CSE 427

Autumn 2015

Motifs: Representation & Discovery

Outline

- Previously: Learning from data
  - MLE: Max Likelihood Estimators
  - EM: Expectation Maximization (MLE w/hidden data)
- These Slides:
  - Bio: Expression & regulation
    - Expression: creation of gene products
    - Regulation: when/where/how much of each gene product; complex and critical
  - Comp: using MLE/EM to find regulatory motifs in biological sequence data

Gene Expression & Regulation

Gene Expression

Recall a *gene* is a DNA sequence for a protein

To say a gene is *expressed* means that it

- is *transcribed* from DNA to RNA
- the mRNA is *processed* in various ways
- is *exported* from the nucleus (eukaryotes)
- is *translated* into protein

A key point: not all genes are expressed all the time, in all cells, or at equal levels

RNA Transcription

Some genes heavily transcribed
(many are not)

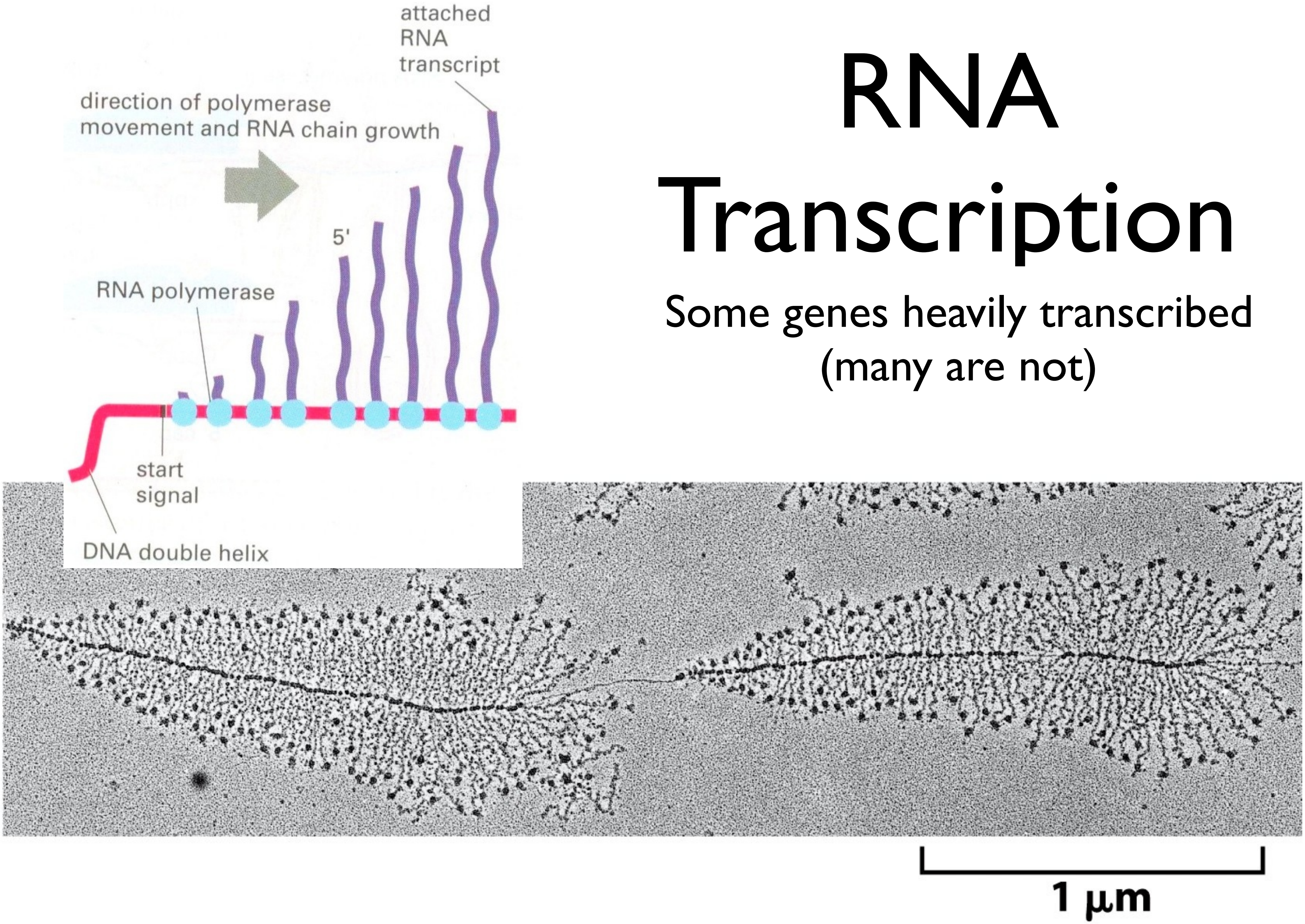


Figure 6-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Regulation

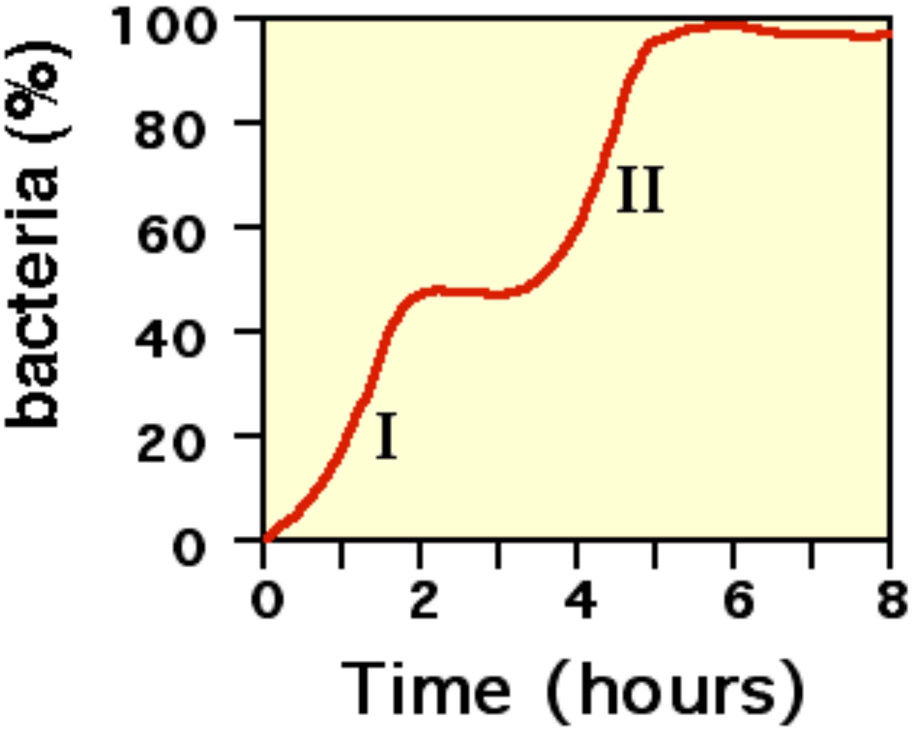
In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes

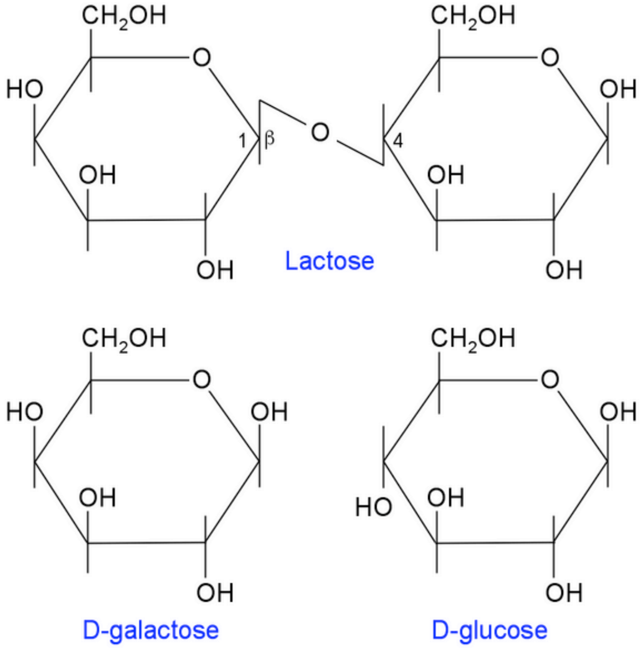
Regulation happens at all steps. E.g., some genes are highly transcribed, some are not transcribed at all, some transcripts can be sequestered then released, or rapidly degraded, some are weakly translated, some are very actively translated, ...

Below, focus on 1st step only:

