# GS 559

## Winter 2010

Lecture 11
Sequence Motifs

Larry Ruzzo

New Web Soon(but old links should redirect):
http://www.cs.washington.edu/homes/ruzzo/courses/gs559/10wi

# 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@uw.edu
554 CSE, 543-6298

Office Hours: Mondays 2:30-3:20, or **by appt**

# Outline

- Bioinformatics:
  - Sequence Motifs
  - Sequence Logos
  - Weight Matrix Models (WMMs)
    - aka Position Specific Scoring Matrices (PSSMs, possums)
    - aka 0th order Markov models
  - Construction, statistics, uses
- Programming:
  - Regular expressions

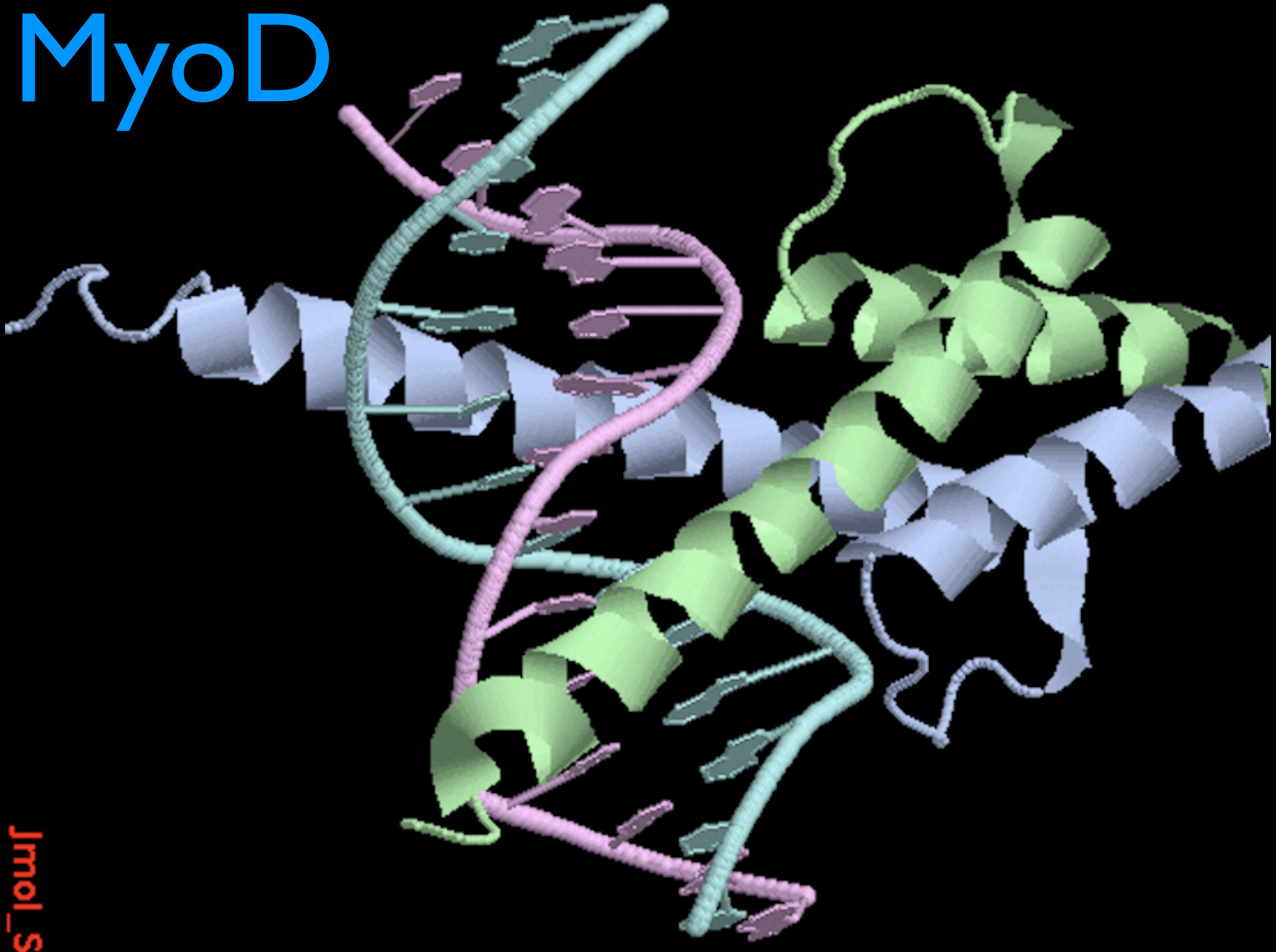
# Motifs

*Motif*: “a recurring salient thematic element”

