GS 559
Winter 2009

Lecture 11
Sequence Motifs

Larry Ruzzo

New Web (but old links should redirect):
http://www.cs.washington.edu/homes/ruzzo/courses/gs559/09wi

Who Am I?

Prof. Computer Science & Engineering
Adjunct Prof., Genome Sciences
Joint Member, FHCRC

Main research interest: noncoding RNA

http://www.cs.washington.edu/homes/ruzzo
ruzzo@cs.washington.edu
554 CSE, 543-6298

Office Hours: Wednesdays 11:00-1:00, or by appt

2 Minute Responses from 2/5

- good lecture today (3)
- really appreciate your straightforward teaching style
- bootstrapping method portion was very informative
- useful to go over which phylogeny method to use (8)

- I like that the problems start easy then get more complex; nice to work up gradually (2)
- slowly getting the function aspect
- I was pretty lost - felt like we jumped into "sort functions" with little explanation
- If I wanted to continue learning about programming & computational biology (building on this course), what should I take? (but I have no comp sci background)
  - *CSE 427/527 for comp bio; 142/143/373/417 for general cs*
  - *GS 540/541, pop gen, phylo*
  - *Biostat Statgen sequence*
  - *BioE*
  - *BHI*
  - *Talk to Bill, Mary or me*

Outline

Bioinformatics:

- Sequence Motifs
- Review - hypothesis testing & maximum likelihood
- Sequence Logos
- Weight Matrix Models (WMMs)
  - aka Position Specific Scoring Matrices (PSSMs, possums)
  - aka 0th order Markov models
- Construction, statistics, uses

Programming:

- Grep and regular expressions

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 $\approx$ 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).

Related Problem: 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

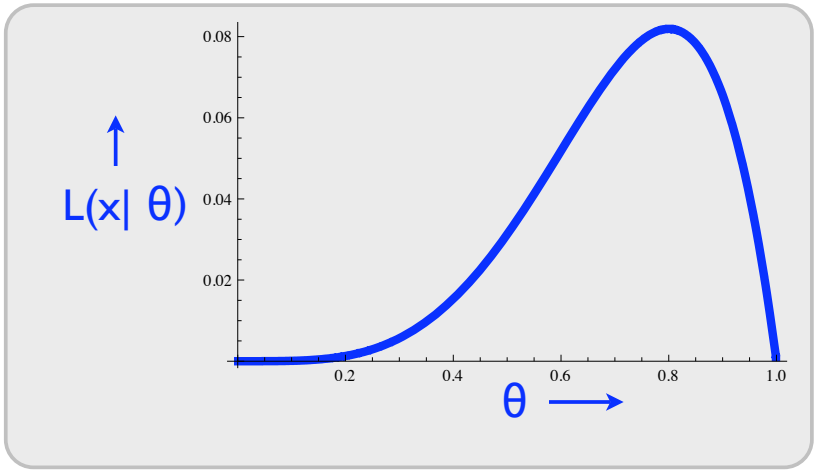
One (of many) approaches to param. est.

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

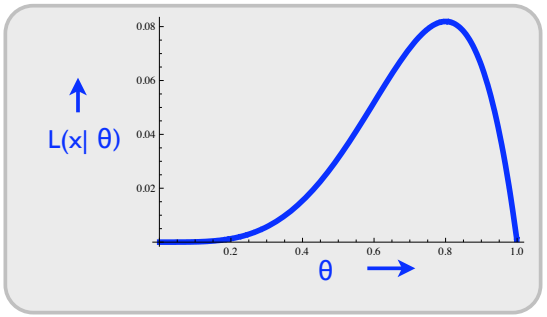
$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

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 1



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

θ = probability of heads

$$L(x_1, x_2, \ldots, x_n \mid \theta) = (1-\theta)^{n_0}\theta^{n_1}$$

$$\log L(x_1, x_2, \ldots, x_n \mid \theta) = n_0 \log(1-\theta) + n_1 \log\theta$$

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

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)

Sequence Motifs

Motif: "a recurring salient thematic element"

E.g., *structural* motifs in proteins (zinc finger, H-T-H, leucine zipper, ... are various DNA binding motifs)

E.g., the DNA *sequence* motifs to which these proteins bind - e.g. , one leucine zipper dimer might bind (with varying affinities) to 10s or 100s or 1000s of similar sequences

E. coli Promoters

"TATA Box" ~ 10bp upstream of transcription start

How to define it?