- transcriptional regulation

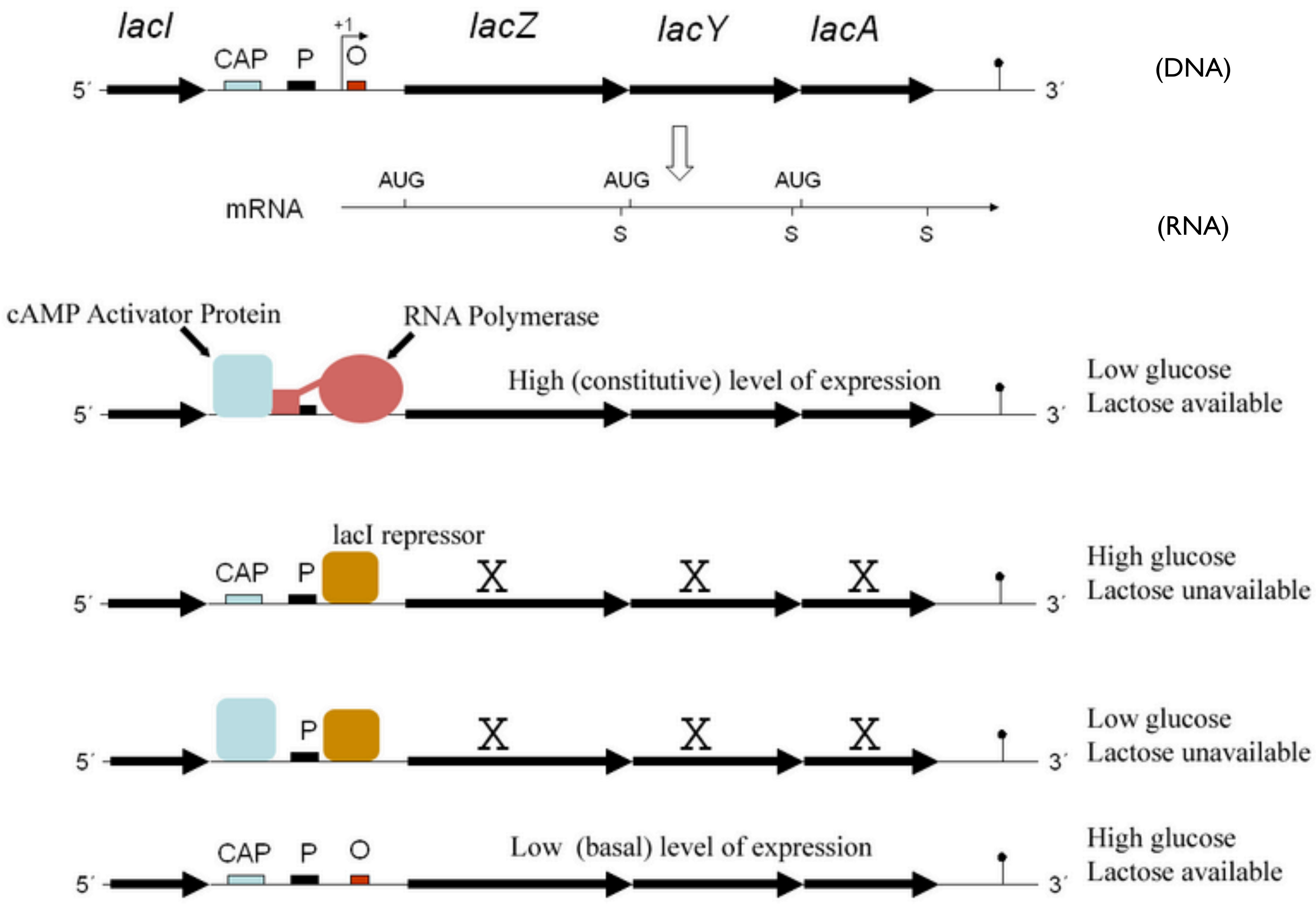
E. coli growth
on glucose + lactose





http://en.wikipedia.org/wiki/Lac_operon

The *lac* Operon and its Control Elements



High (constitutive) level of expression

Low glucose
Lactose available

High glucose
Lactose unavailable

Low glucose
Lactose unavailable

Low (basal) level of expression

High glucose
Lactose available

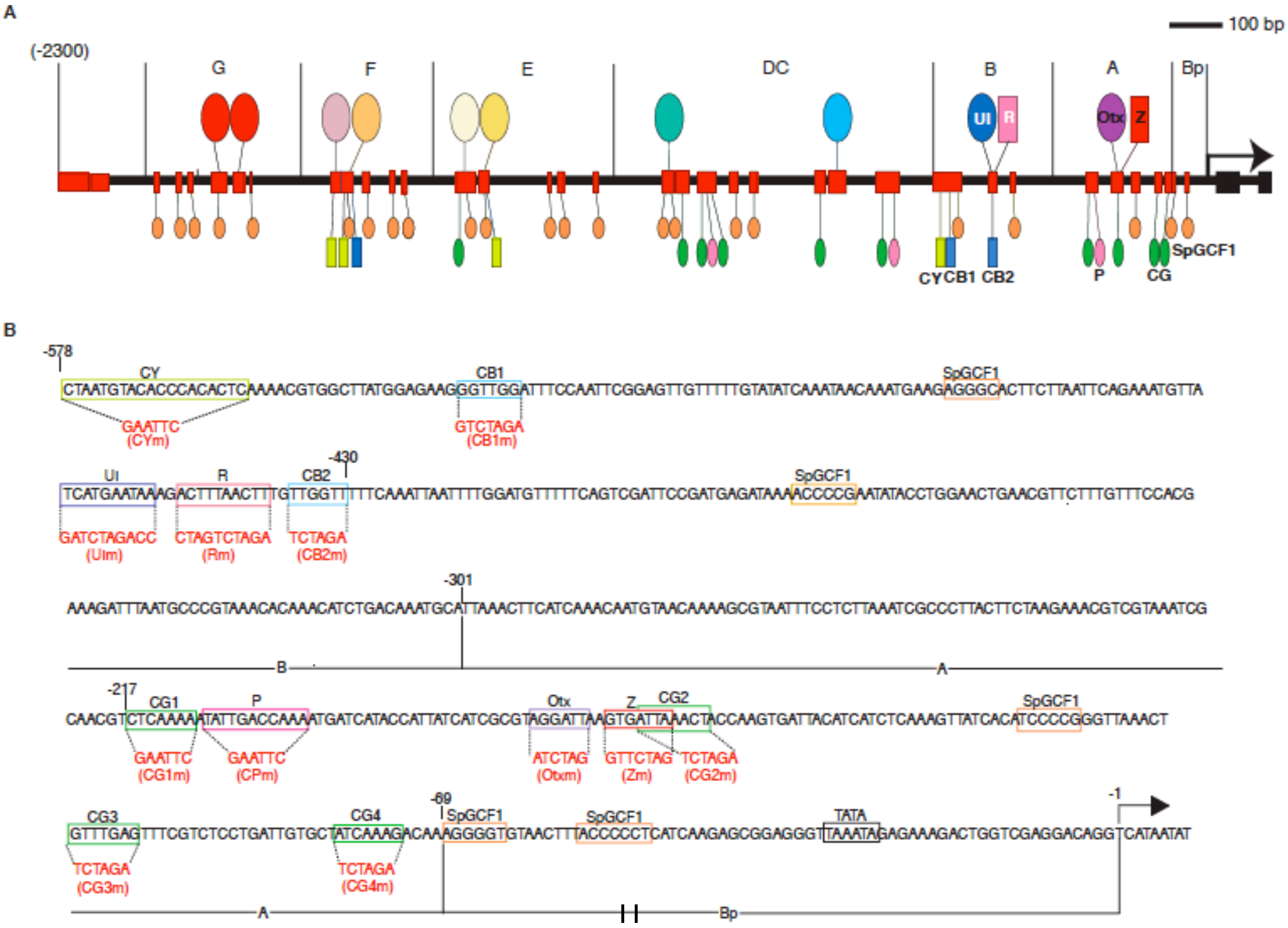
1965 Nobel Prize

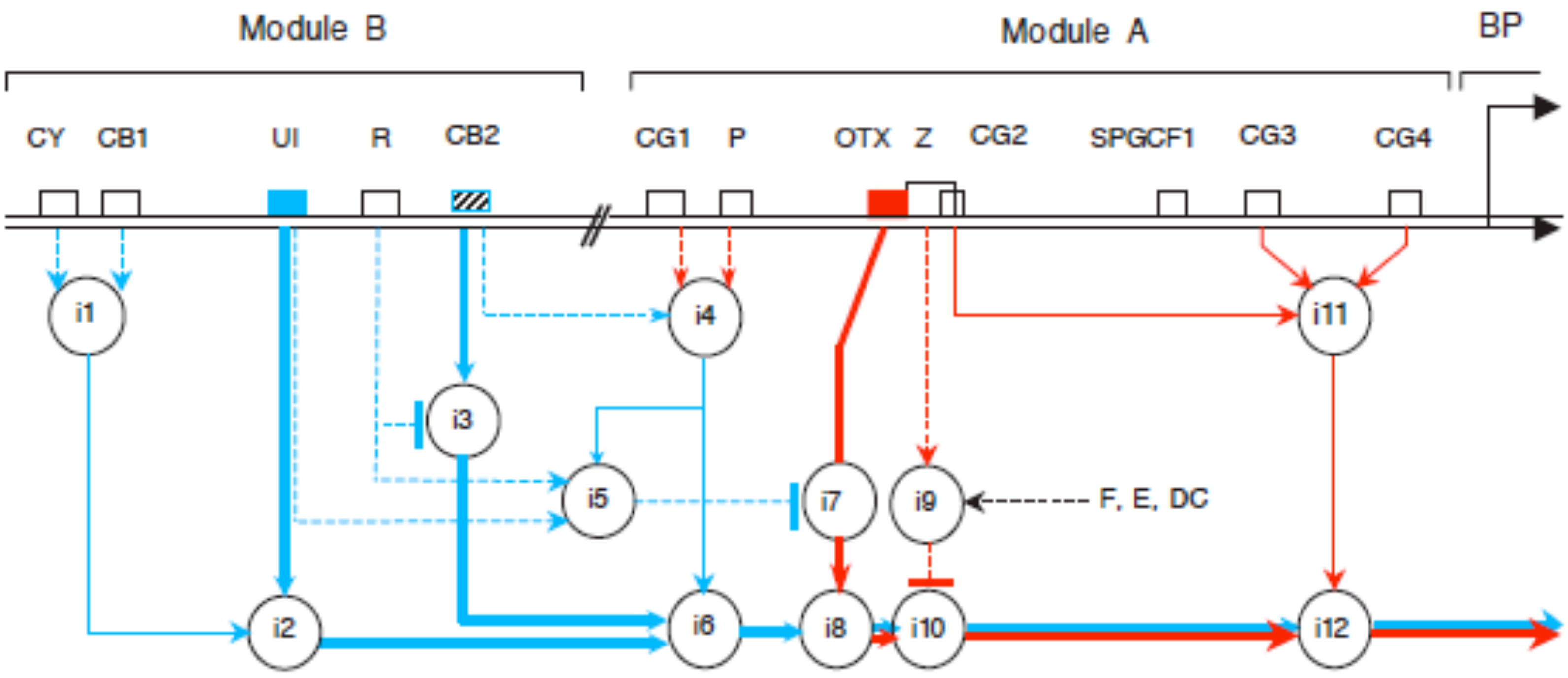
Physiology or Medicine

François Jacob,	Jacques Monod,	André Lwoff
1920-2013	1910-1976	1902-1994

The sea urchin *Strongylocentrotus purpuratus*

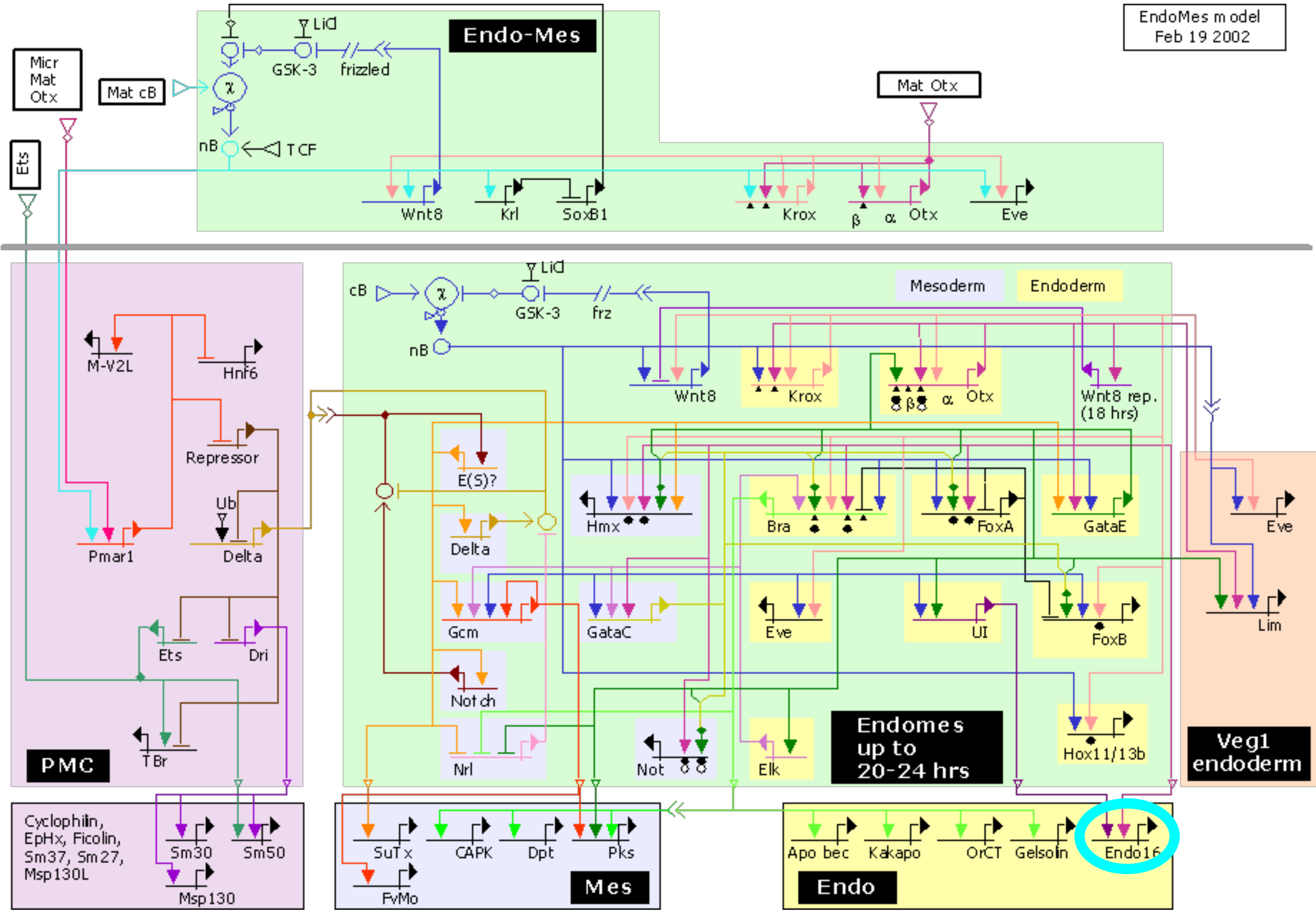
Sea Urchin - Endo16





if CY & CB1	i1 = 1
else	i1 = 0.5
	i2 = i1 · UI (t)
if R	i3 = CB2(t)
else	i3 = k · CB2(t) (1<k<2)
if P & CG1 & CB2	i4 = 2
else	i4 = 0
if UI (t)>threshold & R & i4‡0	i5 = 1
else	i5 = 0
	i6 = i4 · (i2+i3)

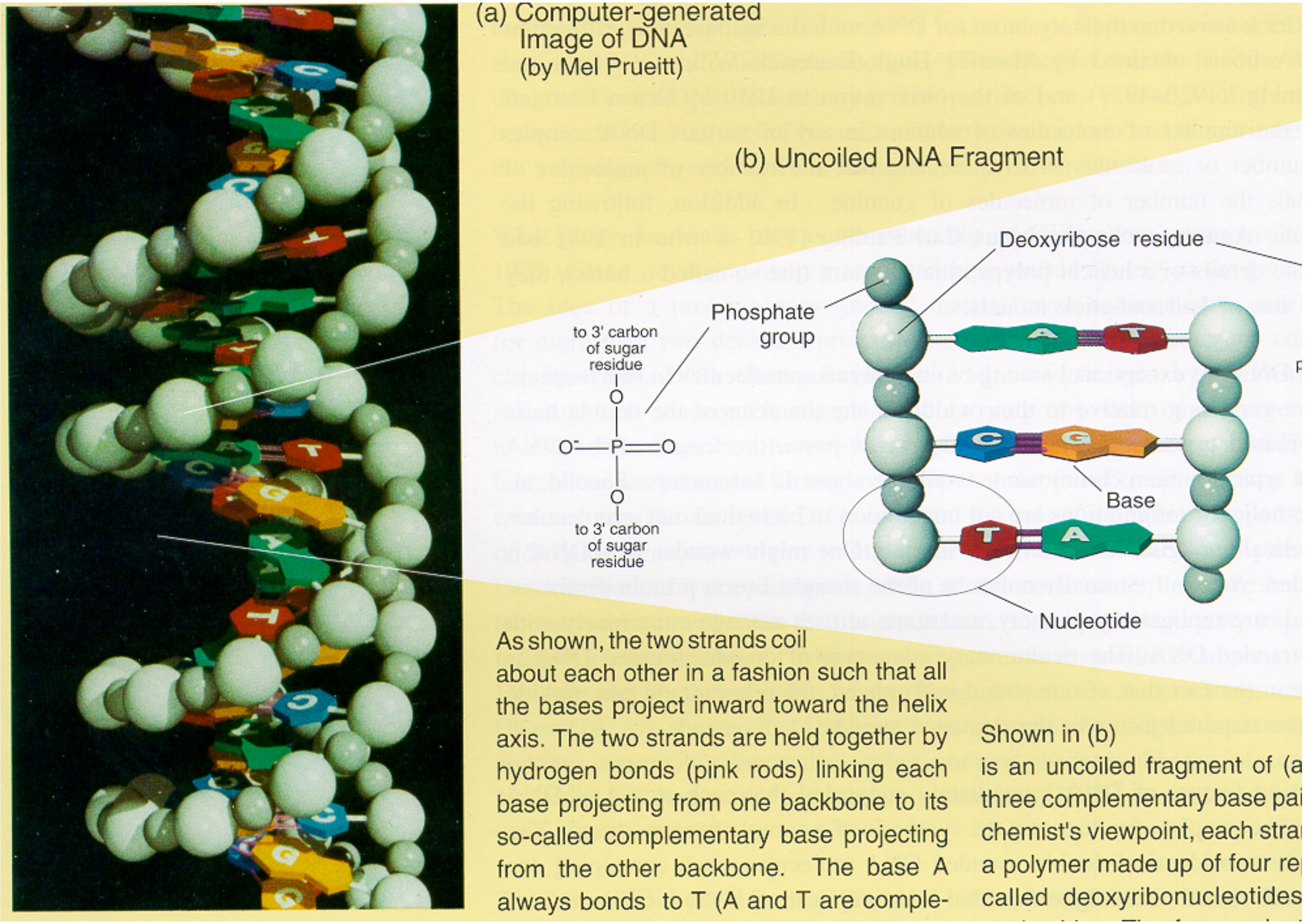
if i5=0	i7 = OTX(t)
else	i7 = 0
	i8 = i6 + i7
if (F or E or DC) & Z	i9 = 1
else	i9 =0
if i9=1	i10 = 0
else	i10 = i8
if (CG2 & CG3 & CG4)	i11 = 2
else	i11 =1
	i12 = i11 · i10


EndoMes model
Feb 19 2002
Endo-Mes
Micr
Mat
Otx
Mat cB
Ets
LiCl
GSK-3
frizzled
nB
TCF
Mat Otx
Wnt8
Krl
SoxB1
Krox
β
α
Otx
Eve
PMC
M-V2L
Hnf6
Repressor
Ub
Pmar1
Delta
Ets
Dri
TBr
cB
LiCl
GSK-3
frz
nB
Mesoderm
Endoderm
Wnt8
Krox
β
α
Otx
Wnt8 rep.
(18 hrs)
E(S)?
Delta
Hmx
Bra
FoxA
GataE
Eve
Gcm
GataC
Eve
Ul
FoxB
Lim
Notch
Nrl
Not
Elk
Endomes
up to
20-24 hrs
Hox11/13b
Veg1
endoderm
Cyclophilin,
EpHx, Ficolin,
Sm37, Sm27,
Msp130L
Sm30
Sm50
Msp130
SuTx
CAPK
Dpt
Pks
FvMo
Mes
Apo bec
Kakapo
OrCT
Gelsolin
Endo16
Endo

DNA Binding Proteins

A variety of DNA binding proteins (so-called “transcription factors”; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes

The Double Helix



(a) Computer-generated Image of DNA (by Mel Prueitt)

(b) Uncoiled DNA Fragment

As shown, the two strands coil about each other in a fashion such that all the bases project inward toward the helix axis. The two strands are held together by hydrogen bonds (pink rods) linking each base projecting from one backbone to its so-called complementary base projecting from the other backbone. The base A always bonds to T (A and T are comple-

Shown in (b)
is an uncoiled fragment of (a
three complementary base pai
chemist's viewpoint, each stra
a polymer made up of four re
called deoxyribonucleotides

Los Alamos Science

In the groove

Different patterns of potential H bonds at edges of different base pairs, accessible esp. in major groove

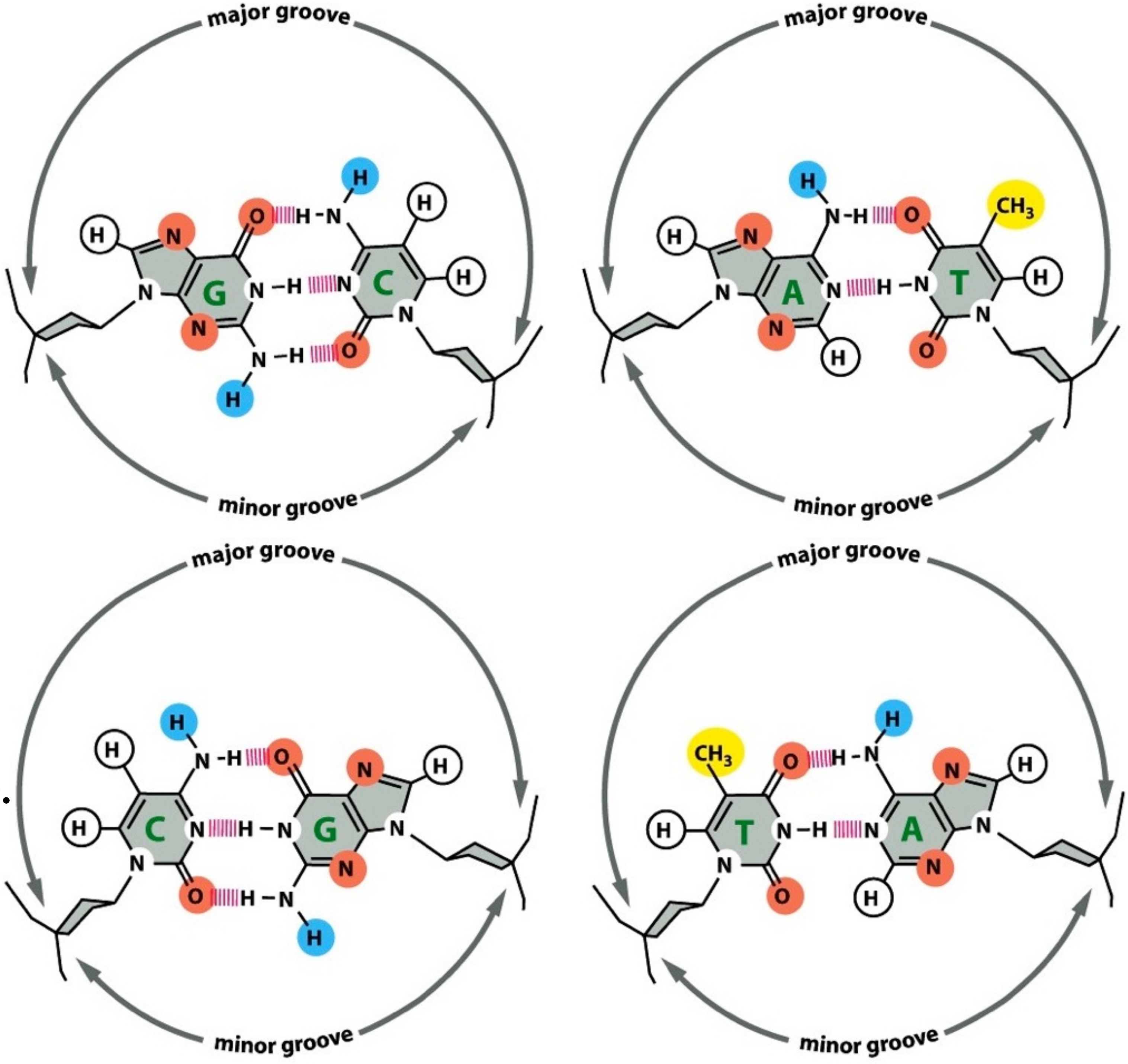


Figure 7-7 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Helix-Turn-Helix DNA Binding Motif

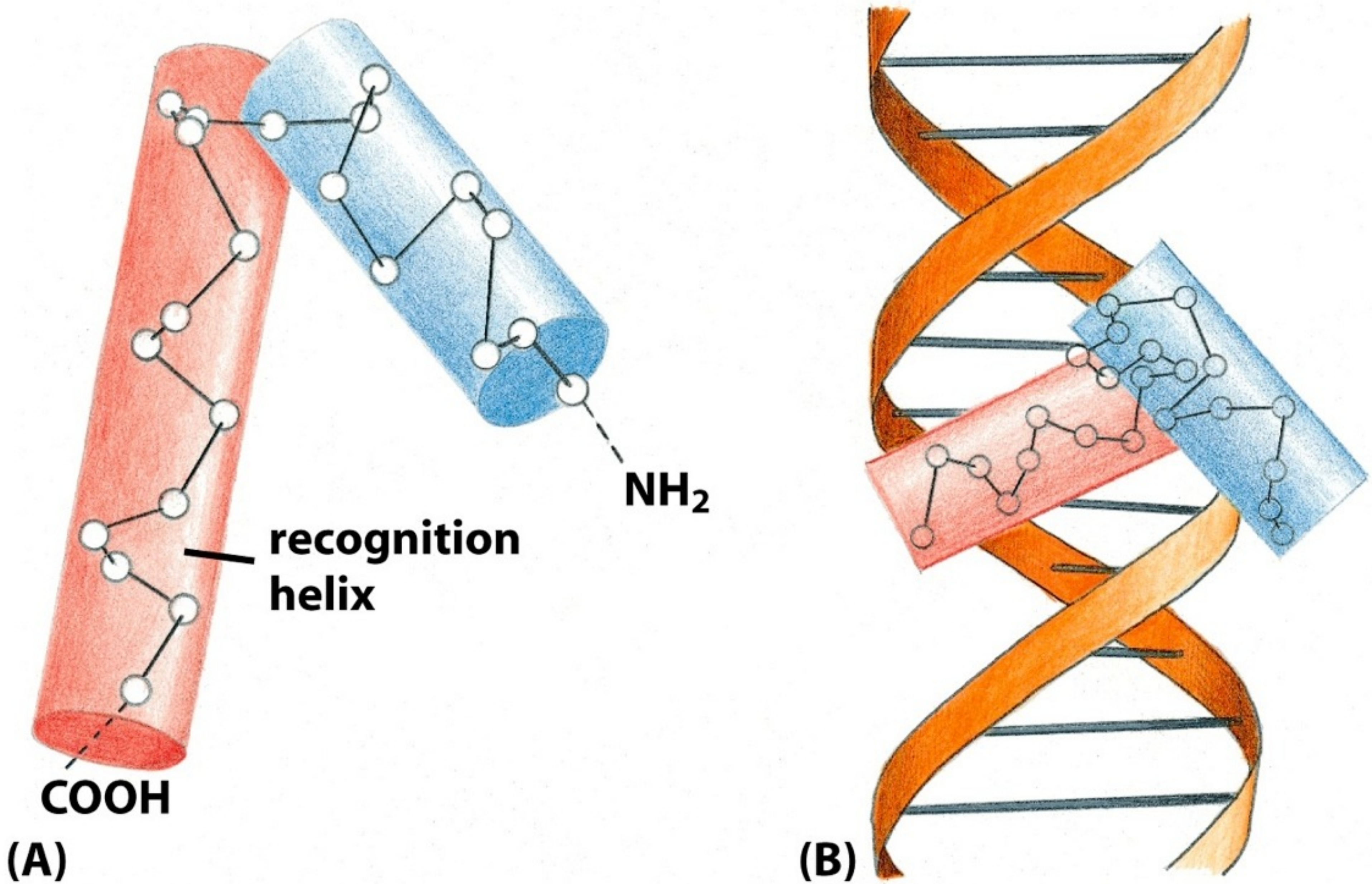


Figure 7-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)

H-T-H Dimers

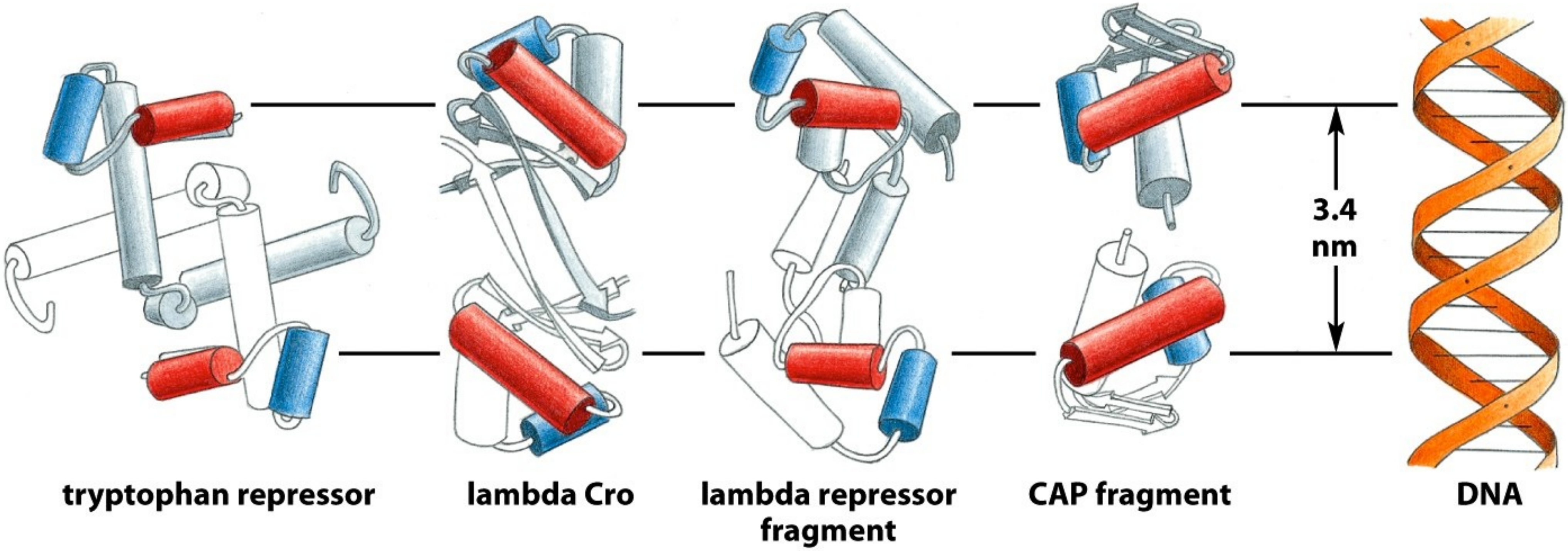
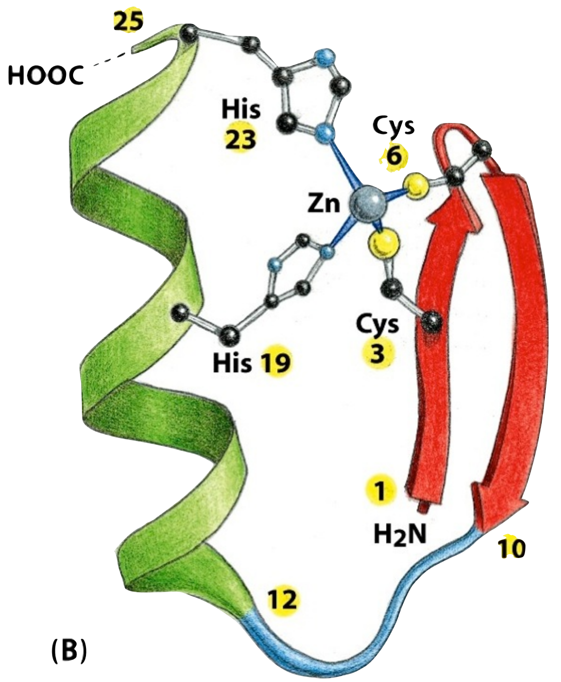