# Motifs

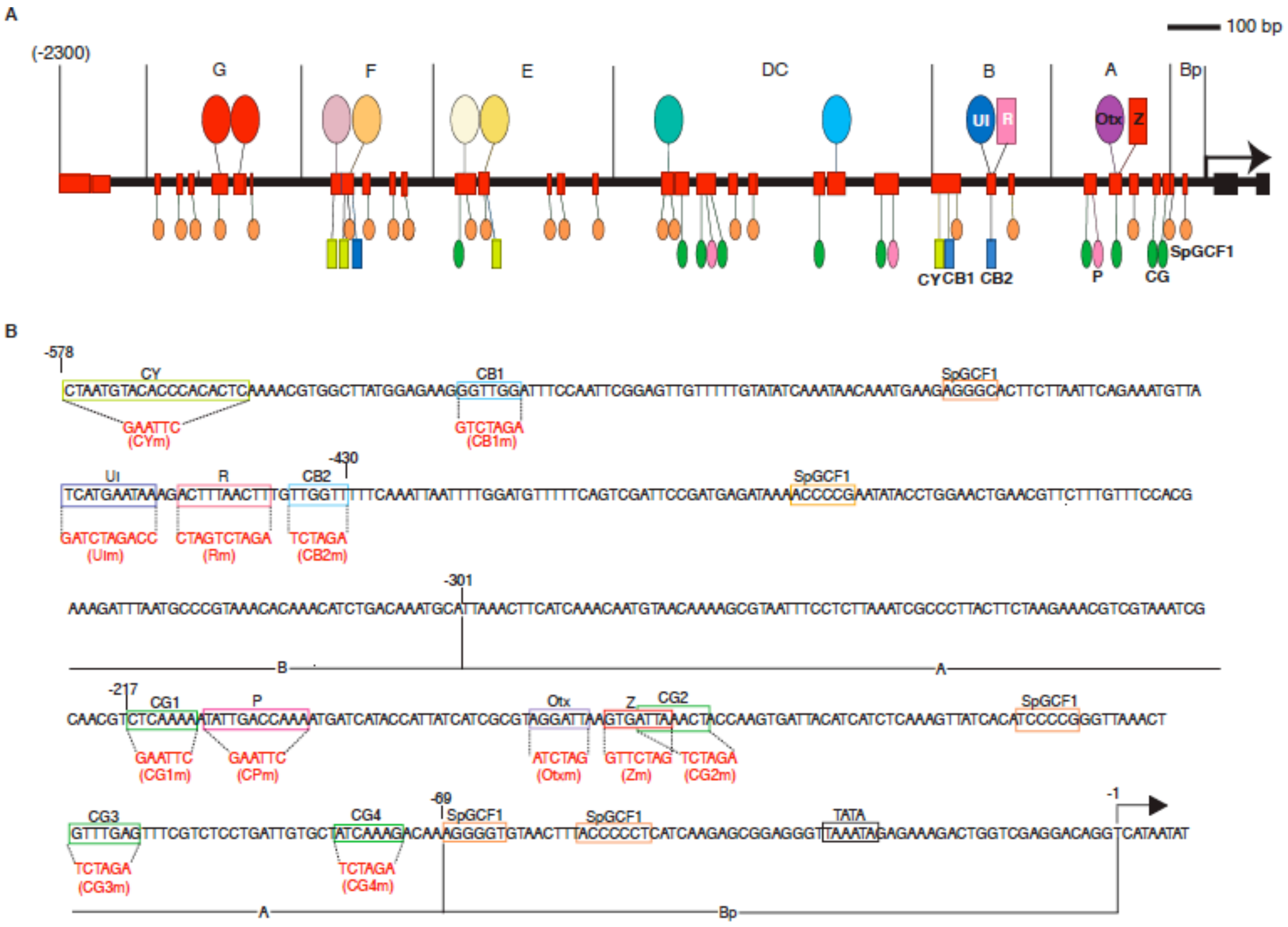
*Motif*: “a recurring salient thematic element”