- *Consensus* is TATAAT
- BUT all differ from it
- Allow k mismatches?
- Equally weighted?
- Wildcards like R,Y? ({A,G}, {C,T}, resp.)

```
TACGAT
TAAAAT
TATACT
GATAAT
TATGAT
TATGTT
```

E. coli Promoters

"TATA Box" - consensus TATAAT

~10bp upstream of transcription start

Not exact: of 168 studied (mid 80's)

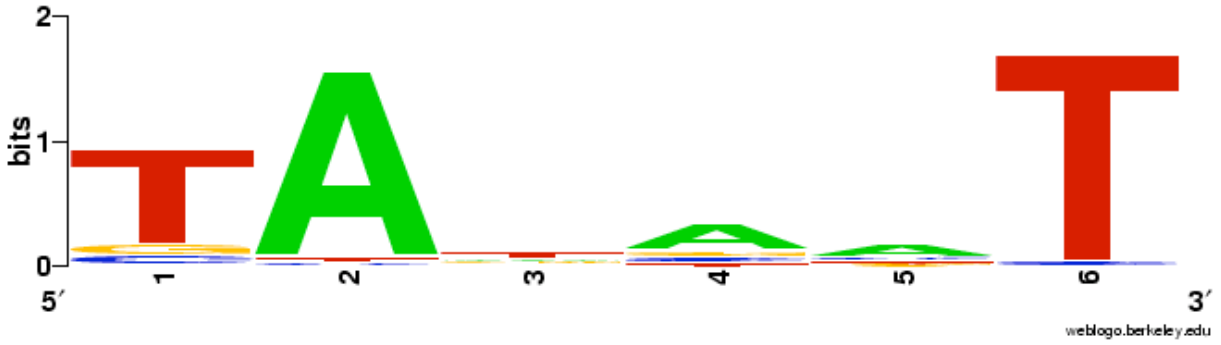
- nearly all had 2/3 of TAxyzT
- 80-90% had all 3
- 50% agreed in each of x,y,z
- no perfect match

(Other common features at -35, etc.)