Figure 7-11 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Bind 2 DNA patches, ~ 1 turn apart
Increases both specificity and affinity

Zinc Finger Motif



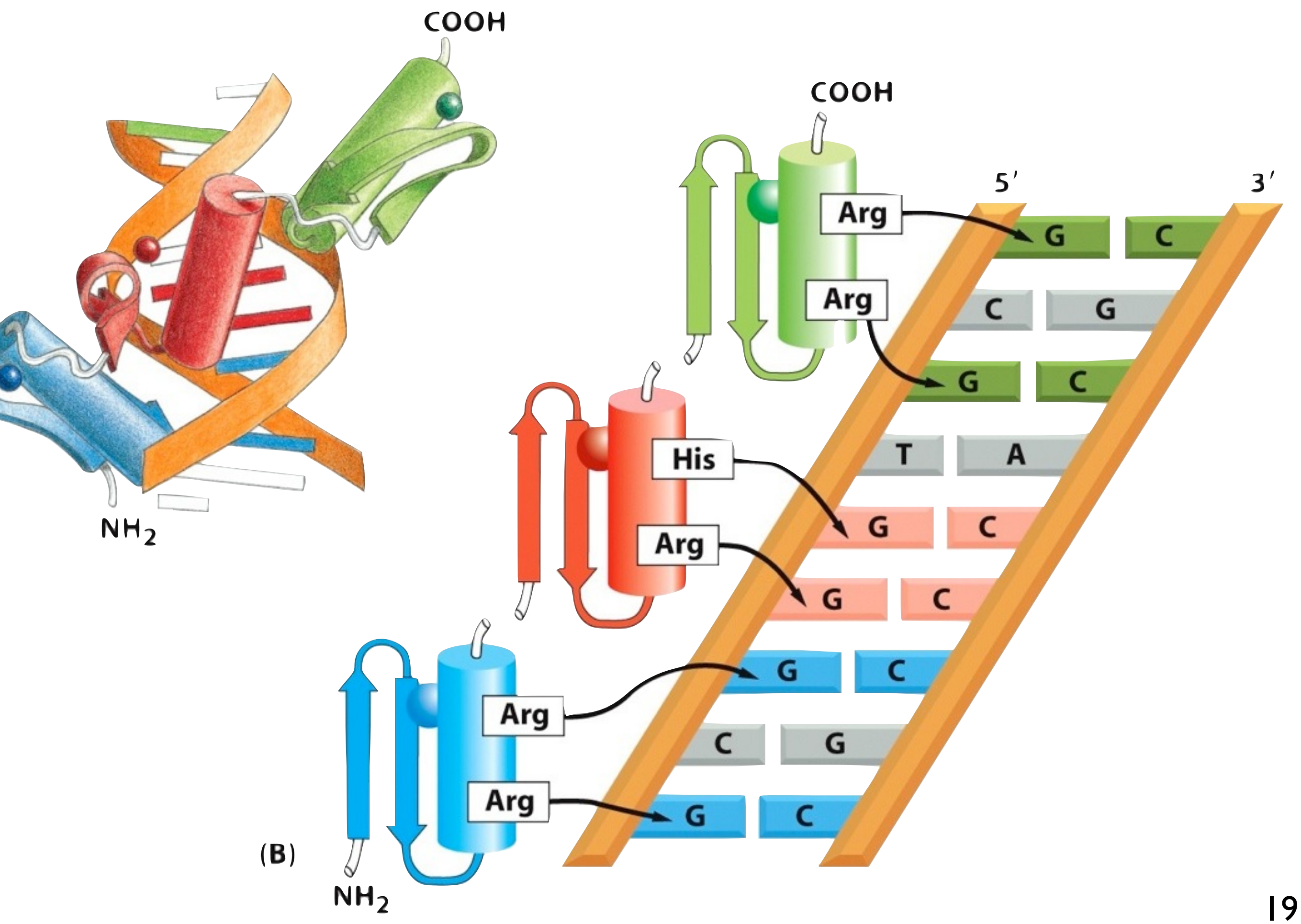


Figure 7-15 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Overheard at the Halloween Party

WWW.PHDCOMICS.COM
© Jorge Cham 10/29/2008

Leucine Zipper Motif

Figure 7-19 Molecular Biology of the Cell 5/e (© Garland Science 2008)

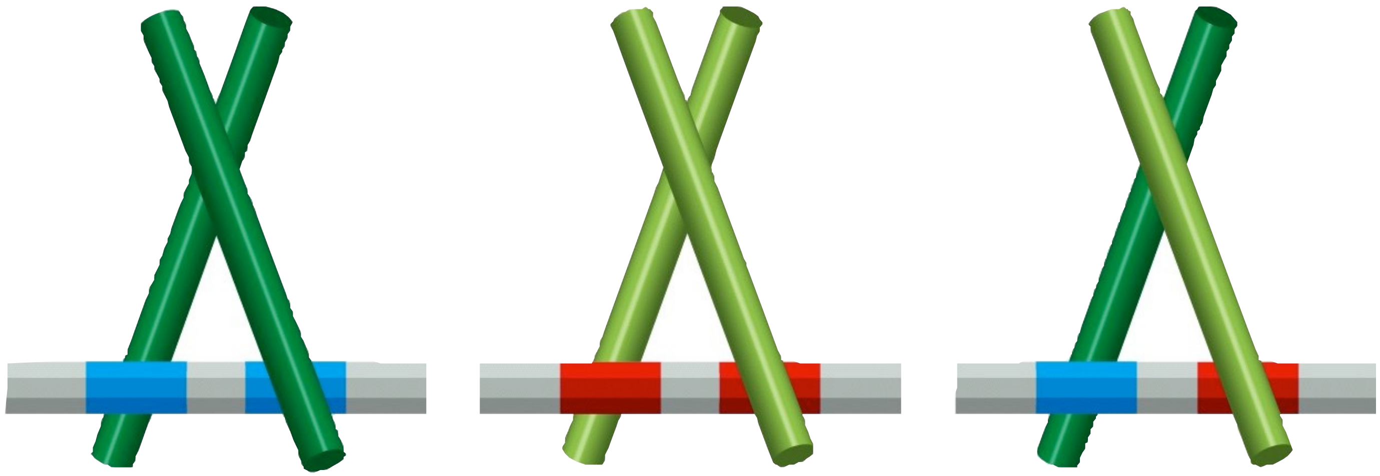


Figure 7-20 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Homo-/hetero-dimers and combinatorial control

MyoD

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

We understand some Protein/DNA interactions

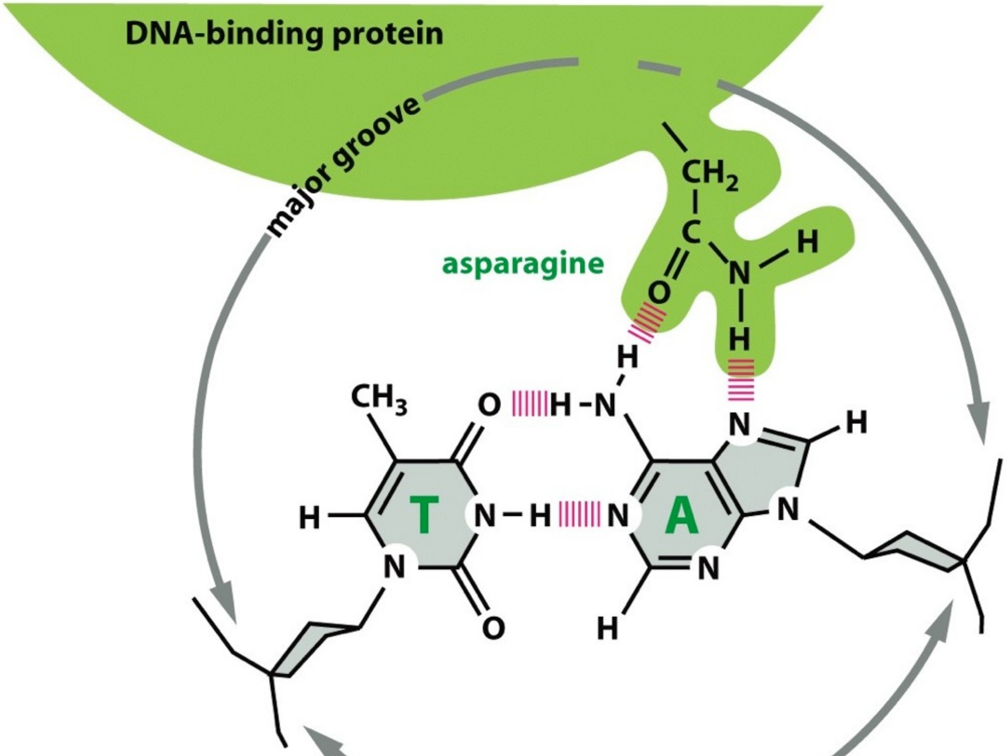


Figure 7-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

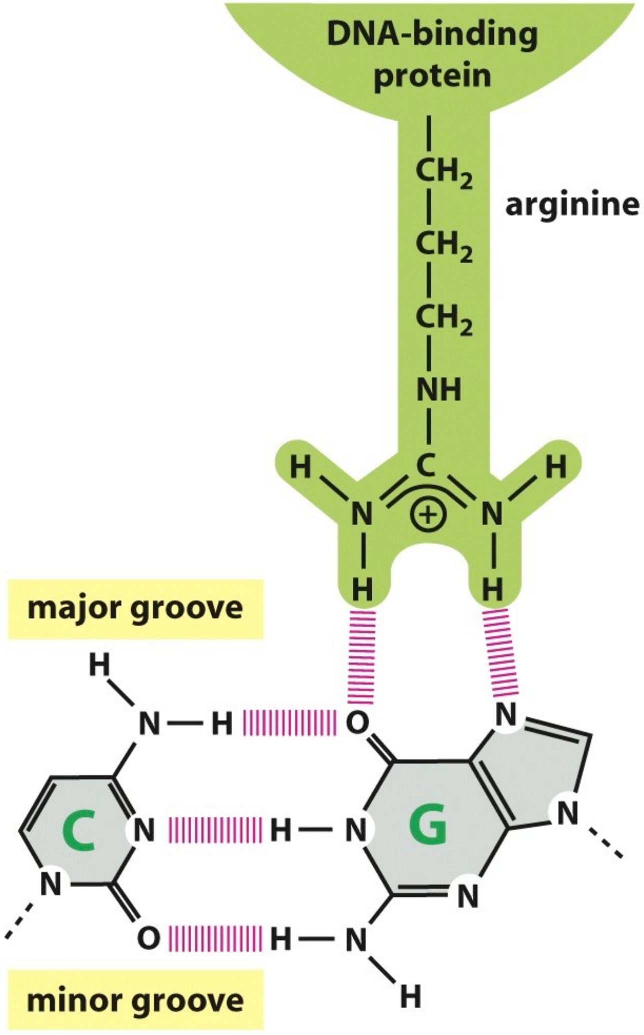
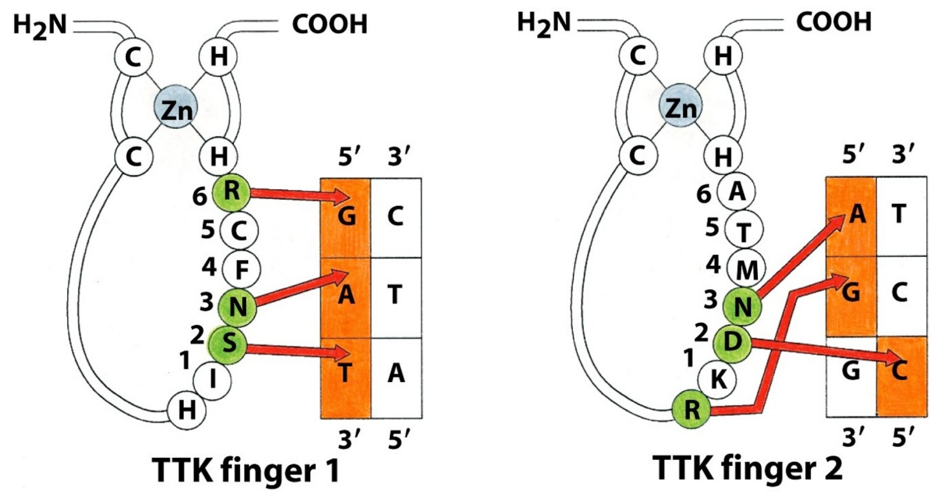


Figure 7-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)

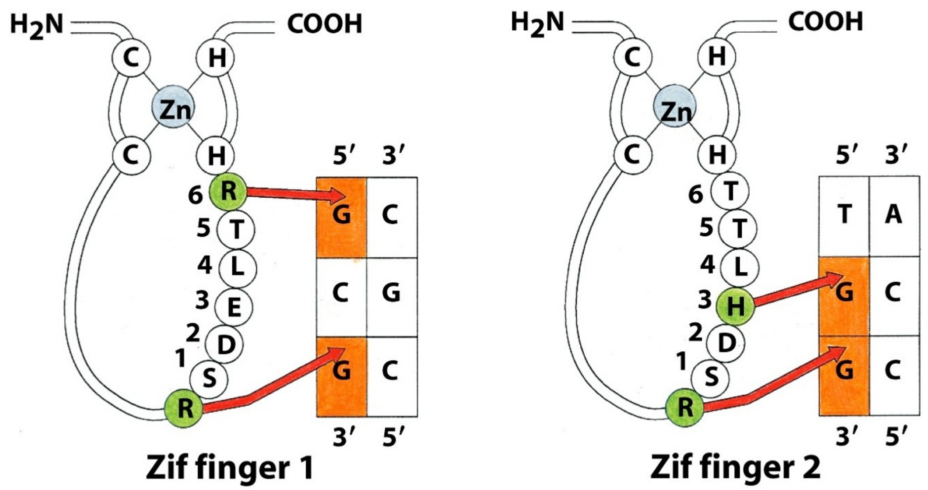
But the overall DNA binding “code” still defies prediction



TTK finger 1

TTK finger 2

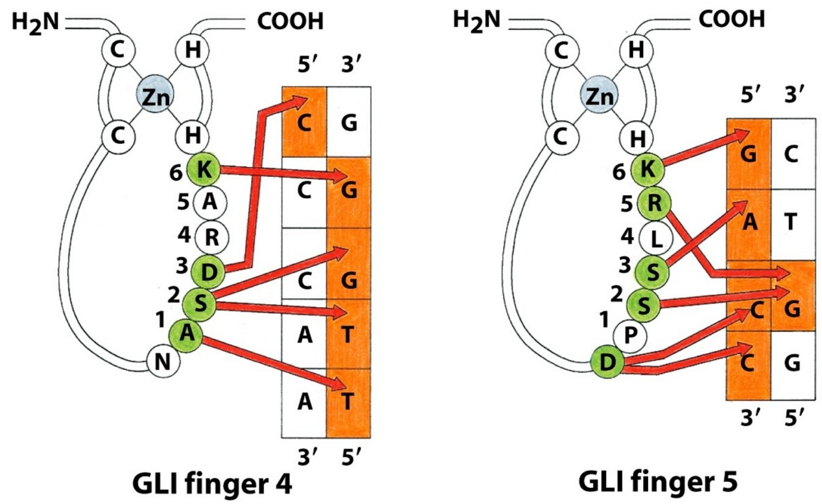
Figure 7-26 part 1 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Zif finger 1

Zif finger 2

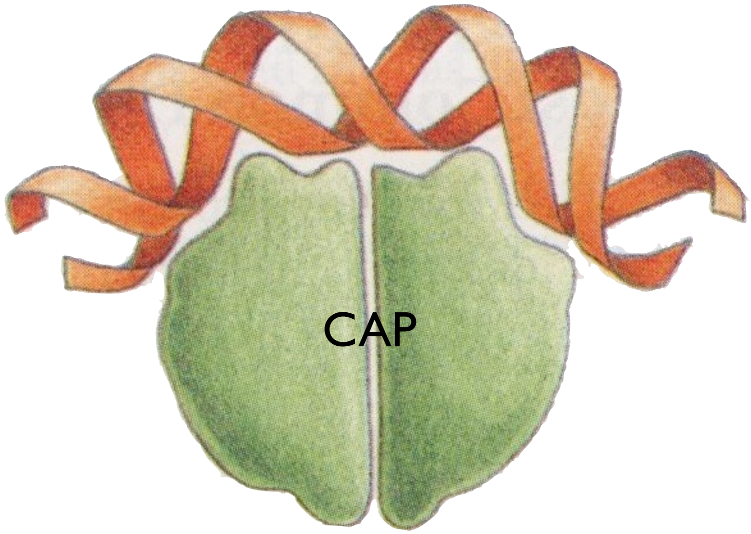
Figure 7-26 part 2 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



GLI finger 4

GLI finger 5

Figure 7-26 part 3 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Summary

Proteins can bind DNA to regulate gene expression (i.e., production of other proteins & themselves)

This is widespread

Complex combinatorial control is possible

Sequence Motifs

Motif: “a recurring salient thematic element”

Last few slides described *structural* motifs in *proteins*

Equally interesting are the *sequence* motifs in *DNA* to which these proteins bind - e.g. , one leucine zipper dimer might bind (with varying affinities) to dozens or hundreds of similar sequences

DNA binding site summary

Complex "code"

Short patches (4-8 bp)

Often near each other (1 turn = 10 bp)

Often reverse-complements (dimer symmetry)

Not perfect matches

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	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96

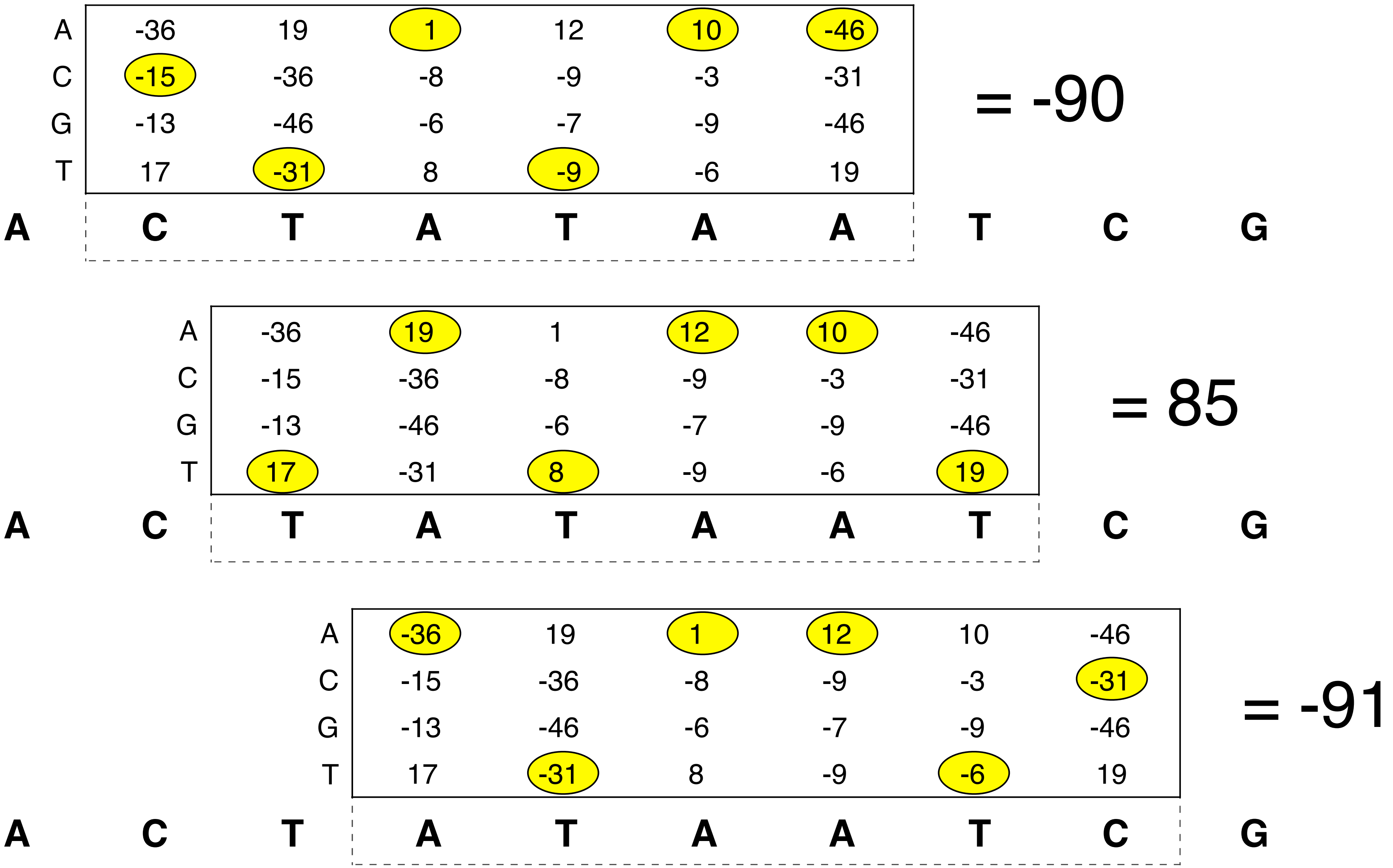
TATA Scores

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

score = 10 $\log_2$ foreground:background odds ratio, 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