# MyoD



http://www.rcsb.org/pdb/explore/jmol.do?structureId=1MDY&bionumber=1

# Sea Urchin - Endo16


A
100 bp
(-2300)
G
F
E
DC
B
A
Bp
UI
R
Otx
Z
SpGCF1
CY
CB1
CB2
P
CG
B
-578
CY
CB1
SpGCF1
CTAATGTACACCCACACTCAAAACGTGGCTTATGGAGAAGGGTTGGATTTCCAATTCGGAGTTGTTTTTGTATATCAAATAACAAATGAAGAGGGCACTTCTTAATTCAGAAATGTTA
GAATTC
(CYm)
GTCTAGA
(CB1m)
-430
UI
R
CB2
SpGCF1
TCATGAATAAAGACTTTAACTTTGTTGGTTTTTCAAATTAATTTTGGATGTTTTTCAGTCGATTCCGATGAGATAAAACCCCGAATATACCTGGAACTGAACGTTCTTTGTTTCCACG
GATCTAGACC
(UIm)
CTAGTCTAGA
(Rm)
TCTAGA
(CB2m)
-301
AAAGATTTAATGCCCGTAAACACAAACATCTGACAAATGCATTAAACTTCATCAAACAATGTAACAAAAGCGTAATTTCCTCTTAAATCGCCCTTACTTCTAAGAAACGTCGTAAATCG
B
A
-217
CG1
P
Otx
Z
CG2
SpGCF1
CAACGTCTCAAAAATATTGACCAAAATGATCATACCATTATCATCGCGTAGGATTAAGTGATTAAACTACCAAGTGATTACATCATCTCAAAGTTATCACATCCCCGGGTTAAACT
GAATTC
(CG1m)
GAATTC
(CPm)
ATCTAG
(Otxm)
GTTCTAG
(Zm)
TCTAGA
(CG2m)
CG3
CG4
-69
SpGCF1
SpGCF1
TATA
-1
GTTTGAGTTTCGTCTCCTGATTGTGCTATCAAAGACAAAGGGGTGTAACTTTACCCCCTCATCAAGAGCGGAGGGTTAAATAGAGAAAGACTGGTCGAGGACAGGTCATAATAT
TCTAGA
(CG3m)
TCTAGA
(CG4m)
A
Bp

# 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

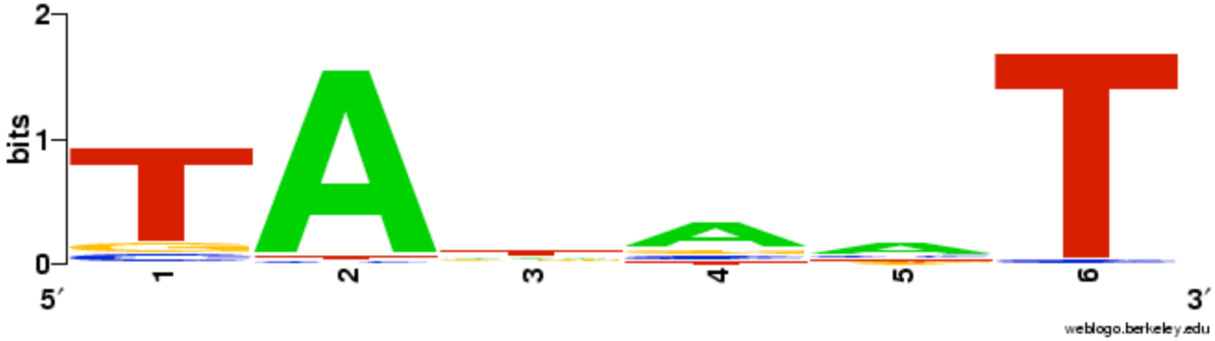
(Other common features at -35, etc.)

# TATA Box Frequencies

| pos<br>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

| pos<br>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<br>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)

