GS 559

Lecture 12a, 2/12/09
Larry Ruzzo

A little more about motifs

Reflections from 2/10

Bioinformatics:

- Motif scanning stuff was very cool
- Good explanation of max likelihood; good use of examples (2)
- I was confused/lost/overwhelmed; a lot of equations; (but I think I got the big picture) (3)

Python:

- Last python hw was a big step up in difficulty. A scary trend? "After all, we all have other stuff to do besides bang our heads against python" (7)
- Do longer, more complex practice problem in class; homework is getting harder, but in-class practice is not... (2)
- Going through code slowly was "a breath of fresh air"
- What is grep? An re? A module we import? compile? etc.
- How do we use python files *not* in the user folder?
- need more practice with reg exps

Both:

- Print slides portrait, not landscape.
- Post HW solutions online? *they are*
- Lecture was clear, but rushed/class was too short (again). (3) Semesters?
- Real-world examples good, do more (but hard to understand) (2)
- Do more with online databases & tools
- Pls include summary slides for lecture review, like Mary & Bill did
- Appreciate taking time to go over tough stuff slowly, even if we don't finish everything planned

Motifs

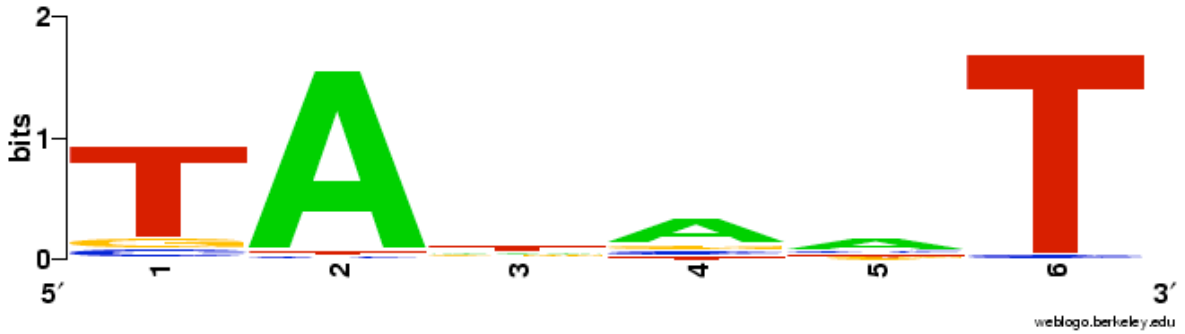
Review, plus a bit more

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

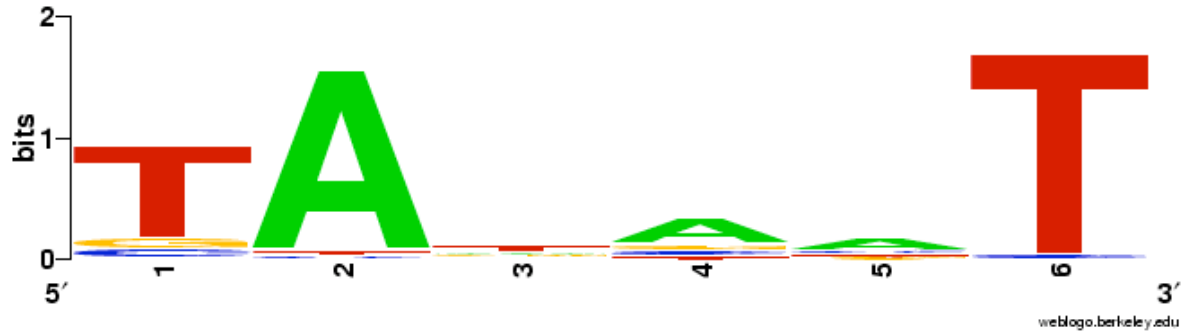
http://weblogo.berkeley.edu



TATA Box Scores

A "Weight Matrix Model" or "WMM"