= -90

A C T A T A A T C G

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

= 85

A C T A T A A T C G

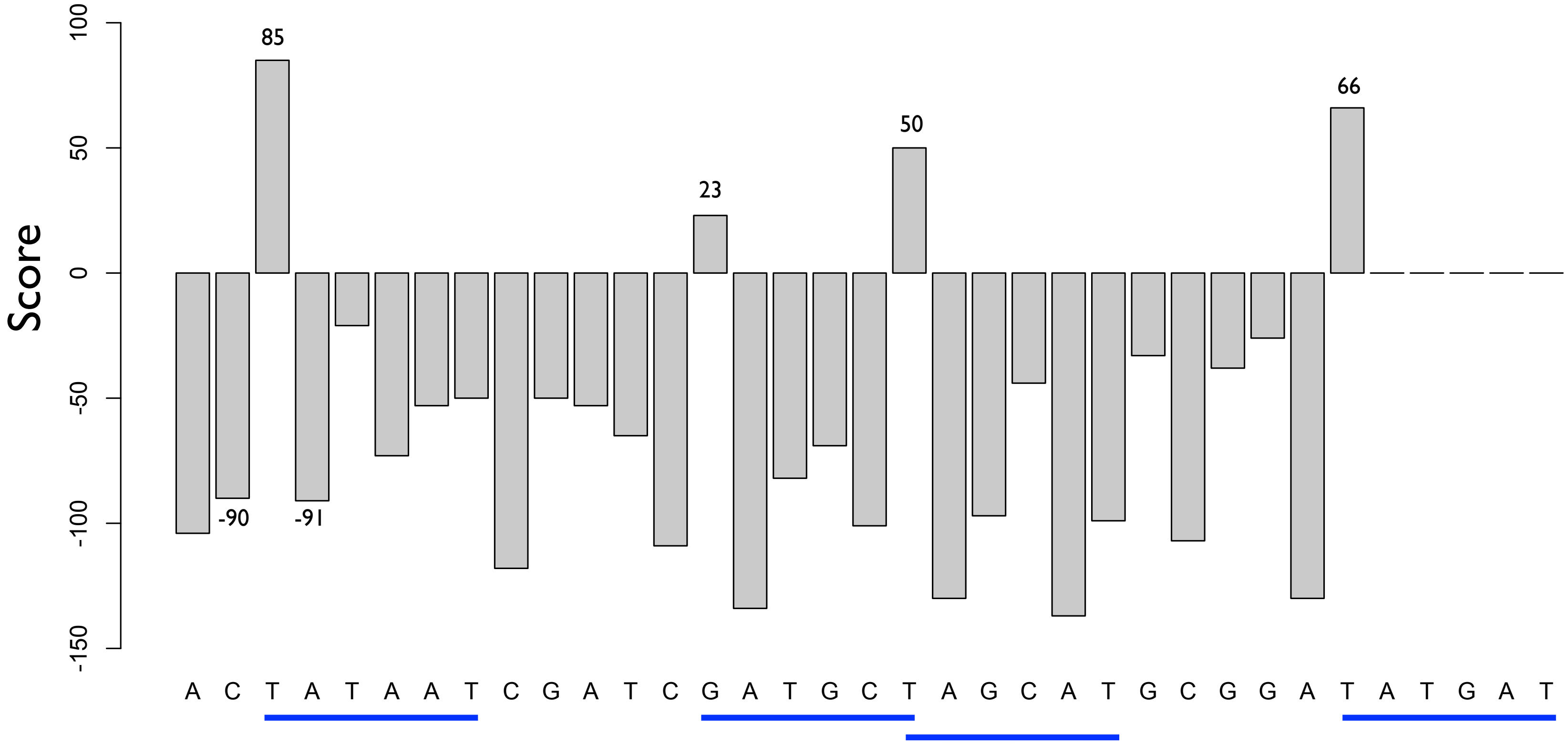
A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

= -91

A C T A T A A T C G

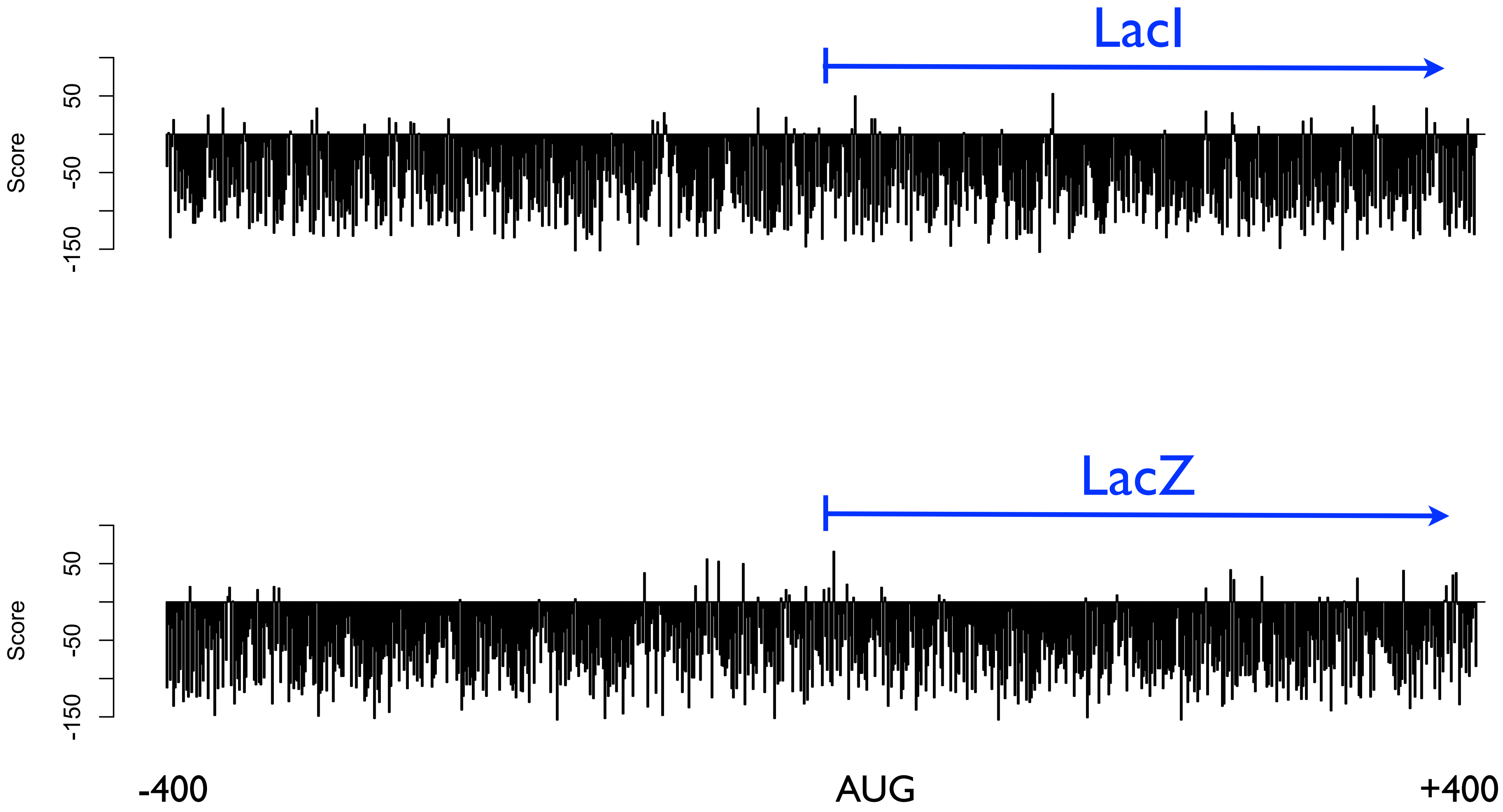
Stormo, Ann. Rev. Biophys. Biophys Chem, 17, 1988, 241-263

Scanning for TATA



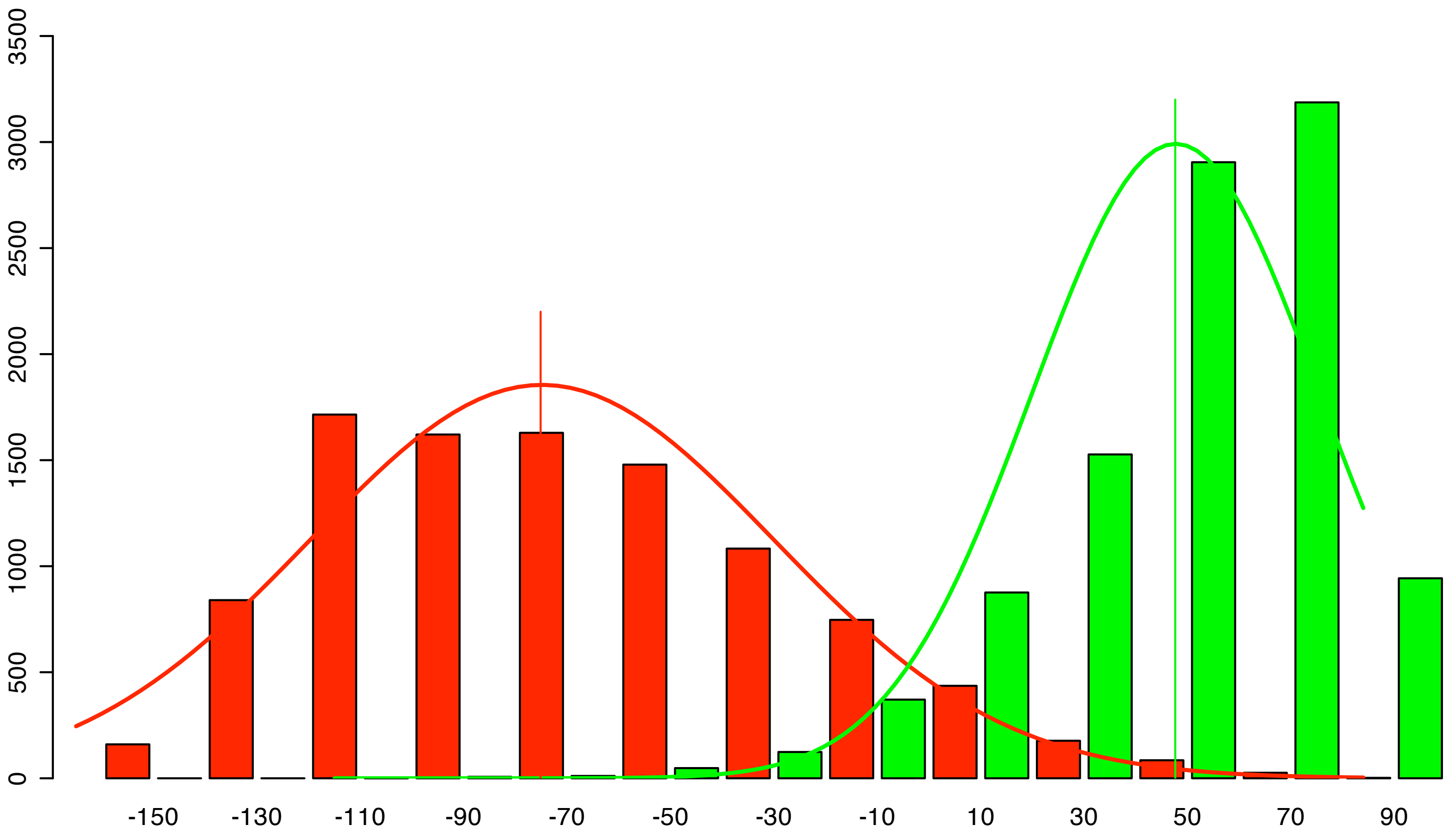
See also slide 60

TATA Scan at 2 genes



Score Distribution

(Simulated)



10^4 random 6-mers from foreground (green) or uniform background (red)

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 $S = B_1 B_2 \ldots B_6$:

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

Assumes *independence*

Neyman-Pearson

Given a sample $x_1, x_2, ..., x_n$, from a distribution $f(...|\Theta)$ with parameter Θ, want to test hypothesis $\Theta = \theta_1$ vs $\Theta = \theta_2$.

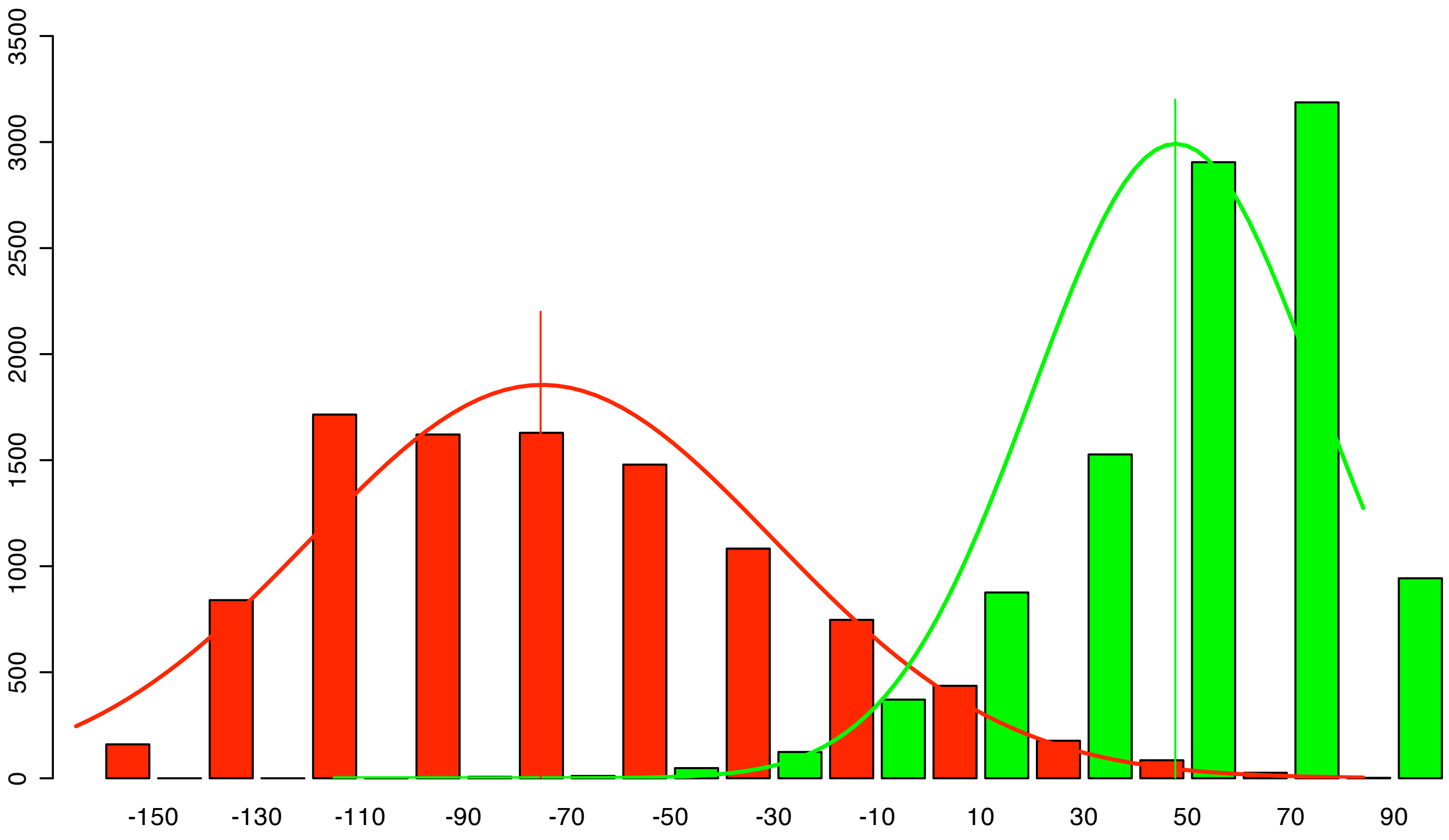
Might as well look at *likelihood ratio:*

$$\frac{f(x_1, x_2, ..., x_n|\theta_1)}{f(x_1, x_2, ..., x_n|\theta_2)} > \tau$$

(or *log likelihood ratio*)

Score Distribution

(Simulated)



10^4 random 6-mers from foreground (green) or uniform background (red)

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 (see HW?)

Weight Matrices: Chemistry

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

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	0.25	0	1
T	0.125	1	0

LLR	Col 1	Col 2	Col 3
A	1.32	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2
T	-1	2	-∞

Log-Likelihood Ratio:

$$\log_2 \frac{f_{x_i,i}}{f_{x_i}}, \;\; f_{x_i} = \frac{1}{4}$$ (uniform background)

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	-∞	3
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).