# 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** |

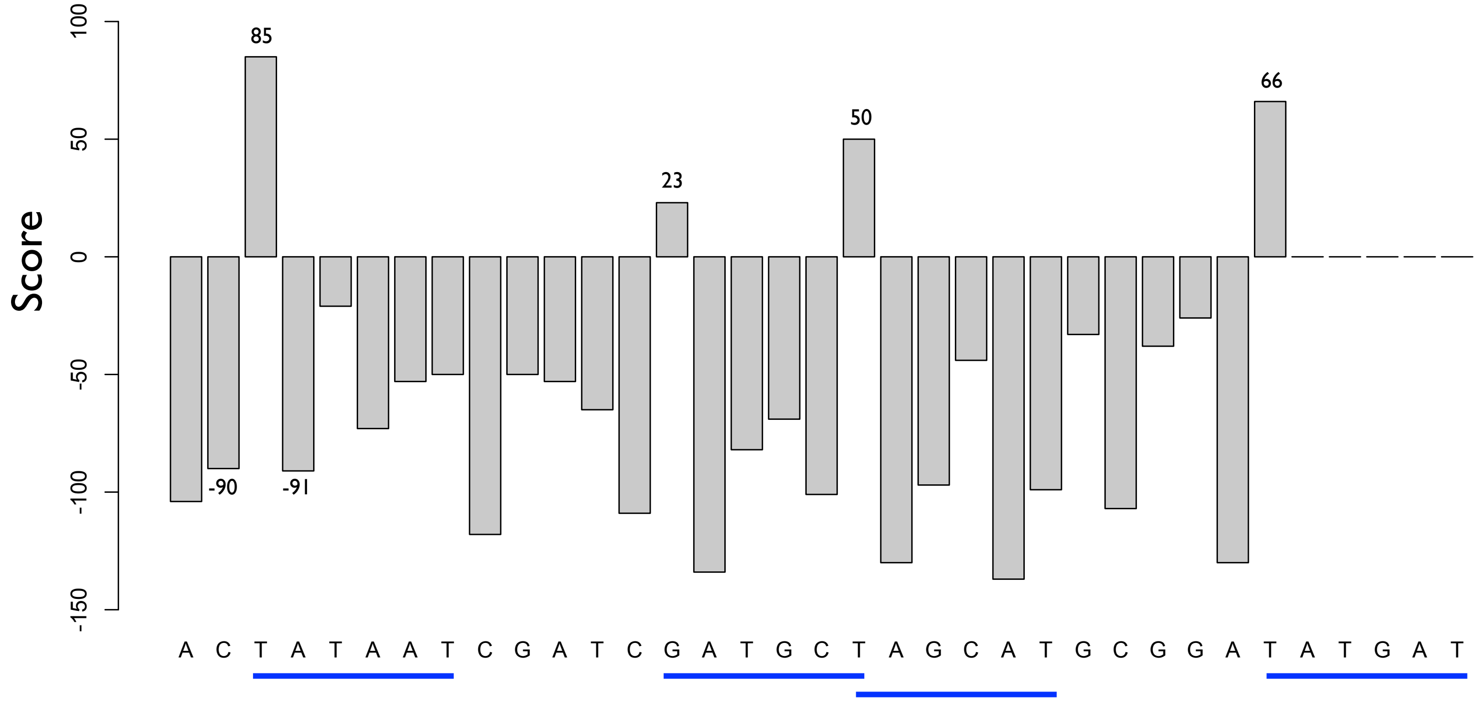
= 85

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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** |

