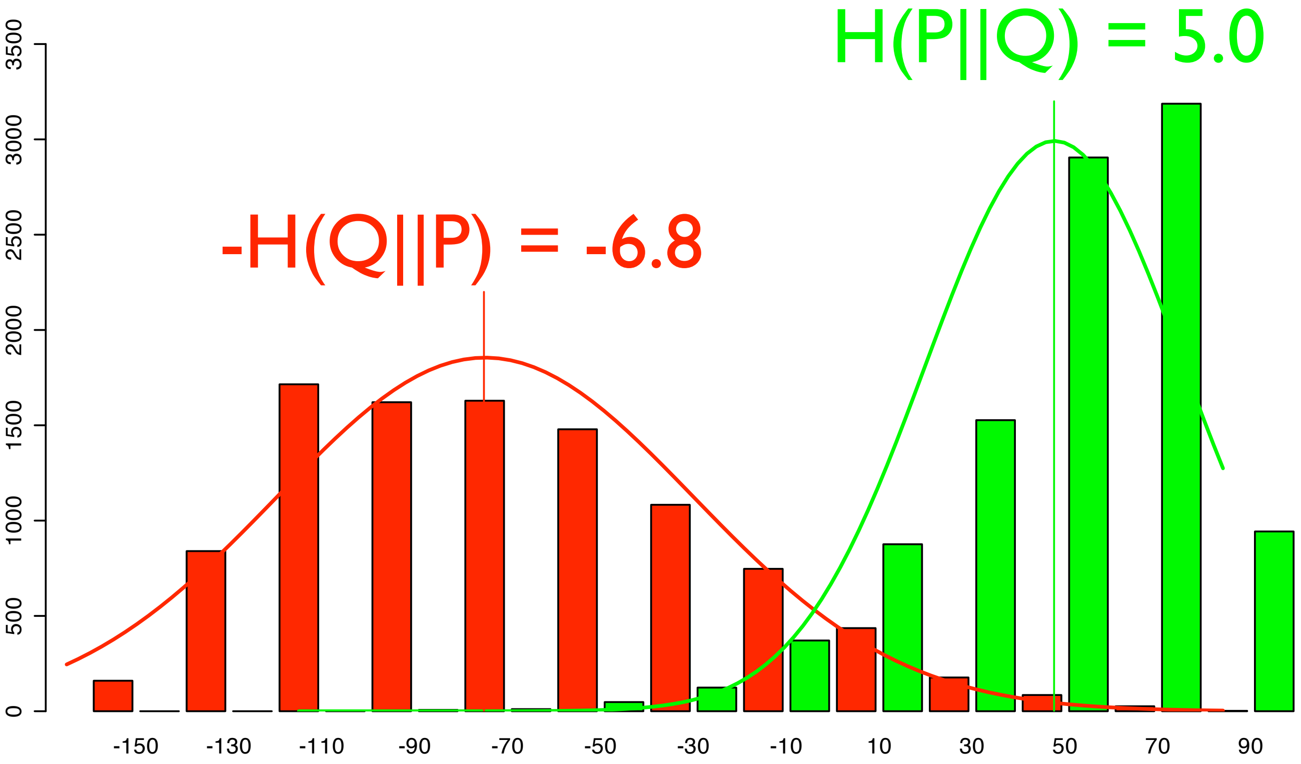
Relative Entropy:

$$H(P||Q) = \sum_{x \in \Omega} P(x) \log_2 \frac{P(x)}{Q(x)}$$



H(P||Q) is *expected log likelihood score* of a sequence randomly chosen from *WMM*;
-H(Q||P) is expected score of *Background*

WMM Scores vs Relative Entropy



On average, foreground model scores > background by 11.8 bits
(score difference of 118 on 10x scale used in examples above).

Calculating H
& H per Column

For WMM, based on the assumption of independence between columns:

$$H(P||Q) \quad = \quad \sum_i H(P_i||Q_i)$$

where Pi and Qi are the WMM/background distributions for column i.

Questions

Which columns of my motif are most informative/uninformative?

How wide is my motif, really?

Per-column relative entropy gives a quantitative way to look at such questions

Another WMM example

8 Sequences:

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

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	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2.00
T	-1.00	2.00	-∞

Log-Likelihood Ratio:

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

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	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2.00
T	-1.00	2.00	-∞

Non-uniform

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

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	-∞	-∞	
C	-∞	-∞	-∞	
G	0	-∞	2.00	
T	-1.00	2.00	-∞	
RelEnt	0.70	2.00	2.00	4.70

Non-uniform

LLR	Col 1	Col 2	Col 3	
A	0.74	-∞	-∞	
C	-∞	-∞	-∞	
G	1.00	-∞	3.00	
T	-1.58	1.42	-∞	
RelEnt	0.51	1.42	3.00	4.93

Today’s Summary

It’s important to account for background

Log likelihood scoring naturally does: log(freq/background freq)

Relative Entropy measures “dissimilarity” of two distributions; “information content”; average score difference between foreground & background. Full motif & per column

Motif 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)

Broadly useful approaches - not just for motifs