Relative Entropy

AKA Kullback-Liebler Divergence,
AKA Information Content

Intuitively “distance”, but technically not, since it’s asymmetric

Given distributions P, Q

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

Notes:

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

Undefined if $0 = Q(x) < P(x)$

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

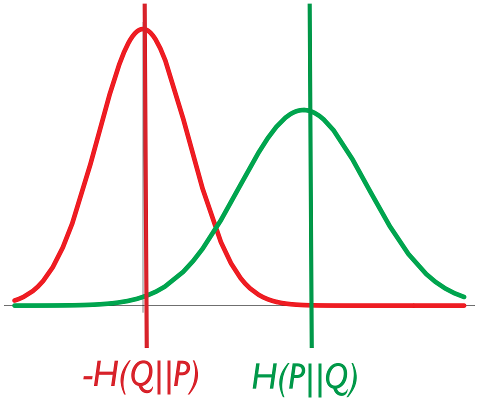
For any fixed length sequence *x*, let

P(x) = Prob. of *x* according to WMM

Q(x) = Prob. of *x* according to background

Relative Entropy:

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

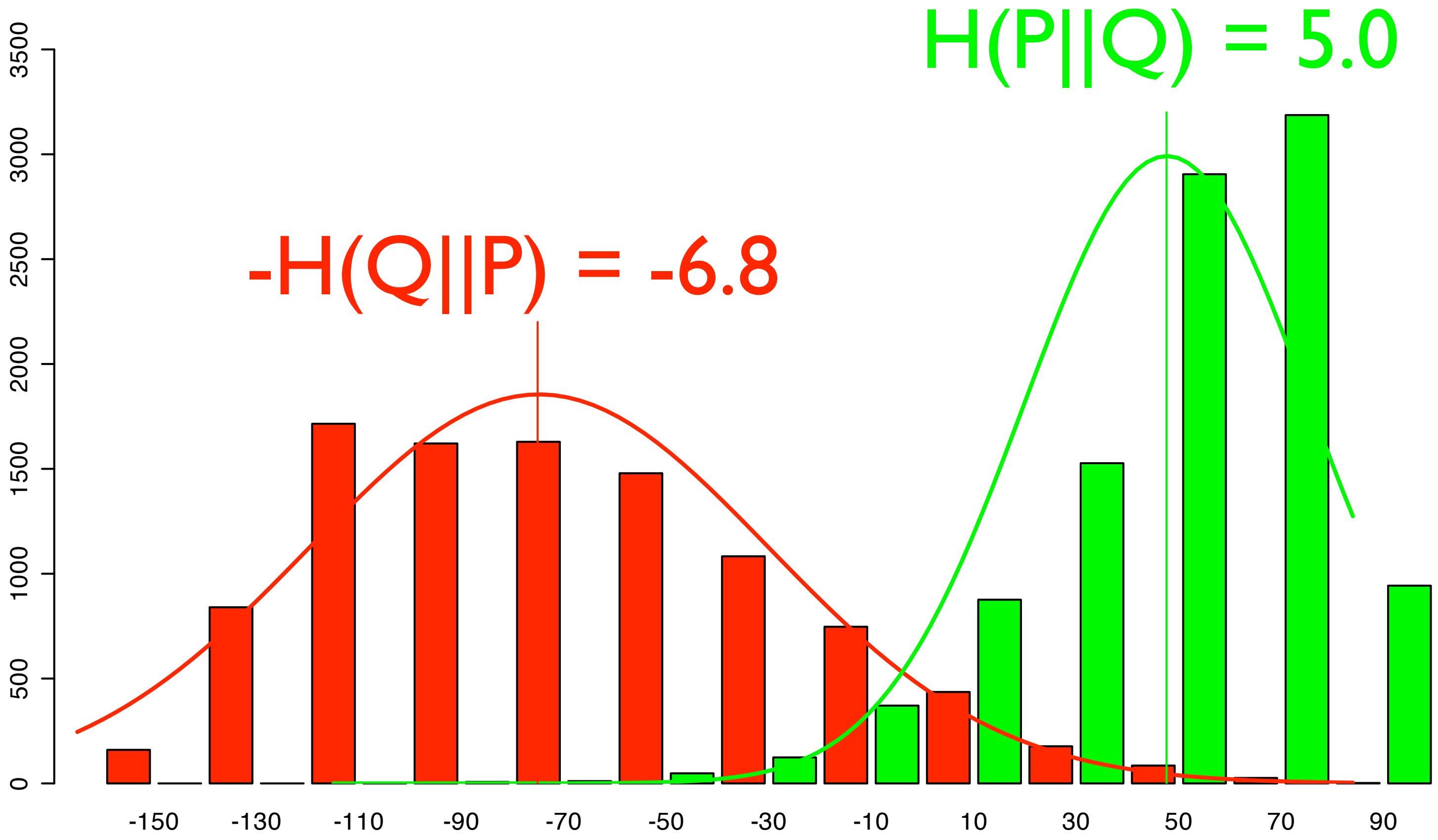


H(P||Q) is *expected log likelihood score* of a sequence randomly chosen from WMM *(wrt background)*;

-H(Q||P) is expected score of *Background* *(wrt WMM)*

Expected score difference: *H(P||Q)* + *H(Q||P)*

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).

For a WMM:

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

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

Proof: exercise

Hint: Use the assumption of independence between WMM columns

WMM Example, cont.

Freq.	Col 1	Col 2	Col 3
A	0.625	0	0
C	0	0	0
G	0.25	0	1
T	0.125	1	0

Uniform

LLR	Col 1	Col 2	Col 3	
A	1.32	-∞	-∞	
C	-∞	-∞	-∞	
G	0	-∞	2	
T	-1	2	-∞	
RelEnt	0.7	2	2	4.7

Non-uniform

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

Pseudocounts

Are the $-\infty$'s 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

WMM Summary

Weight Matrix Model (aka Position Weight Matrix, PWM, Position Specific Scoring Matrix, PSSM, “possum”, 0th order Markov model)

Simple statistical model assuming independence between adjacent positions

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

To scan: add LLRs per position, compare to threshold

Generalizations to higher order models (i.e., letter frequency per position, conditional on neighbor) also possible, with enough training data (k^{th} order MM)

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 ... rest of lecture.

Motif Discovery

Unfortunately, finding a site of max relative entropy in a set of unaligned sequences is NP-hard [Akutsu]

Motif Discovery: 4 example approaches

Brute Force

Greedy search

Expectation Maximization

Gibbs sampler

Brute Force

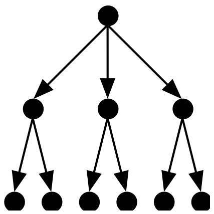
Input:

- Motif length L, plus sequences $s_1, s_2, ..., s_k$ (all of length $n+L-1$, say), each with one instance of an unknown motif

Algorithm:

- Build all k-tuples of length L subsequences, one from each of $s_1, s_2, ..., s_k$ (n^k such tuples)
- Compute relative entropy of each
- Pick best

Brute Force, II



Input:

Motif length L, plus seqs $s_1, s_2, \ldots, s_k$ (all of length n+L-1, say), each with one instance of an unknown motif

Algorithm in more detail:

Build singletons: each len L subseq of each $s_1, s_2, \ldots, s_k$ (nk sets)

Extend to pairs: len L subseqs of each pair of seqs ($n^2\binom{k}{2}$ sets)

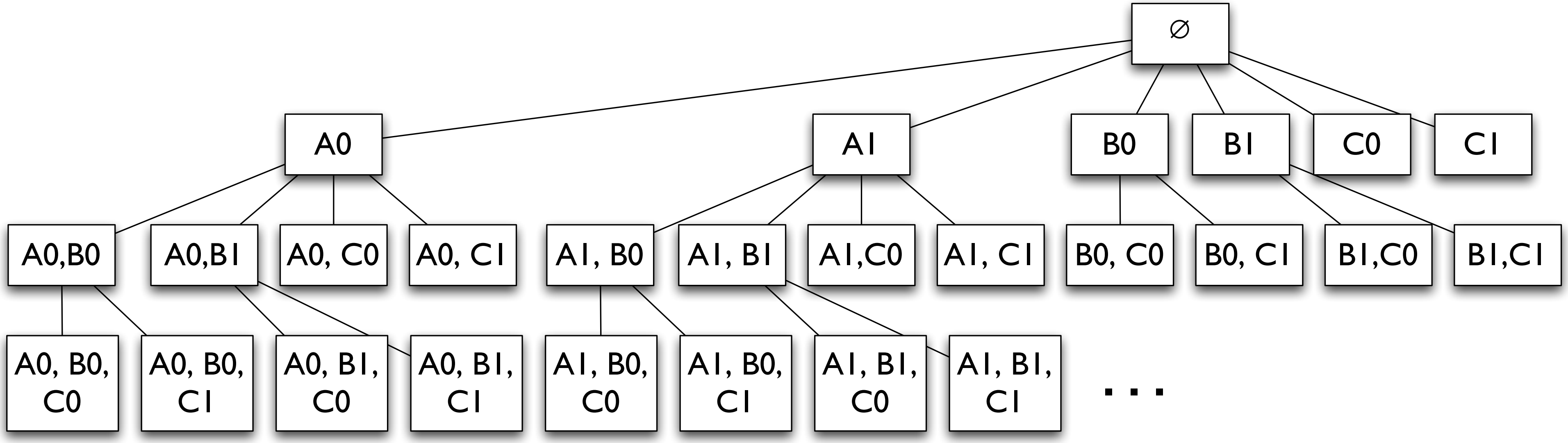
Then triples: len L subseqs of each triple of seqs ($n^3\binom{k}{3}$ sets)

Repeat until all have k sequences ($n^k\binom{k}{k}$ sets)

$(n+1)^k$ in total; compute relative entropy of each; pick best

problem:
astronomically sloooow

Example



∅

A0 A1 B0 B1 C0 C1

A0,B0 A0,B1 A0, C0 A0, C1 A1, B0 A1, B1 A1,C0 A1, C1 B0, C0 B0, C1 B1,C0 B1,C1

A0, B0, C0 A0, B0, C1 A0, B1, C0 A0, B1, C1 A1, B0, C0 A1, B0, C1 A1, B1, C0 A1, B1, C1 . . .

Three sequences (A, B, C), each with two possible motif positions (0,1)

Greedy Best-First

[Hertz, Hartzell & Stormo, 1989, 1990]

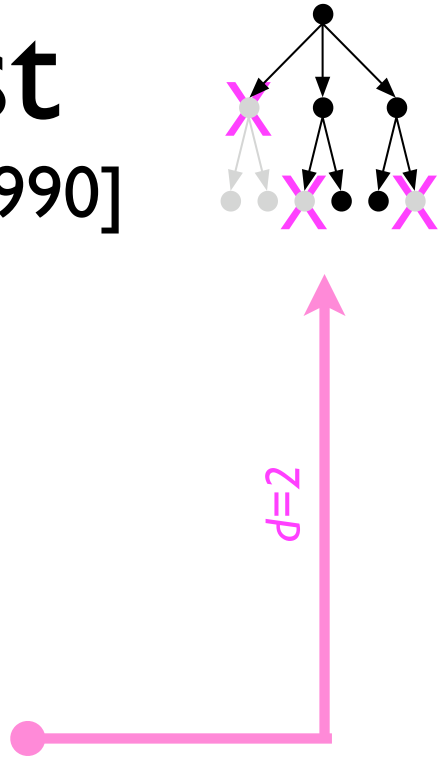
Input:

Sequences $s_1, s_2, \ldots, s_k$; motif length L;

"breadth" d, say d = 1000

Algorithm:

As in brute, but discard all but best d relative entropies at each stage



usual "greedy" problems

Expectation Maximization

[MEME, Bailey & Elkan, 1995]

Input (as above):

Sequences $s_1, s_2, ..., s_k$; motif length l; background model; again assume one instance per sequence (variants possible)

Algorithm: EM

Visible data: the sequences

Hidden data: where's the motif

$$Y_{i,j} = \begin{cases} 1 & \text{if motif in sequence } i \text{ begins at position } j \\ 0 & \text{otherwise} \end{cases}$$

Parameters θ: The WMM

MEME Outline

Typical EM algorithm:

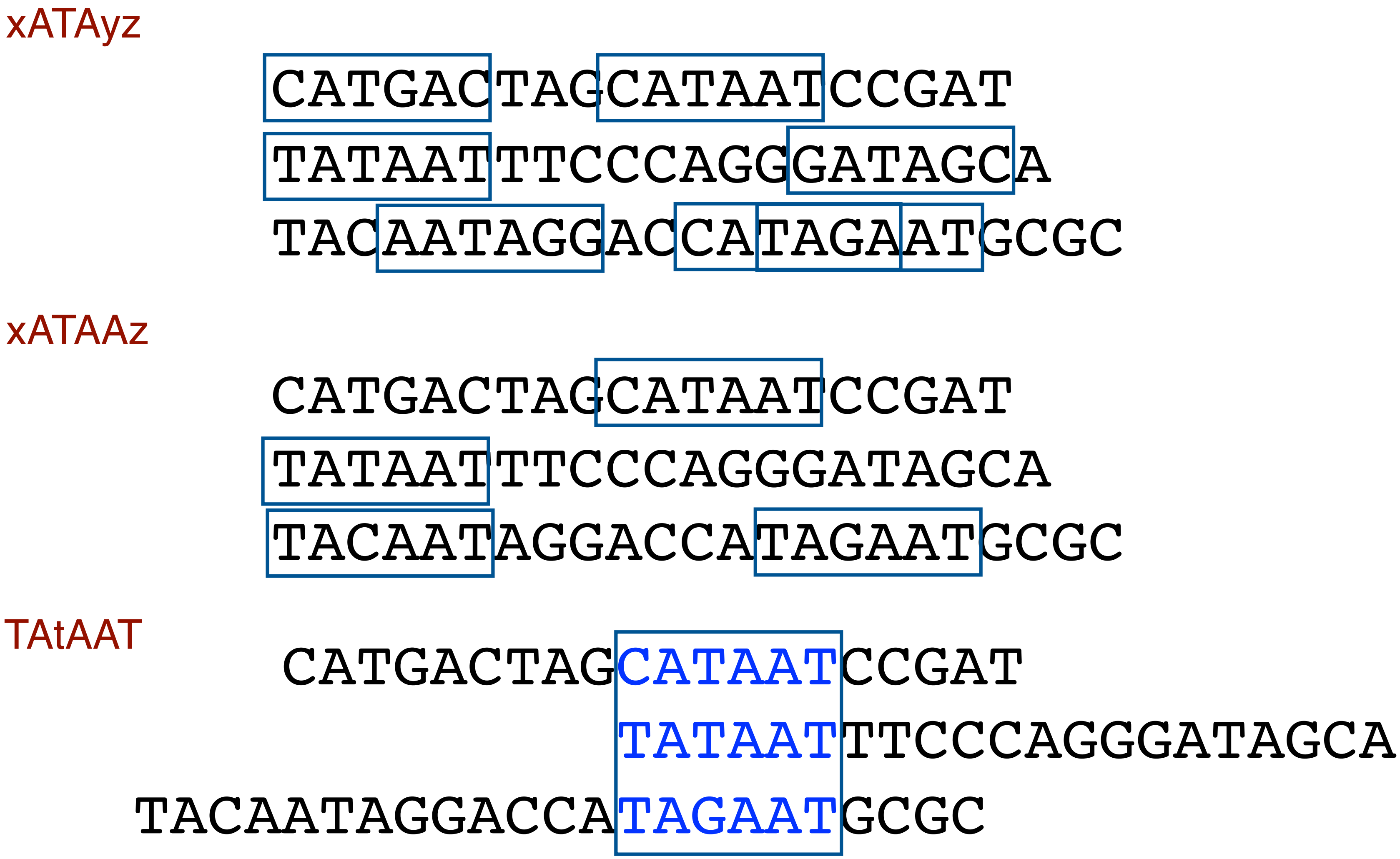
Parameters θ^t at t^{th} iteration, used to estimate where the motif instances are (the hidden variables)

Use those estimates to re-estimate the parameters θ to maximize likelihood of observed data, giving θ^{t+1}

Repeat

Key: given a few good matches to best motif, expect to pick more

Cartoon Example



xATAyz

CATGACTAGCATAATCCGAT
TATAATTTCCCAGGGATAGCA
TACAATAGGACCATAGAATGCGC

xATAAz

CATGACTAGCATAATCCGAT
TATAATTTCCCAGGGATAGCA
TACAATAGGACCATAGAATGCGC

TAtAAT

CATGACTAG**CATAAT**CCGAT
TATAATTTCCCAGGGATAGCA
TACAATAGGACCA**TAGAAT**GCGC

Expectation Step

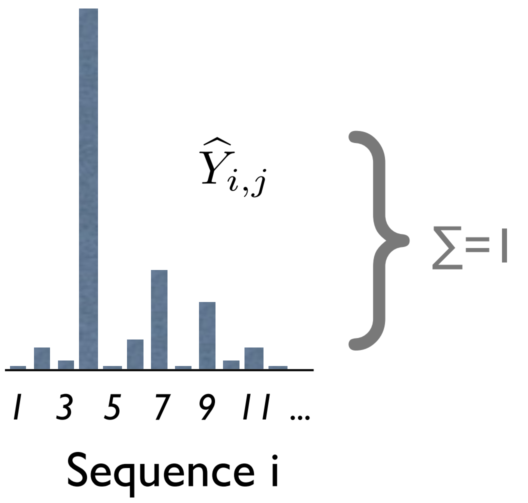
(where are the motif instances?)

$$\begin{aligned}
\widehat{Y}_{i,j} &= E(Y_{i,j} \mid s_i, \theta^t) \\
&= P(Y_{i,j} = 1 \mid s_i, \theta^t) \\
&= P(s_i \mid Y_{i,j} = 1, \theta^t) \frac{P(Y_{i,j}=1|\theta^t)}{P(s_i|\theta^t)} \\
&= cP(s_i \mid Y_{i,j} = 1, \theta^t) \\
&= c' \prod_{k=1}^{l} P(s_{i,j+k-1} \mid \theta^t)
\end{aligned}$$

$E = 0 \cdot P(0) + 1 \cdot P(1)$

Bayes

where c' is chosen so that $\sum_j \widehat{Y}_{i,j} = 1$.



Recall slide 32

Maximization Step

(what is the motif?)

Find θ maximizing expected log likelihood:

$$
\begin{aligned}
Q(\theta \mid \theta^t) &= E_{Y\sim\theta^t}[\log P(s, Y \mid \theta)] \\
&= E_{Y\sim\theta^t}[\log \textstyle\prod_{i=1}^{k} P(s_i, Y_i \mid \theta)] \\
&= E_{Y\sim\theta^t}[\textstyle\sum_{i=1}^{k} \log P(s_i, Y_i \mid \theta)] \\
&= E_{Y\sim\theta^t}[\textstyle\sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} Y_{i,j} \log P(s_i, Y_{i,j} = 1 \mid \theta)] \\
&= E_{Y\sim\theta^t}[\textstyle\sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} Y_{i,j} \log(P(s_i \mid Y_{i,j} = 1, \theta) P(Y_{i,j} = 1 \mid \theta))] \\
&= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} E_{Y\sim\theta^t}[Y_{i,j}] \log P(s_i \mid Y_{i,j} = 1, \theta) + C \\
&= \textstyle\sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} \widehat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C
\end{aligned}
$$

From E-Step

M-Step (cont.)

$$Q(\theta \mid \theta^t) \quad = \quad \textstyle\sum_{i=1}^{k} \sum_{j=1}^{|s_i|-l+1} \widehat{Y}_{i,j} \log P(s_i \mid Y_{i,j} = 1, \theta) + C$$

Exercise: Show this is maximized by "counting" letter frequencies over all possible motif instances, with counts weighted by $\widehat{Y}_{i,j}$, again the "obvious" thing.

s_1 : ACGGATT. . .

$\cdots$

s_k : GC. . . TCGGAC

$\widehat{Y}_{1,1}$	ACGG
$\widehat{Y}_{1,2}$	CGGA
$\widehat{Y}_{1,3}$	GGAT
$\vdots$	$\vdots$
$\widehat{Y}_{k,l-1}$	CGGA
$\widehat{Y}_{k,l}$	GGAC

Initialization

1. Try *every* motif-length substring, and use as initial θ a WMM with, say, 80% of weight on that sequence, rest uniform
2. Run a few iterations of each
3. Run best few to convergence

(Having a supercomputer helps):

http://meme.sdsc.edu/

Another Motif Discovery Approach

The Gibbs Sampler

Lawrence, *et al.* “Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Sequence Alignment,” *Science* 1993

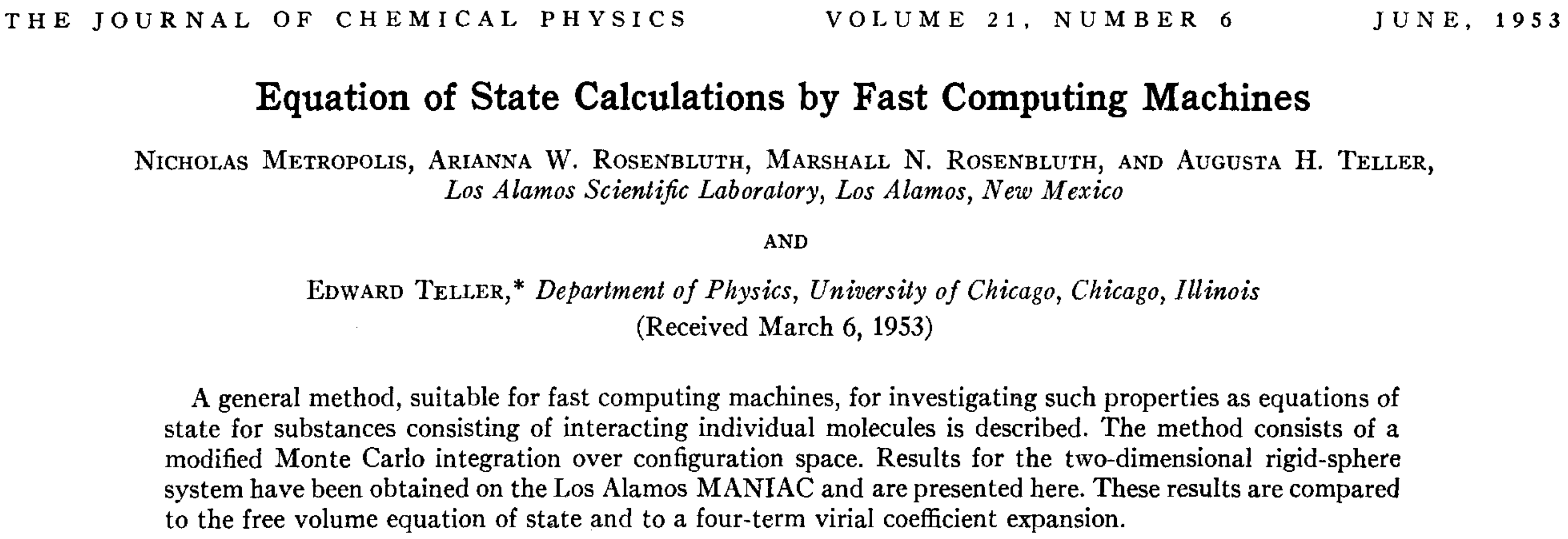
Some History

Geman & Geman, IEEE PAMI 1984

Hastings, Biometrika, 1970

Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, "Equations of State Calculations by Fast Computing Machines," J. Chem. Phys. 1953