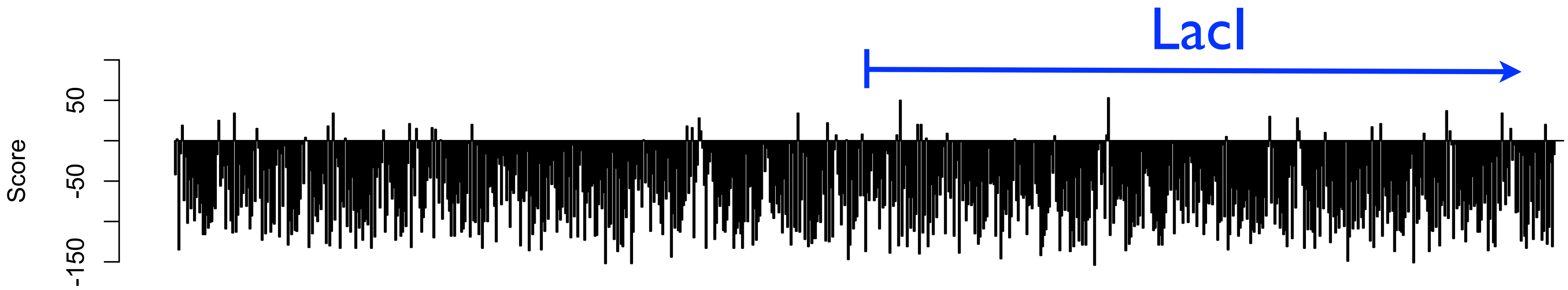
= -91

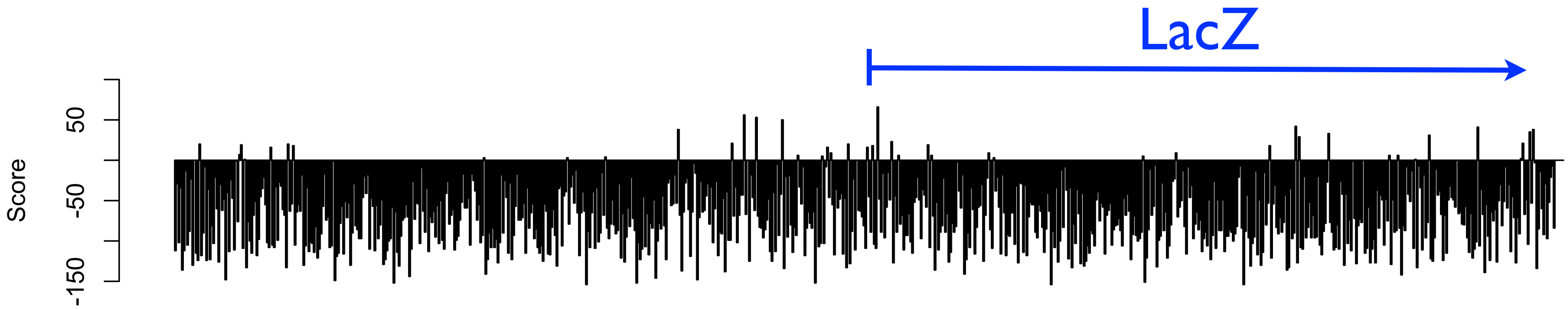
Stormo, Ann. Rev. Biophys. Biophys Chem, 17, 1988, 241-263

# Scanning for TATA



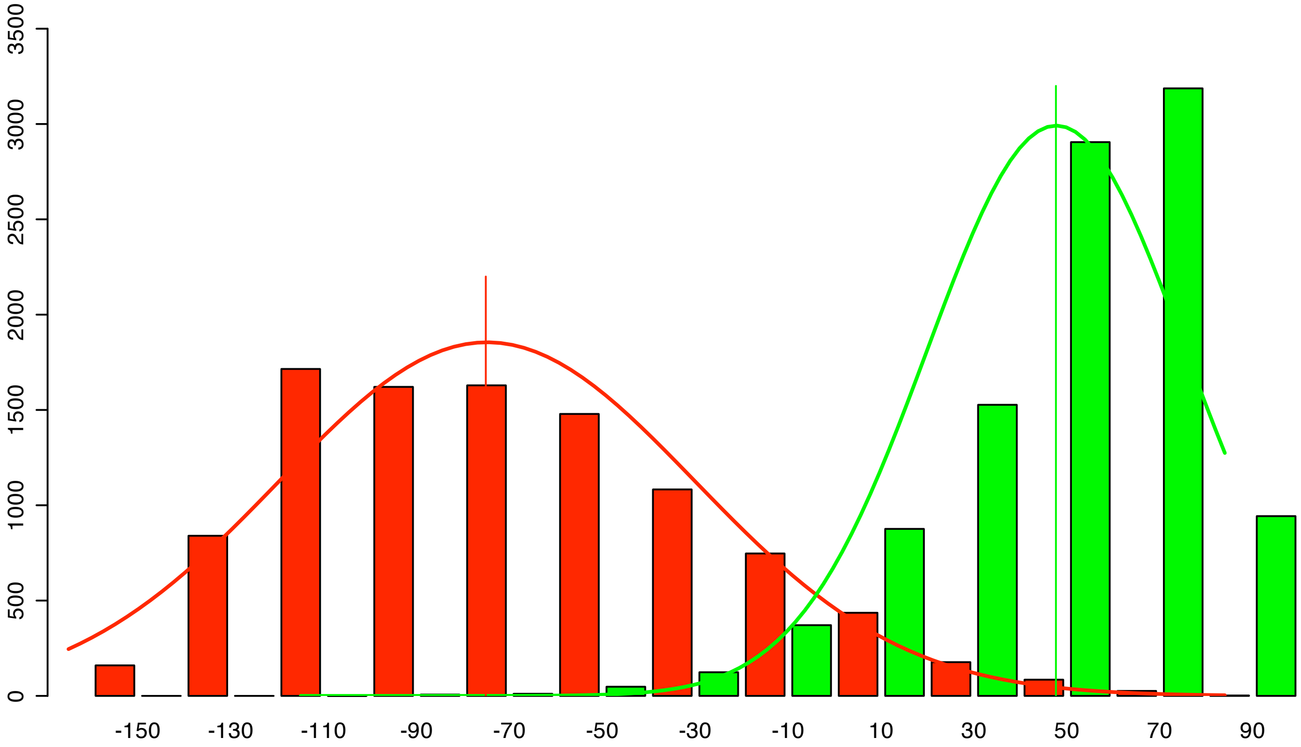
# TATA Scan at 2 genes





# Score Distribution

(Simulated)



# 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]

# What's best WMM?

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

Answer: count frequencies per position.

More justification next time, but if you saw 900 Heads in1000 coin flips, you'd perhaps estimate

P(Heads) = 900/1000

# 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

# 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, ...