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



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**

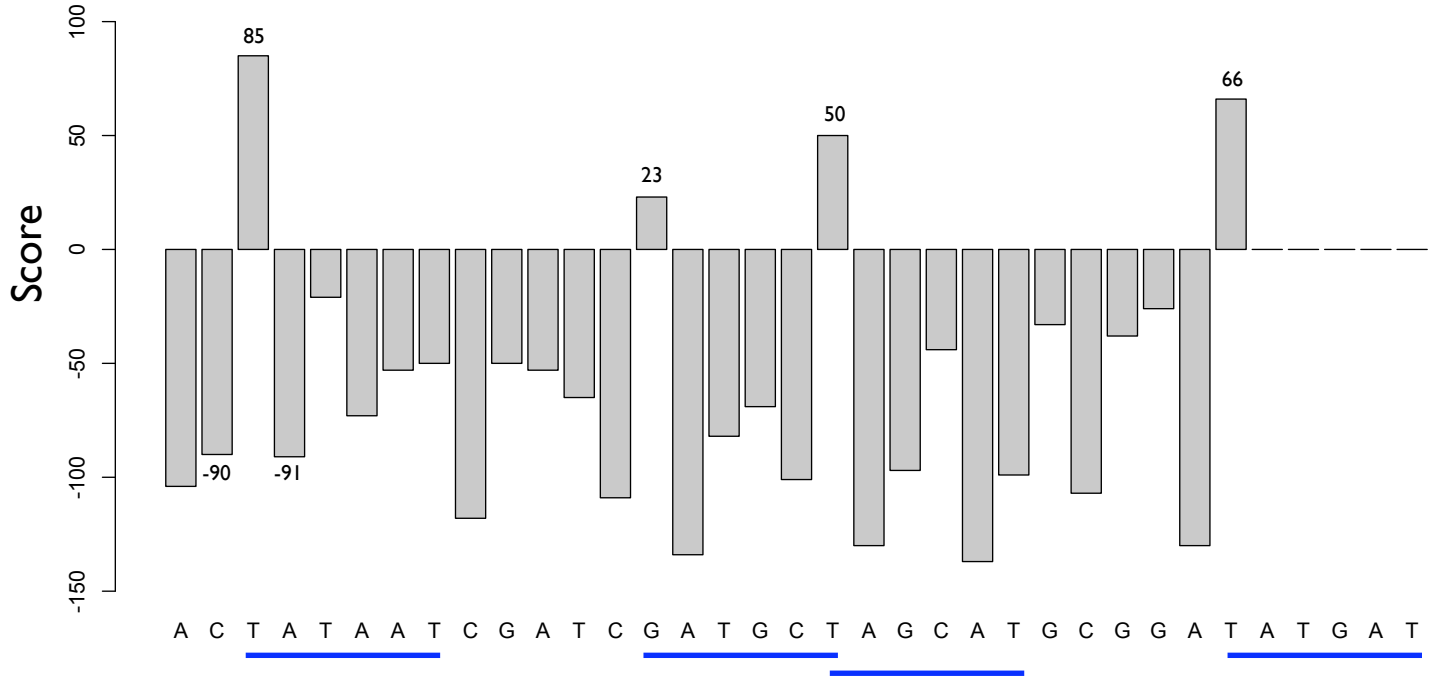
= 85

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

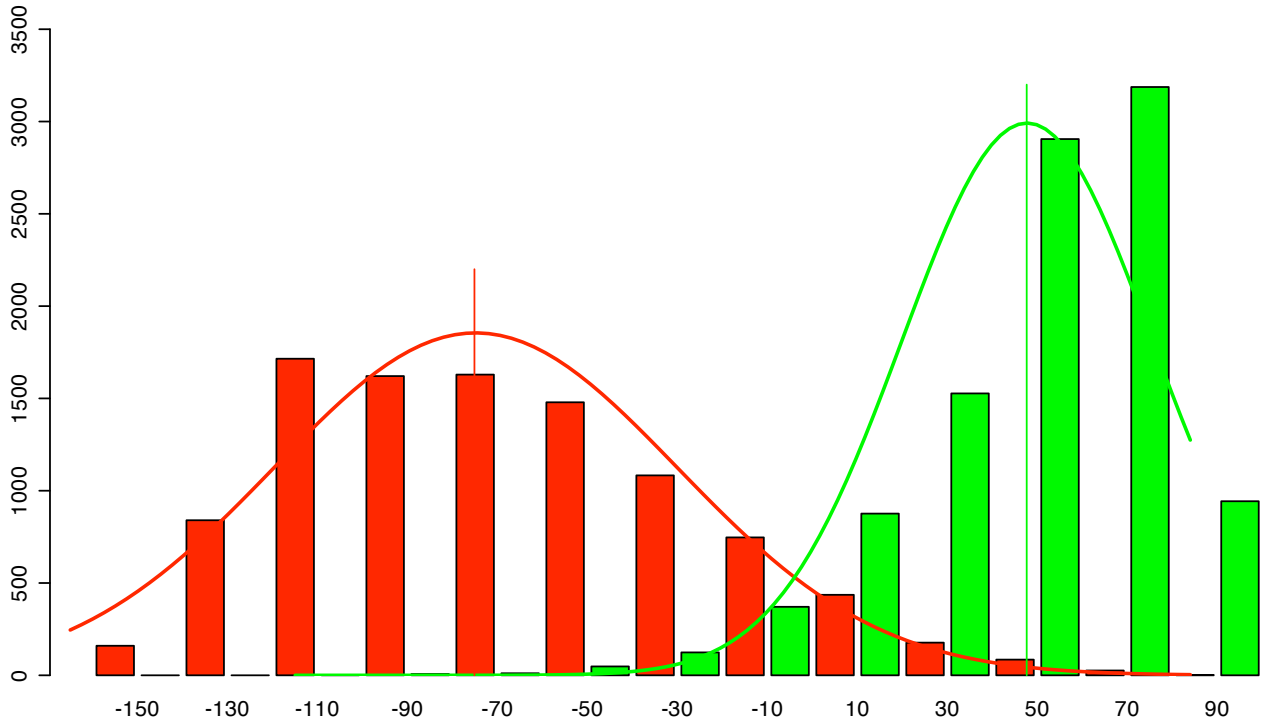
Stormo, Ann. Rev. Biophys. Biophys Chem, 17, 1988, 241-263