Josiah Williard Gibbs, 1839-1903, American physicist, a pioneer of thermodynamics



THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

Markov Chain Monte Carlo

“The Monte Carlo method was a profound innovation, moving away from the continuous variables and differential equations that had dominated the physical sciences since the time of Newton”

p405

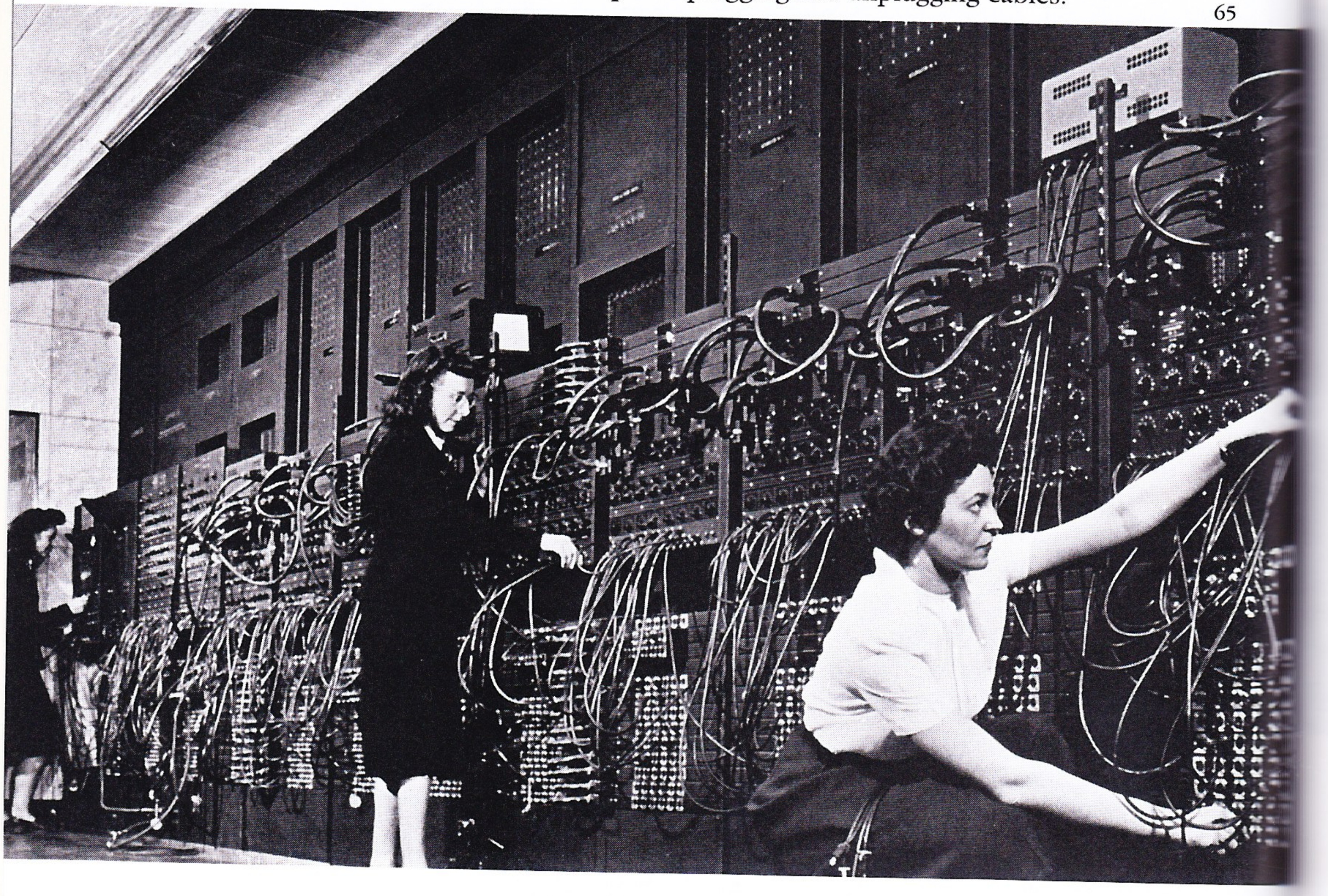
Marshall N. Rosenbluth, 5 February 1927 · 28 September 2003
Author(s): Freeman Dyson
Source: Proceedings of the American Philosophical Society, Vol. 150, No. 2 (Jun., 2006), pp. 403- 407
Published by: American Philosophical Society

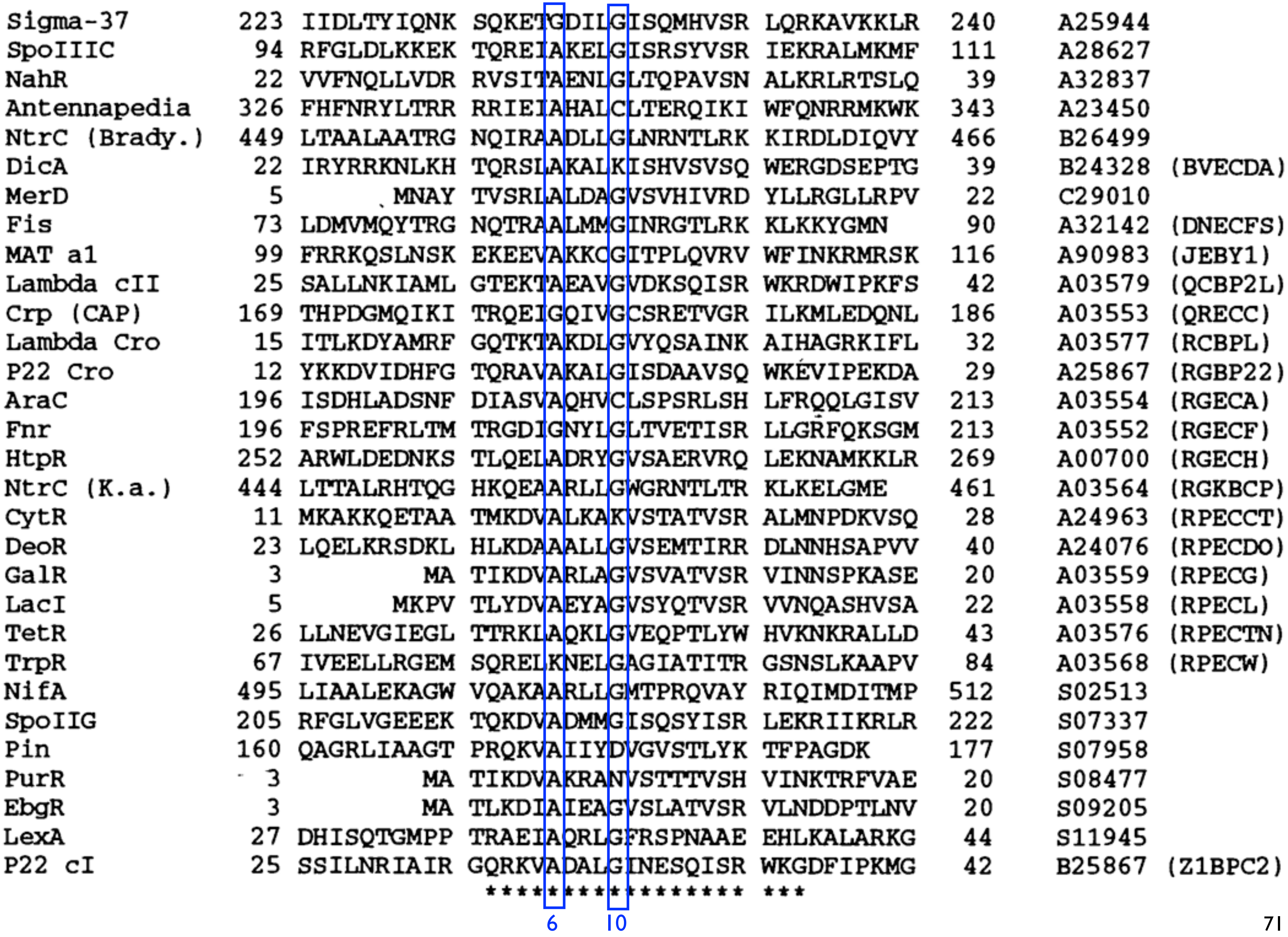
66. Stanislaw Ulam, Richard Feynman and John von Neumann at Los Alamos, late 1940s. Von Neumann invented stored programming; Ulam proved Teller's Super unworkable with hand calculations that the new computers confirmed.

Ulam also pioneered Monte Carlo methods

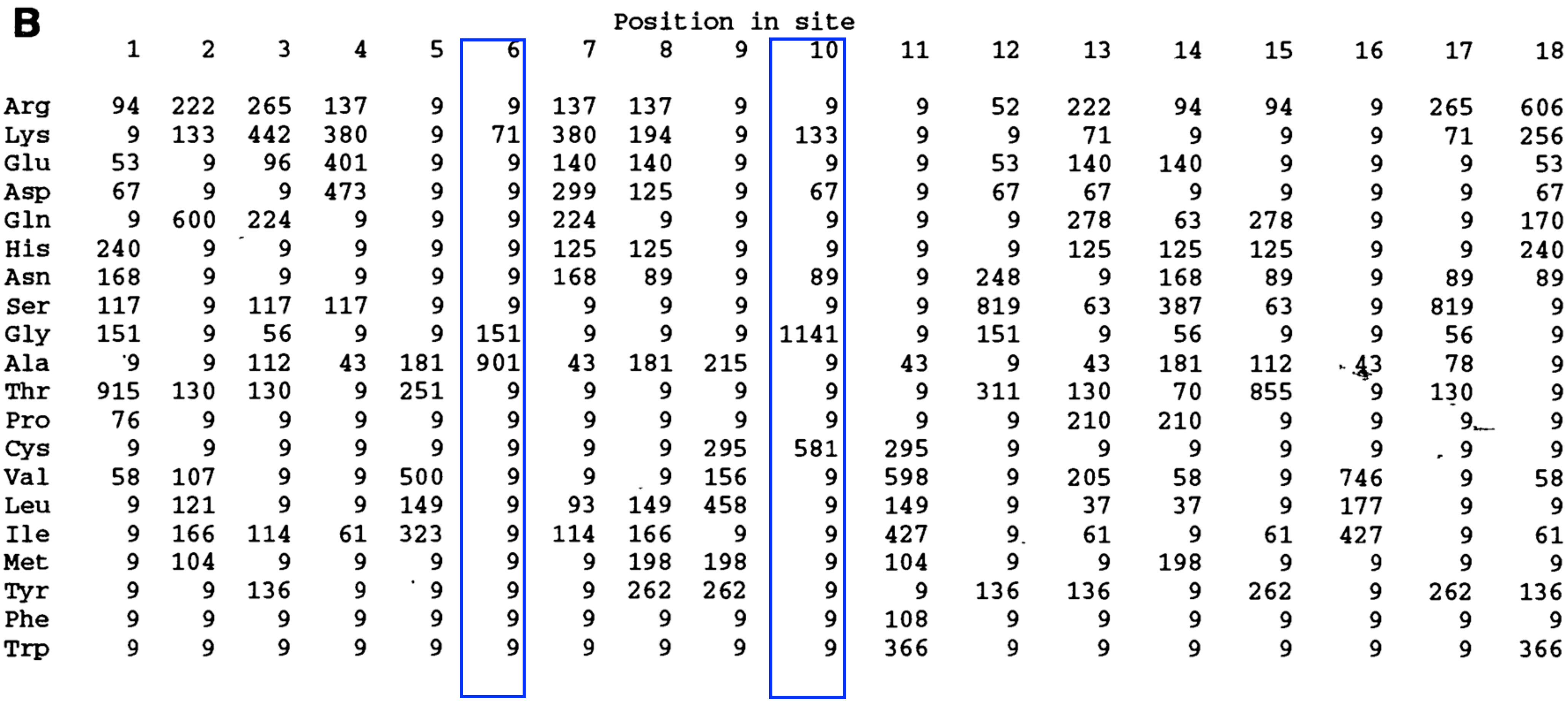
63. Computer power paced H-bomb progress; H-bomb work pushed early computer development. Mathematician Nicholas Metropolis (right, with Paul Stein) developed the MANIAC at Los Alamos. 63

65. Programming the earlier ENIAC required plugging and unplugging cables.





```
Sigma-37        223 IIDLTYIQNK SQKETGDILG ISQMHVSR LQRKAVKKLR 240    A25944
SpoIIIC          94 RFGLDLKKEK TQREIAKELG ISRSYVSR IEKRALMKMF 111    A28627
NahR             22 VVFNQLLVDR RVSITAENLG LTQPAVSN ALKRLRTSLQ  39    A32837
Antennapedia    326 FHFNRYLTRR RRIEIAHALC LTERQIKI WFQNRRMKWK 343    A23450
NtrC (Brady.)   449 LTAALAATRG NQIRAADLLG LNRNTLRK KIRDLDIQVY 466    B26499
DicA             22 IRYRRKNLKH TQRSLAKALK ISHVSVSQ WERGDSEPTG  39    B24328 (BVECDA)
MerD              5       MNAY TVSRLALDAG VSVHIVRD YLLRGLLRPV  22    C29010
Fis              73 LDMVMQYTRG NQTRAALMMG INRGTLRK KLKKYGMN    90    A32142 (DNECFS)
MAT a1           99 FRRKQSLNSK EKEEVAKKCG ITPLQVRV WFINKRMRSK 116    A90983 (JEBY1)
Lambda cII       25 SALLNKIAML GTEKTAEAVG VDKSQISR WKRDWIPKFS  42    A03579 (QCBP2L)
Crp (CAP)       169 THPDGMQIKI TRQEIGQIVG CSRETVGR ILKMLEDQNL 186    A03553 (QRECC)
Lambda Cro       15 ITLKDYAMRF GQTKTAKDLG VYQSAINK AIHAGRKIFL  32    A03577 (RCBPL)
P22 Cro          12 YKKDVIDHFG TQRAVAKALG ISDAAVSQ WKEVIPEKDA  29    A25867 (RGBP22)
AraC            196 ISDHLADSNF DIASVAQHVC LSPSRLSH LFRQQLGISV 213    A03554 (RGECA)
Fnr             196 FSPREFRLTM TRGDIGNYLG LTVETISR LLGRFQKSGM 213    A03552 (RGECF)
HtpR            252 ARWLDEDNKS TLQELADRYG VSAERVRQ LEKNAMKKLR 269    A00700 (RGECH)
NtrC (K.a.)     444 LTTALRHTQG HKQEAARLLG WGRNTLTR KLKELGME   461    A03564 (RGKBCP)
CytR             11 MKAKKQETAA TMKDVALKAK VSTATVSR ALMNPDKVSQ  28    A24963 (RPECCT)
DeoR             23 LQELKRSDKL HLKDAAALLG VSEMTIRR DLNNHSAPVV  40    A24076 (RPECDO)
GalR              3         MA TIKDVARLAG VSVATVSR VINNSPKASE  20    A03559 (RPECG)
LacI              5       MKPV TLYDVAEYAG VSYQTVSR VVNQASHVSA  22    A03558 (RPECL)
TetR             26 LLNEVGIEGL TTRKLAQKLG VEQPTLYW HVKNKRALLD  43    A03576 (RPECTN)
TrpR             67 IVEELLRGEM SQRELKNELG AGIATITR GSNSLKAAPV  84    A03568 (RPECW)
NifA            495 LIAALEKAGW VQAKAARLLG MTPRQVAY RIQIMDITMP 512    S02513
SpoIIG          205 RFGLVGEEEK TQKDVADMMG ISQSYISR LEKRIIKRLR 222    S07337
Pin             160 QAGRLIAAGT PRQKVAIIYD VGVSTLYK TFPAGDK    177    S07958
PurR              3         MA TIKDVAKRAN VSTTTVSH VINKTRFVAE  20    S08477
EbgR              3         MA TLKDIAIEAG VSLATVSR VLNDDPTLNV  20    S09205
LexA             27 DHISQTGMPP TRAEIAQRLG FRSPNAAE EHLKALARKG  44    S11945
P22 cI           25 SSILNRIAIR GQRKVADALG INESQISR WKGDFIPKMG  42    B25867 (Z1BPC2)
                               ***************** ***
```



B

	Position in site																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Arg	94	222	265	137	9	9	137	137	9	9	9	52	222	94	94	9	265	606
Lys	9	133	442	380	9	71	380	194	9	133	9	9	71	9	9	9	71	256
Glu	53	9	96	401	9	9	140	140	9	9	9	53	140	140	9	9	9	53
Asp	67	9	9	473	9	9	299	125	9	67	9	67	67	9	9	9	9	67
Gln	9	600	224	9	9	9	224	9	9	9	9	9	278	63	278	9	9	170
His	240	9	9	9	9	9	125	125	9	9	9	9	125	125	125	9	9	240
Asn	168	9	9	9	9	9	168	89	9	89	9	248	9	168	89	9	89	89
Ser	117	9	117	117	9	9	9	9	9	9	9	819	63	387	63	9	819	9
Gly	151	9	56	9	9	151	9	9	9	1141	9	151	9	56	9	9	56	9
Ala	9	9	112	43	181	901	43	181	215	9	43	9	43	181	112	43	78	9
Thr	915	130	130	9	251	9	9	9	9	9	9	311	130	70	855	9	130	9
Pro	76	9	9	9	9	9	9	9	9	9	9	9	210	210	9	9	9	9
Cys	9	9	9	9	9	9	9	9	295	581	295	9	9	9	9	9	9	9
Val	58	107	9	9	500	9	9	9	156	9	598	9	205	58	9	746	9	58
Leu	9	121	9	9	149	9	93	149	458	9	149	9	37	37	9	177	9	9
Ile	9	166	114	61	323	9	114	166	9	9	427	9	61	9	61	427	9	61
Met	9	104	9	9	9	9	9	198	198	9	104	9	9	198	9	9	9	9
Tyr	9	9	136	9	9	9	9	262	262	9	9	136	136	9	262	9	262	136
Phe	9	9	9	9	9	9	9	9	9	9	108	9	9	9	9	9	9	9
Trp	9	9	9	9	9	9	9	9	9	9	366	9	9	9	9	9	9	366

6

10

How to Average

An old problem:

- n random variables: $x_1, x_2, \ldots, x_k$
- Joint distribution (p.d.f.): $P(x_1, x_2, \ldots, x_k)$
- Some function: $f(x_1, x_2, \ldots, x_k)$
- <u>Want</u> Expected Value: $E(f(x_1, x_2, \ldots, x_k))$

How to Average

$$E(f(x_1, x_2, \ldots, x_k)) \quad = \\ \int_{x_1}\int_{x_2}\cdots\int_{x_k} f(x_1, x_2, \ldots, x_k) \cdot P(x_1, x_2, \ldots, x_k) dx_1 dx_2 \ldots dx_k$$