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

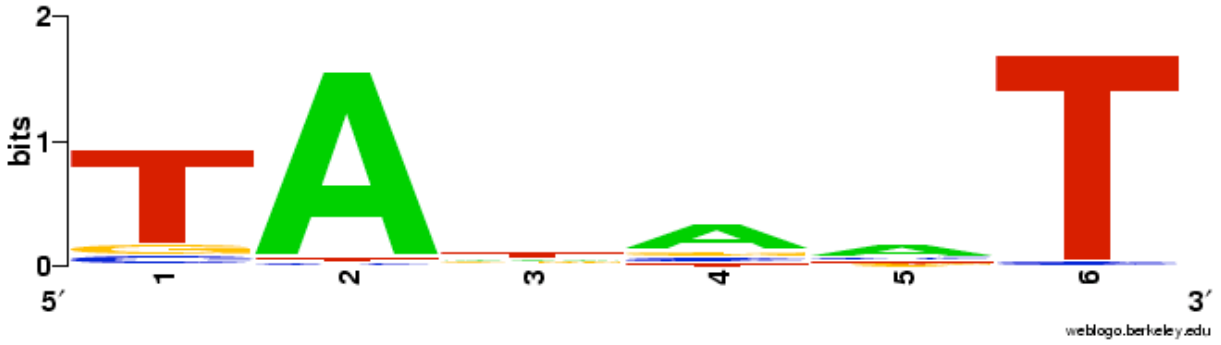
Sequence Logo



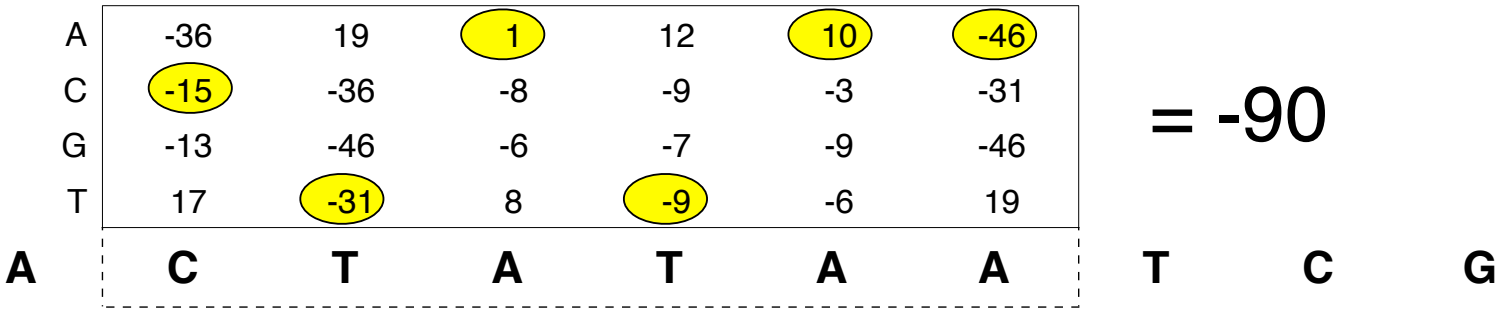
TATA Box Scores

A "Weight Matrix Model" or "WMM"

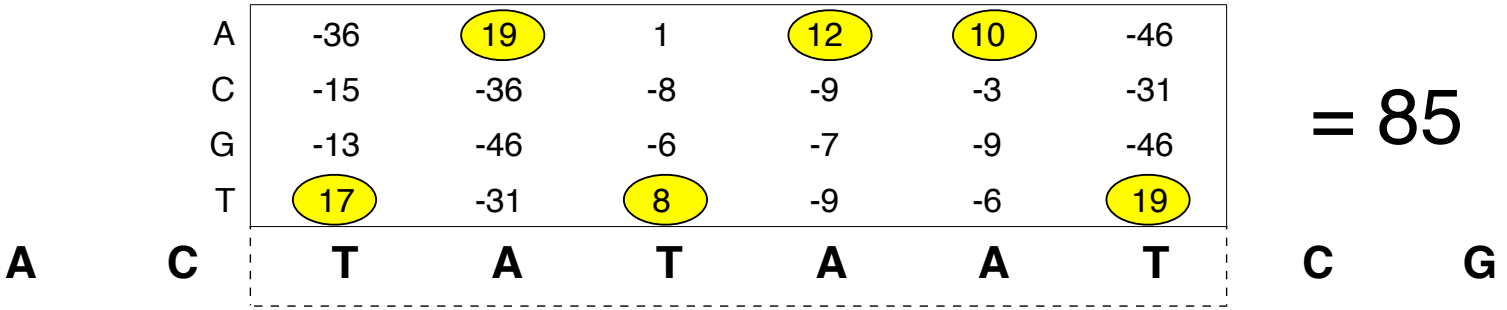
pos base	1	2	3	4	5	6
A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46 (?)
T	17	-31	8	-9	-6	19