Scanning for TATA



Score Distribution

(Simulated)



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 $S = B_1 B_2 ... B_6$:

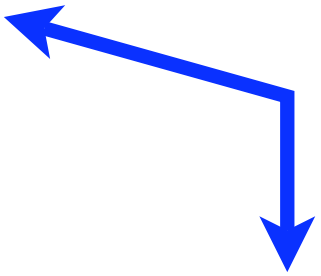
$$\log\left(\frac{P(S|\text{“tata”})}{P(S|\text{“non-tata”})}\right) = \log\frac{\prod_{i=1}^{6} f_{B_i,i}}{\prod_{i=1}^{6} f_{B_i}} = \sum_{i=1}^{6} \log\frac{f_{B_i,i}}{f_{B_i}}$$

Assumes *independence*

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)



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)

Another WMM example

8 Sequences:

ATG
ATG
ATG
ATG
ATG
GTG
GTG
TTG

Log-Likelihood Ratio:

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

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

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	-∞

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:

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

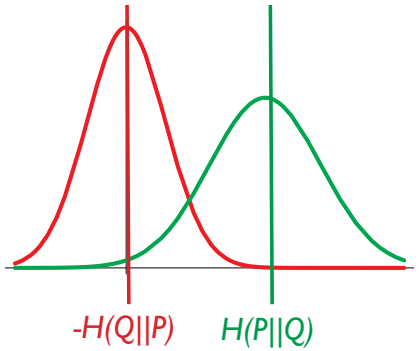
$$\text{Undefined if } 0 = Q(x) < P(x)$$

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

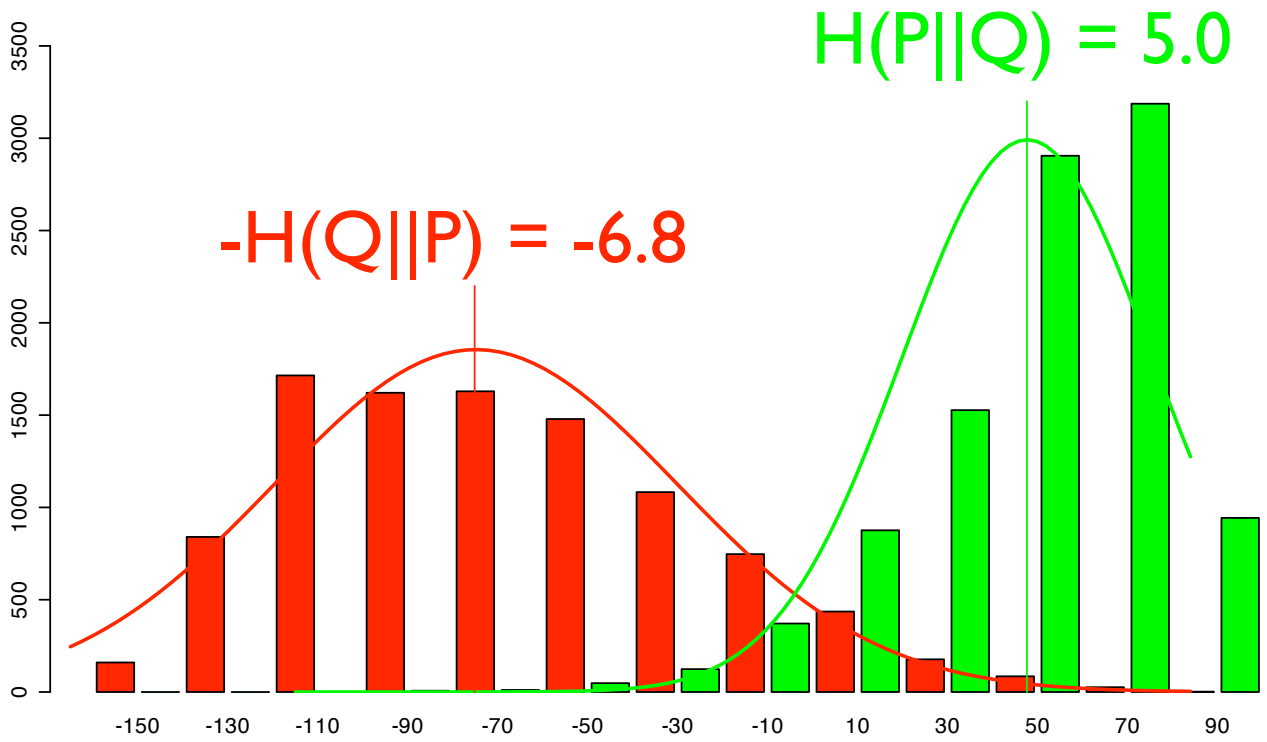
Relative Entropy:

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



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).

More 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 questions like these

For WMM, you can show (based on the assumption of independence between columns), that :

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

where P_i and Q_i are the WMM/background distributions for column i.

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

Pseudocounts

Freq/count of 0 ⇒ $-\infty$ 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

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