Approach 1: direct integration
  (rarely solvable analytically, esp. in high dim)
Approach 2: numerical integration
  (often difficult, e.g., unstable, esp. in high dim)
Approach 3: Monte Carlo integration
  sample $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots \vec{x}^{(n)} \sim P(\vec{x})$ and average:

$$E(f(\vec{x})) \approx \frac{1}{n}\sum_{i=1}^{n} f(\vec{x}^{(i)})$$

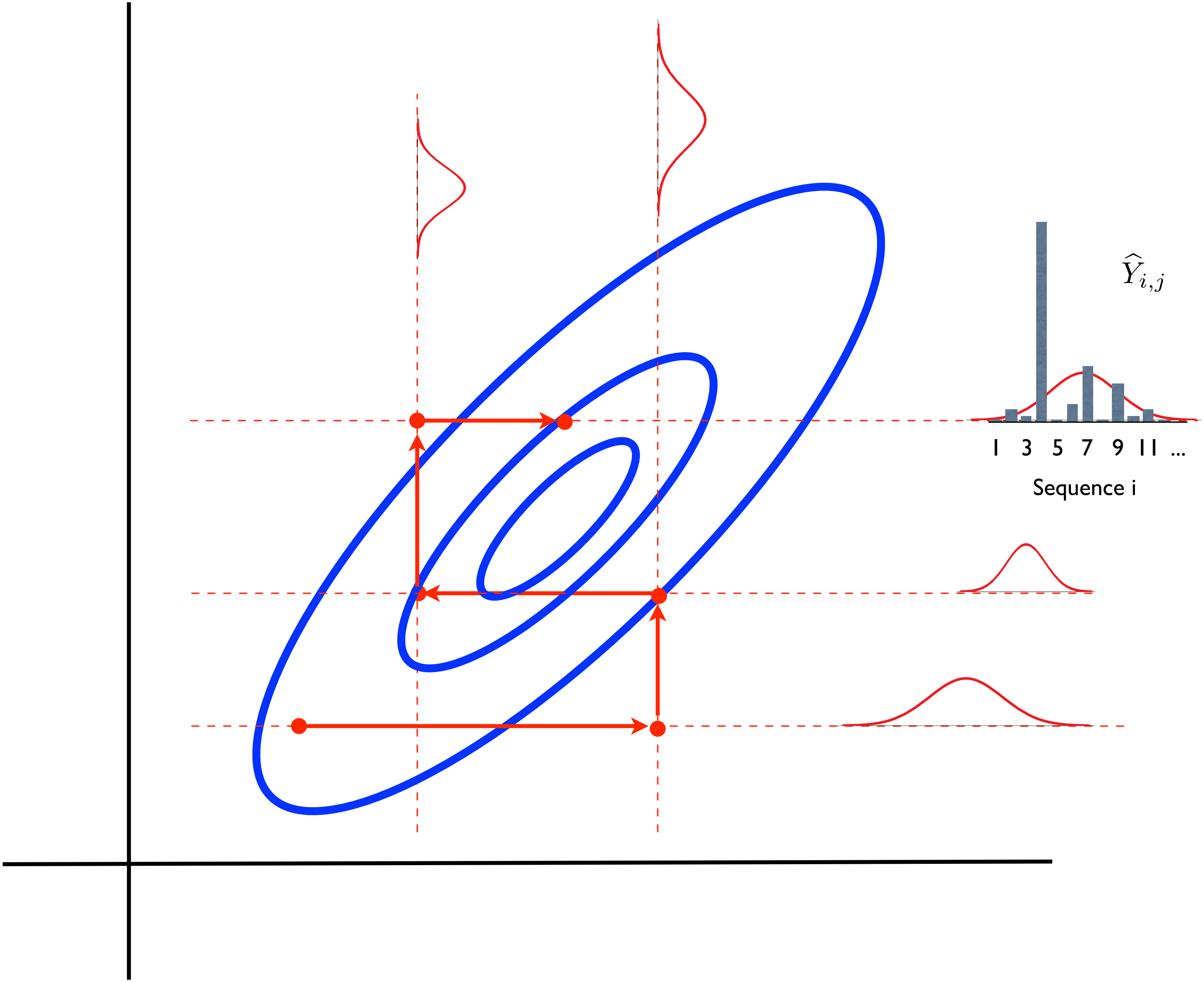
Markov Chain Monte Carlo (MCMC)

- *Independent* sampling also often hard, but *not required* for expectation
- MCMC $\vec{X}_{t+1} \sim P(\vec{X}_{t+1} \mid \vec{X}_t)$ w/ stationary dist = *P*
- Simplest & most common: Gibbs Sampling

$$P(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$

- Algorithm

```
for t = 1 to ∞
    for i = 1 to k do :
        $x_{t+1,i} \sim P(x_{t+1,i} \mid \overbrace{x_{t+1,1}, x_{t+1,2}, \ldots, x_{t+1,i-1}}^{t+1}, \overbrace{x_{t,i+1}, \ldots, x_{t,k}}^{t})$
```


$\widehat{Y}_{i,j}$
1 3 5 7 9 11 ...
Sequence i

Input: again assume sequences $s_1, s_2, ..., s_k$ with one length w motif per sequence

Motif model: WMM

Parameters: Where are the motifs? for $1 \leq i \leq k$, have $1 \leq x_i \leq |s_i|-w+1$

"Full conditional": to calc

$$P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$

build WMM from motifs in all sequences except *i*, then calc prob that motif in i^{th} seq occurs at *j* by usual "scanning" alg.

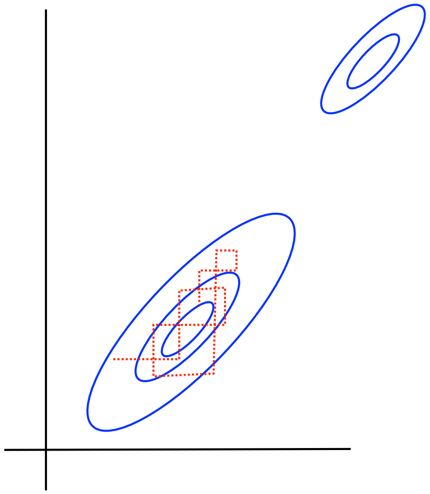
Overall Gibbs Alg

```
Randomly initialize $x_i$'s
for $t$ = 1 to ∞
  for $i$ = 1 to $k$
    discard motif instance from $s_i$;
    recalc WMM from rest
    for $j$ = 1 ... $|s_i|-w+1$
      calculate prob that $i^{th}$ motif is at $j$:
        $P(x_i = j \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$
      pick new $x_i$ according to that distribution
```

Similar to MEME, but it would *average* over, rather than *sample* from

Issues

- Burnin - how long must we run the chain to reach stationarity?
- Mixing - how long a post-burnin sample must we take to get a good sample of the stationary distribution? In particular:
  - Samples are not independent; may not "move" freely through the sample space
  - Many isolated modes

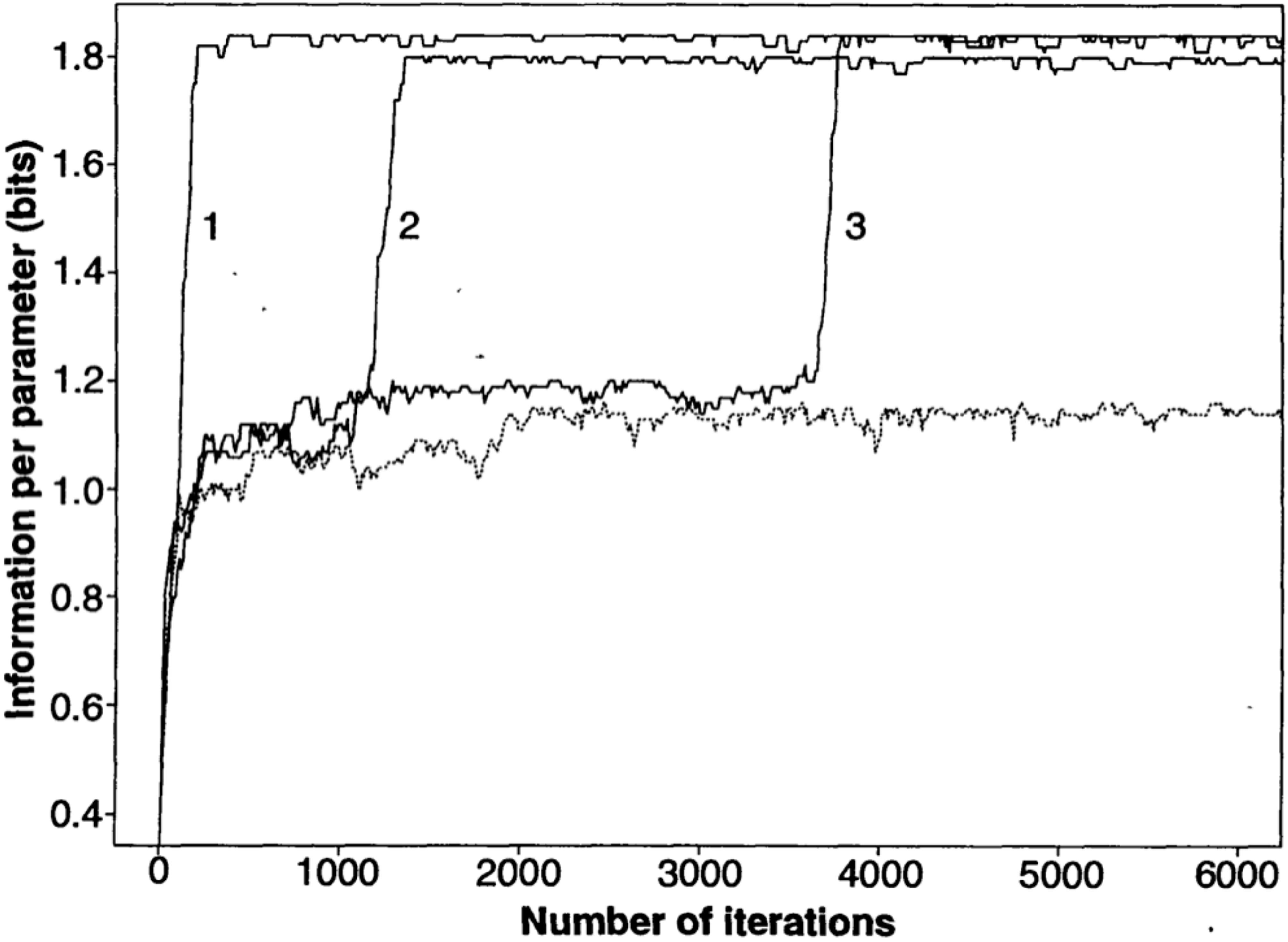


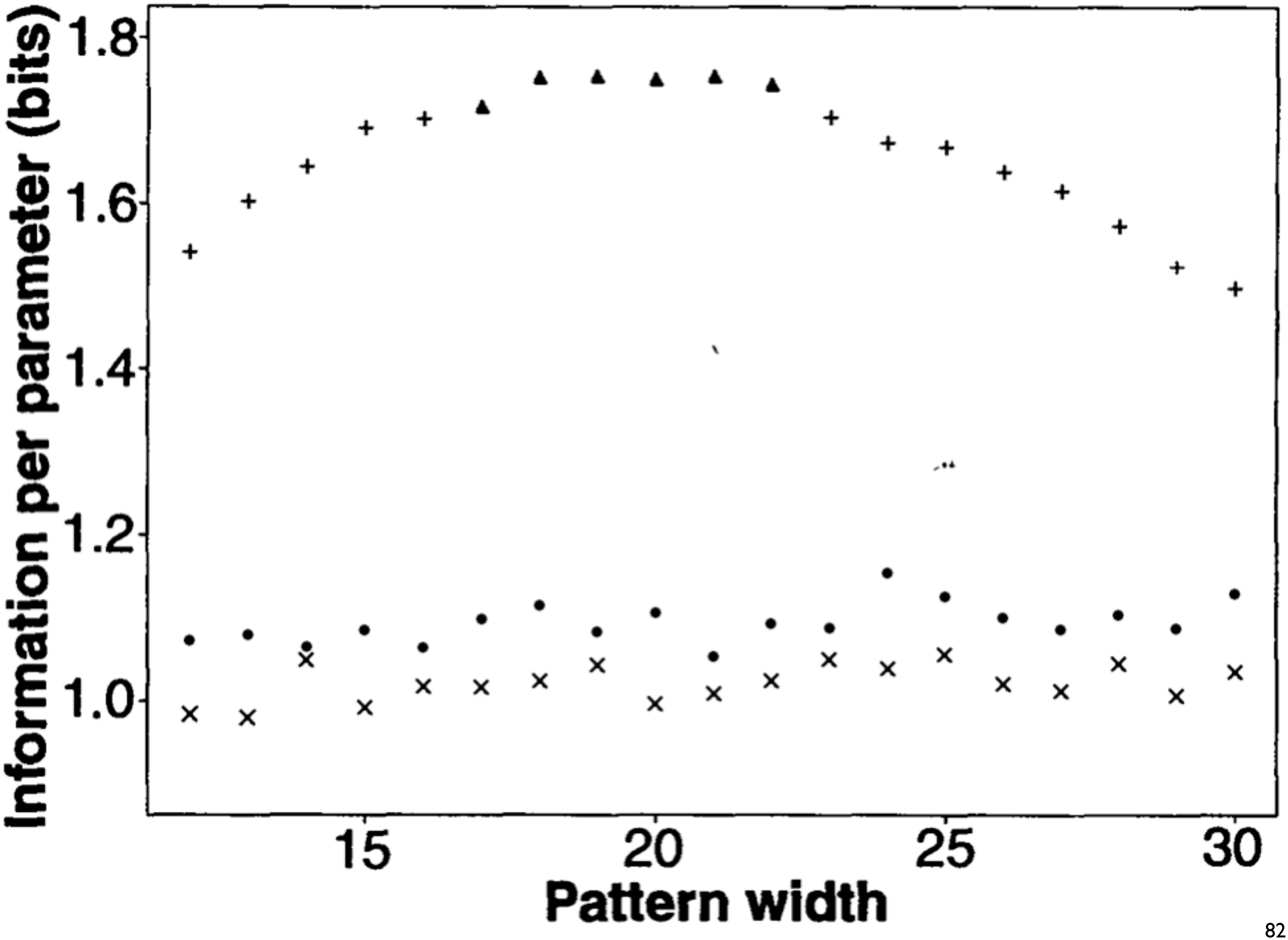
Variants & Extensions

“Phase Shift” - may settle on suboptimal solution that overlaps part of motif. Periodically try moving all motif instances a few spaces left or right.

Algorithmic adjustment of pattern width: Periodically add/remove flanking positions to maximize (roughly) average relative entropy per position

Multiple patterns per string


Information per parameter (bits)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
1
2
3
0
1000
2000
3000
4000
5000
6000
Number of iterations


Information per parameter (bits)
1.8
1.6
1.4
1.2
1.0
15
20
25
30
Pattern width

Bioinformatics Advance Access published November 26, 2013

BIOINFORMATICS **ORIGINAL PAPER**

2013, pages 1–9
doi:10.1093/bioinformatics/btt615

Sequence analysis

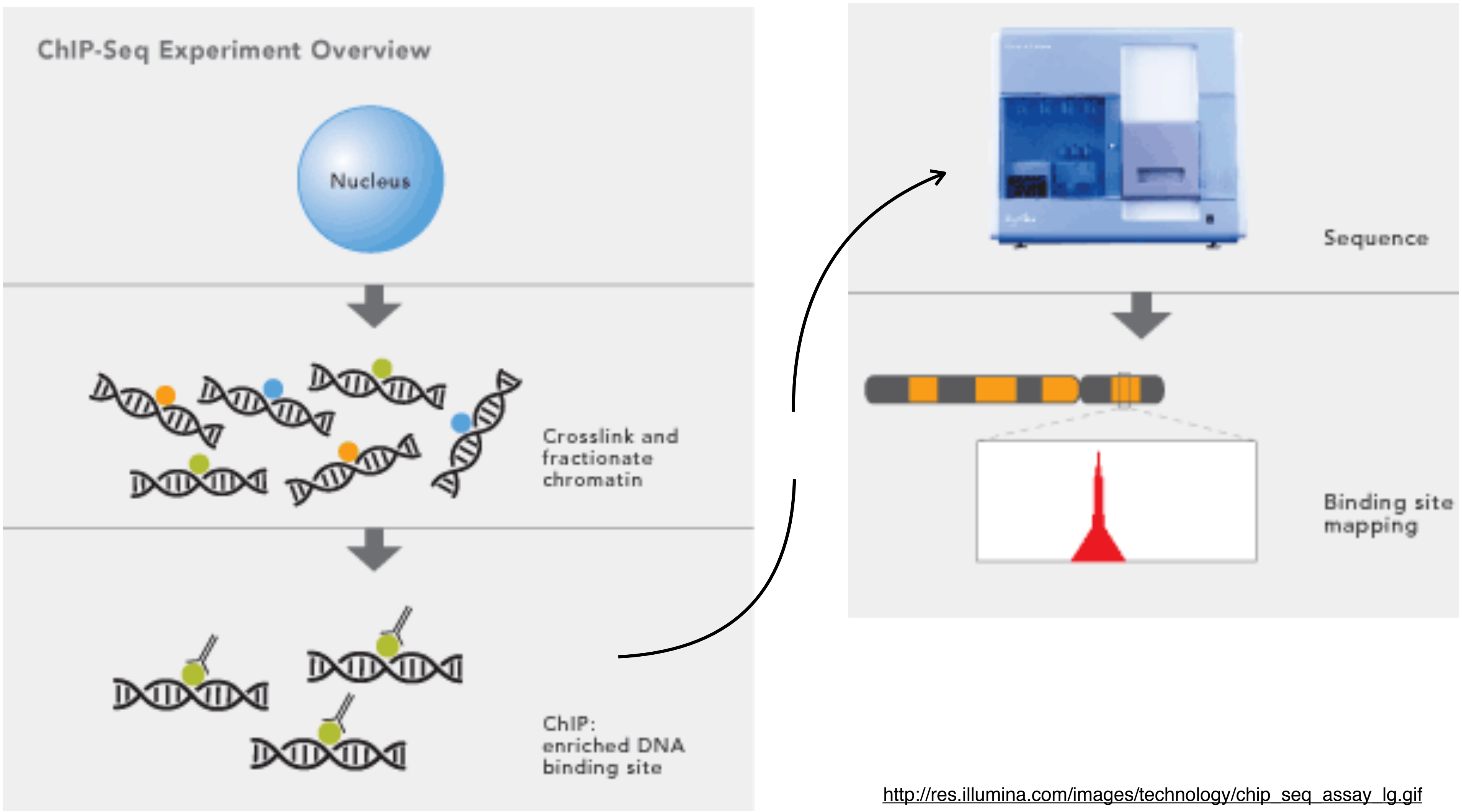
Advance Access publication October 25, 2013

Discriminative motif analysis of high-throughput dataset

Zizhen Yao[1,*], Kyle L. MacQuarrie[1,2], Abraham P. Fong[3,4], Stephen J. Tapscott[1,3,5], Walter L. Ruzzo[6,7,8] and Robert C. Gentleman[9]

[1]Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [2]Molecular and Cellular Biology Program, University of Washington, Seattle, Washington, 98105, USA, [3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [4]Department of Pediatrics, School of Medicine, [5]Department of Neurology, School of Medicine, University of Washington, Seattle, Washington, 98105, USA, [6]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, [7]Department of Computer Science and Engineering, [8]Department of Genome Sciences, University of Washington, Seattle, Washington, 98105, USA and [9]Bioinformatics and Computational Biology, Genentech, South San Francisco, CA 94080, USA

ChIP-seq



http://res.illumina.com/images/technology/chip_seq_assay_lg.gif

TF Binding Site Motifs From ChIPseq

LOTS of data

  E.g. 10^3–10^5 sites, hundreds of reads each (plus perhaps even more nonspecific)