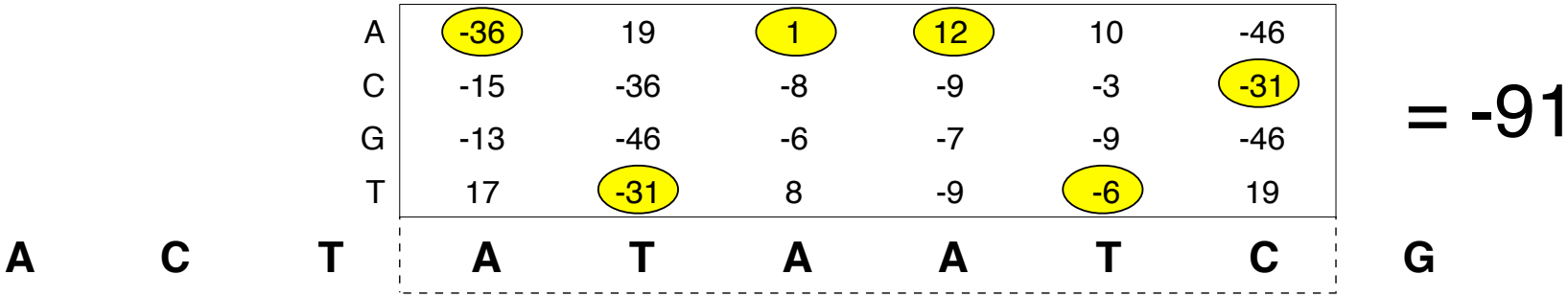
Scanning for TATA



= -90



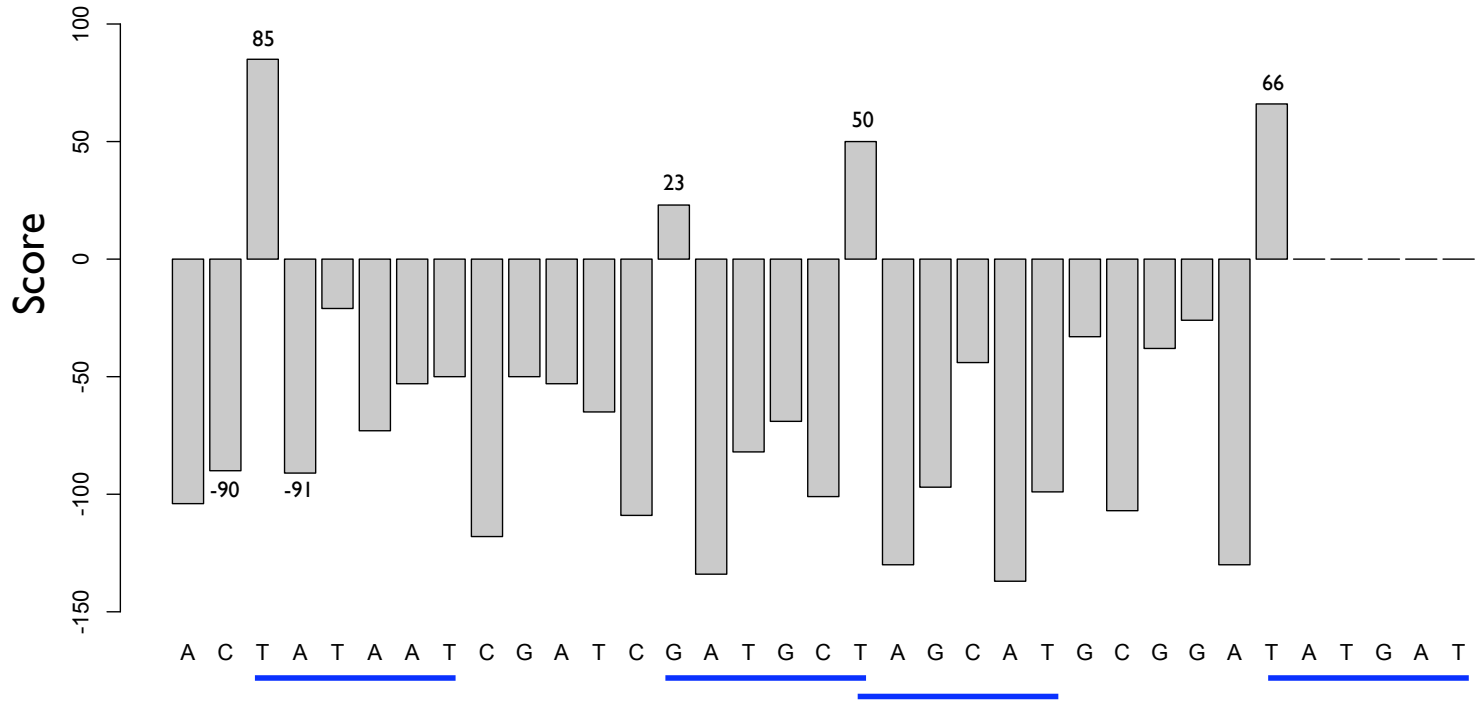
= 85



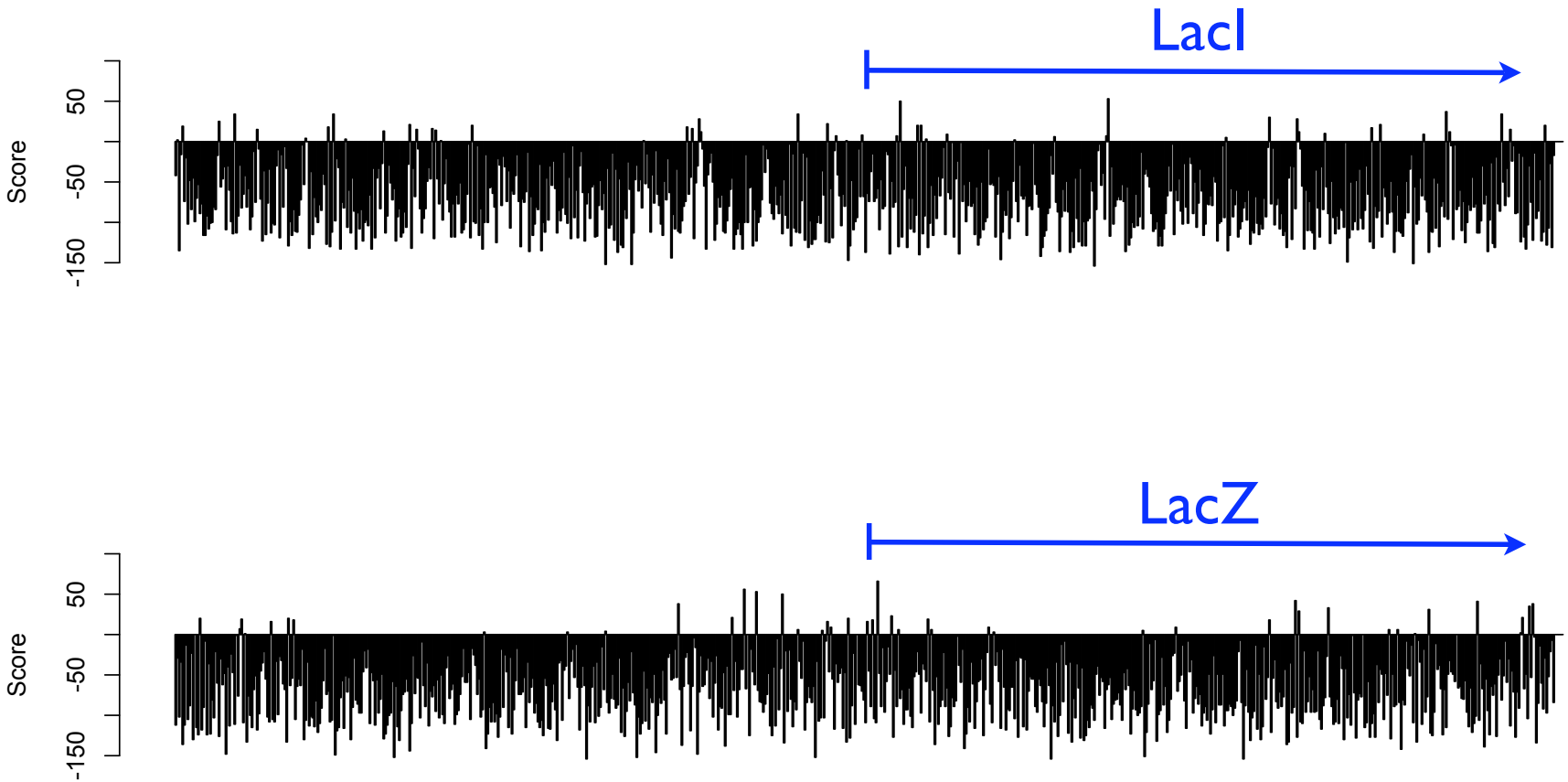
= -91

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

Scanning for TATA

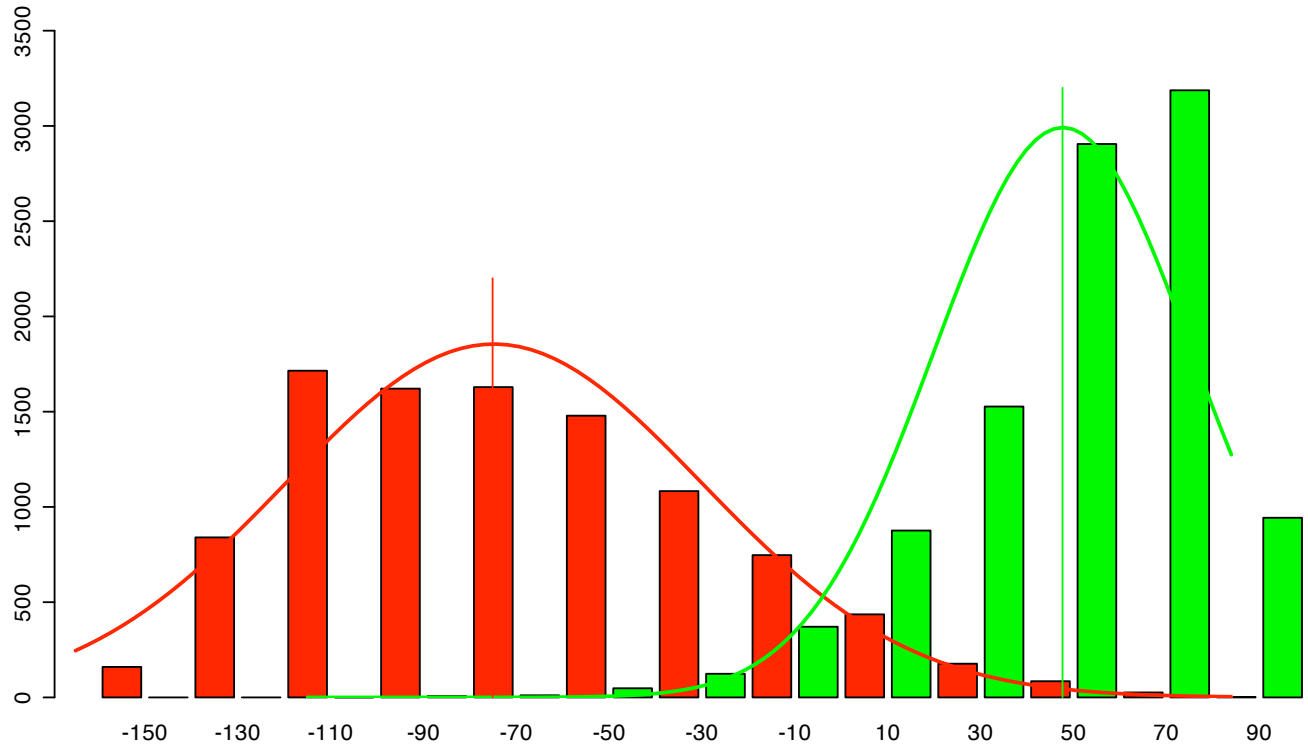


TATA Scan at 2 genes



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*

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: like coin flips or dice rolls, count frequencies per position.

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]

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

How-to Questions

Given aligned motif instances, build model?

- Frequency counts (above, maybe w/ pseudocounts)

Given a model, find (probable) instances

- Scanning, as above

Given unaligned strings thought to contain a motif, find it? (e.g., upstream regions of co-expressed genes)

- Hard ... maybe another lecture.

WMM Summary

Weight Matrix Model (aka Position Specific Scoring Matrix, PSSM, "possum", 0th order Markov models)

Simple statistical model assuming independence between adjacent positions

To build: align, count (+ pseudocount) letter frequency per position, log likelihood ratio to background

To scan: add per position scores, compare to threshold, slide

Databases & tools: Transfac, Jaspar, MEME/MAST, ...