Motif variability

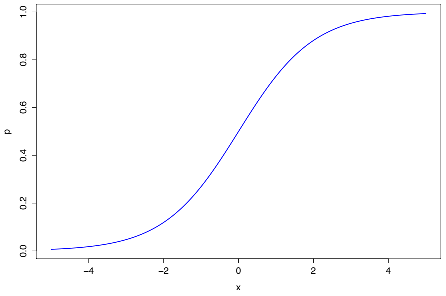
Co-factor binding sites

Method Outline

Discriminative model – foreground/background

Logistic regression – *x* = motif count

$$log \frac{p}{1-p} = \beta_0 + \beta_1 x$$

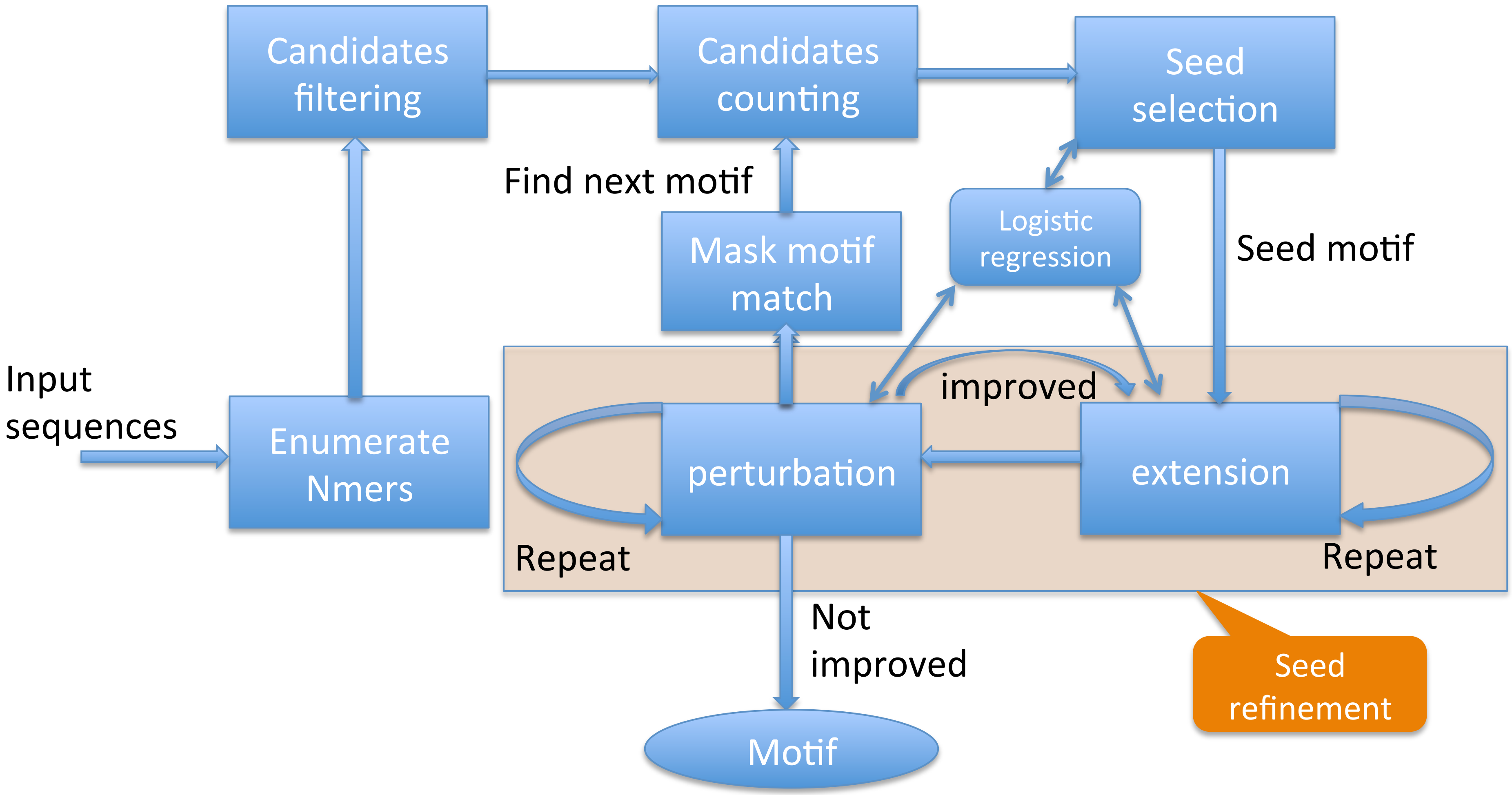


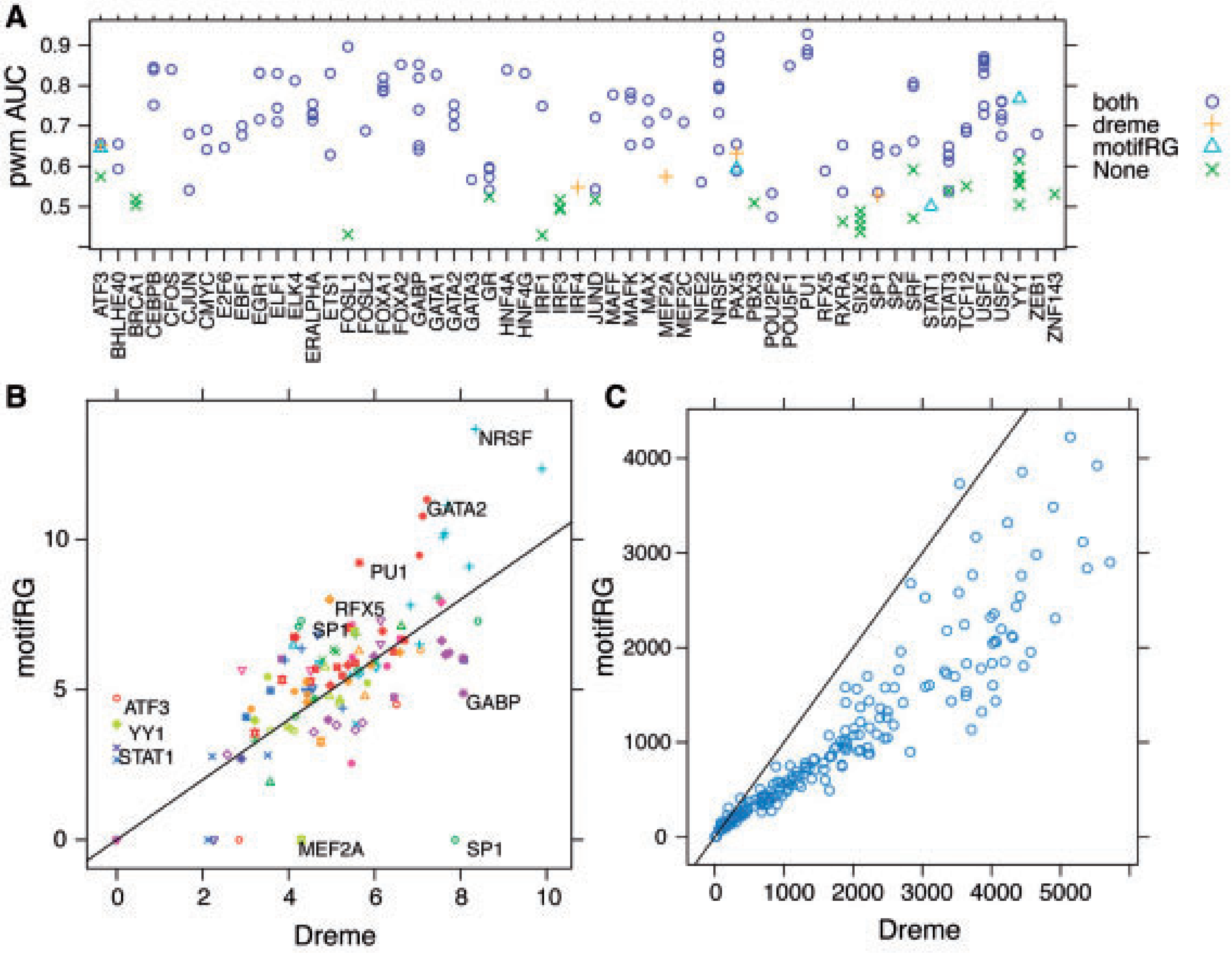
IUPAC patterns – e.g., “R” = A or G

seed/extend/perturb

Z-scores

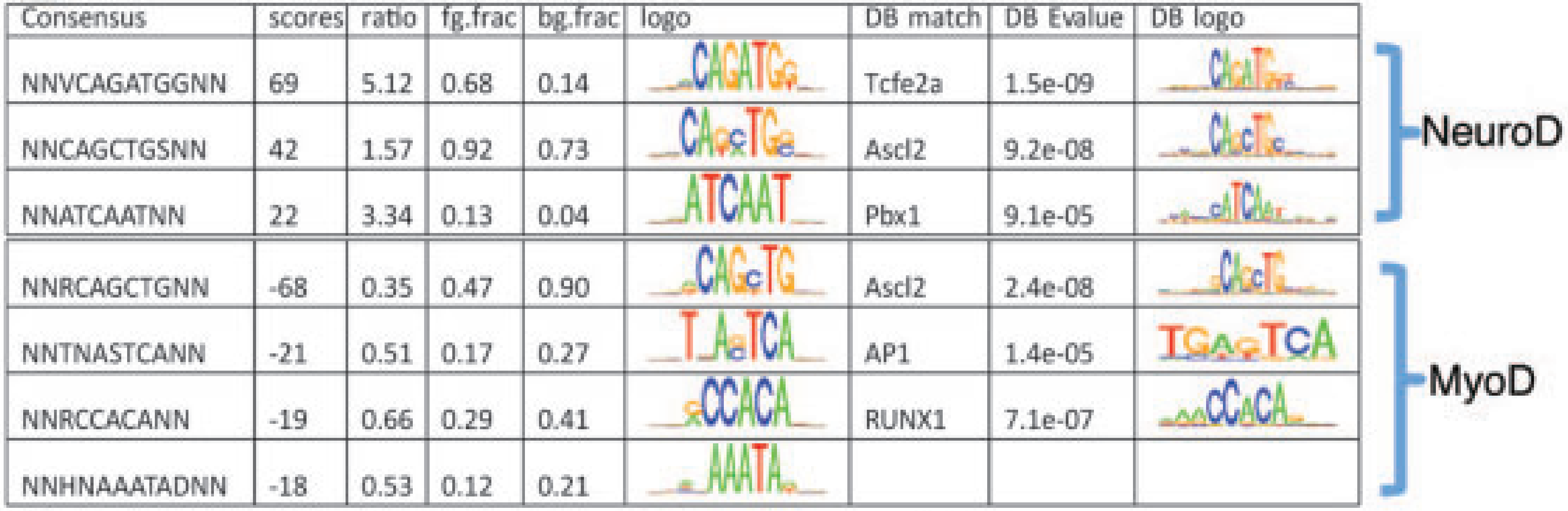
Method Outline




A
pwm AUC
0.9
0.8
0.7
0.6
0.5
both
dreme
motifRG
None
ATF3
BHLHE40
BRCA1
CEBPB
CFOS
CJUN
CMYC
E2F6
EBF1
EGR1
ELF1
ELK4
ERALPHA
ETS1
FOSL1
FOSL2
FOXA1
FOXA2
GABP
GATA1
GATA2
GATA3
GR
HNF4A
HNF4G
IRF1
IRF3
IRF4
JUND
MAFF
MAFK
MAX
MEF2A
MEF2C
NFE2
NRSF
PAX5
PBX3
POU2F2
POU5F1
PU1
RFX5
RXRA
SIX5
SP1
SP2
SRF
STAT1
STAT3
TCF12
USF1
USF2
YY1
ZEB1
ZNF143
B
motifRG
Dreme
10
5
0
0
2
4
6
8
10
NRSF
GATA2
PU1
RFX5
SP1
ATF3
YY1
STAT1
GABP
MEF2A
SP1
C
motifRG
Dreme
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000

A1

A2



Consensus	scores	ratio	fg.frac	bg.frac	logo	DB match	DB Evalue	DB logo
NNVCAGATGGNN	69	5.12	0.68	0.14		Tcfe2a	1.5e-09	
NNCAGCTGSNN	42	1.57	0.92	0.73		Ascl2	9.2e-08	
NNATCAATNN	22	3.34	0.13	0.04		Pbx1	9.1e-05	
NNRCAGCTGNN	-68	0.35	0.47	0.90		Ascl2	2.4e-08	
NNTNASTCANN	-21	0.51	0.17	0.27		AP1	1.4e-05	
NNRCCACANN	-19	0.66	0.29	0.41		RUNX1	7.1e-07	
NNHNAAATADNN	-18	0.53	0.12	0.21				

NATURE BIOTECHNOLOGY VOLUME 23 NUMBER 1 JANUARY 2005

Assessing computational tools for the discovery of transcription factor binding sites

Martin Tompa[1,2], Nan Li[1], Timothy L Bailey[3], George M Church[4], Bart De Moor[5], Eleazar Eskin[6], Alexander V Favorov[7,8], Martin C Frith[9], Yutao Fu[9], W James Kent[10], Vsevolod J Makeev[7,8], Andrei A Mironov[7,11], William Stafford Noble[1,2], Giulio Pavesi[12], Graziano Pesole[13], Mireille Régnier[14], Nicolas Simonis[15], Saurabh Sinha[16], Gert Thijs[5], Jacques van Helden[15], Mathias Vandenbogaert[14], Zhiping Weng[9], Christopher Workman[17], Chun Ye[18] & Zhou Zhu[4]

Methodology

13 tools

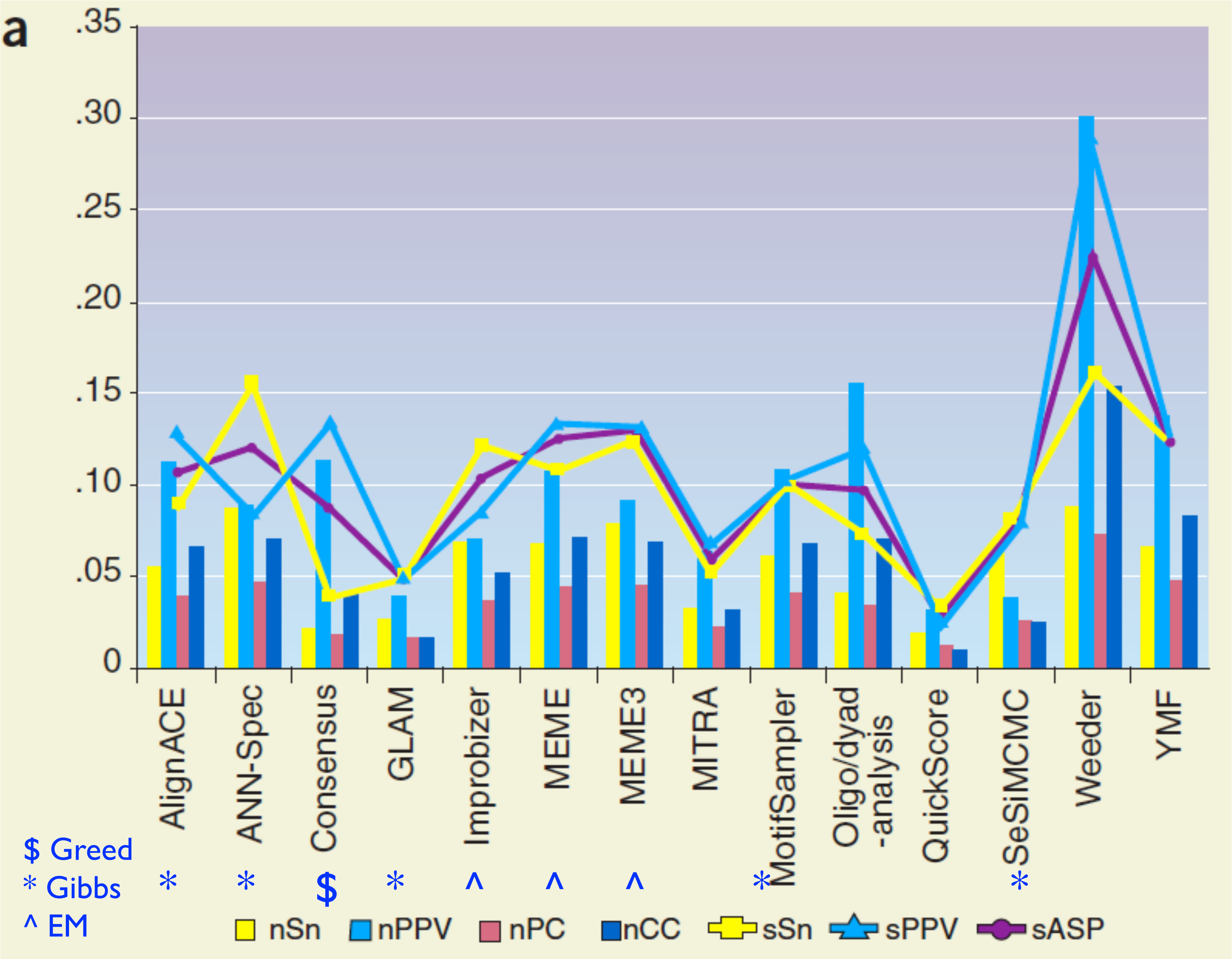
Real ‘motifs’ (Transfac)

56 data sets (human, mouse, fly, yeast)

‘Real’, ‘generic’, ‘Markov’

Expert users, top prediction only

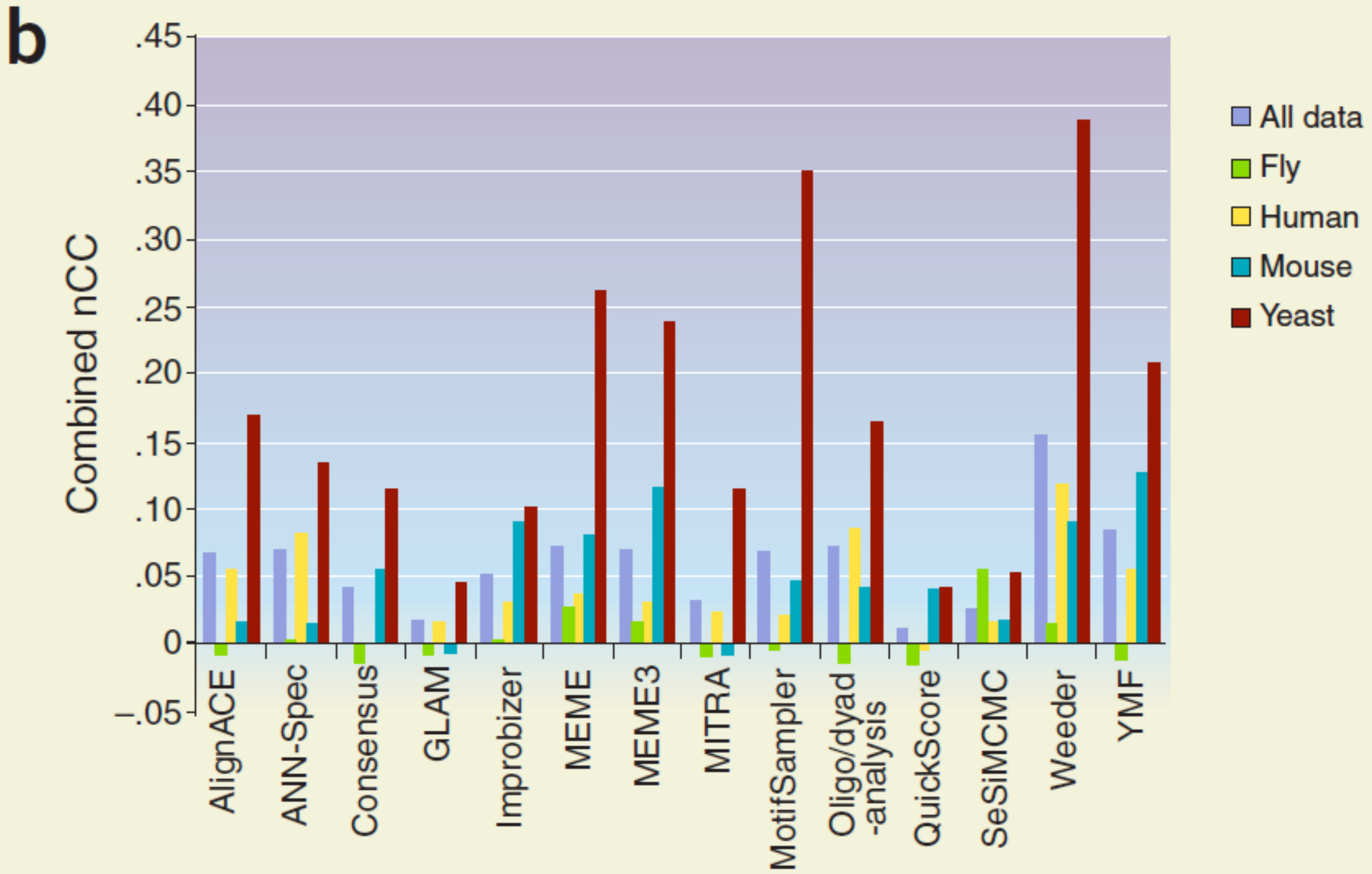
“Blind” – sort of



$ Greed

* Gibbs

^ EM


b
Combined nCC
.45
.40
.35
.30
.25
.20
.15
.10
.05
0
–.05
AlignACE
ANN-Spec
Consensus
GLAM
Improbizer
MEME
MEME3
MITRA
MotifSampler
Oligo/dyad -analysis
QuickScore
SeSiMCMC
Weeder
YMF
All data
Fly
Human
Mouse
Yeast

Lessons

Evaluation is hard (esp. when “truth” is unknown)

Accuracy low

  partly reflects limitations in evaluation methodology (e.g. ≤ 1 prediction per data set; results better in synth data)

  partly reflects difficult task, limited knowledge (e.g. yeast > others)

No clear winner re methods or models

Motif Discovery Summary

Important problem: a key to understanding gene regulation

Hard problem: short, degenerate signals amidst much noise

Many variants have been tried, for representation, search, and discovery. We looked at only a few:

- WMMs + wildcard patterns for representation & search
- Greedy, MEME and Gibbs for discovery

Still much room for improvement. *Comparative genomics*, i.e. cross-species comparison is very promising