

CSEP 527

5 – 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

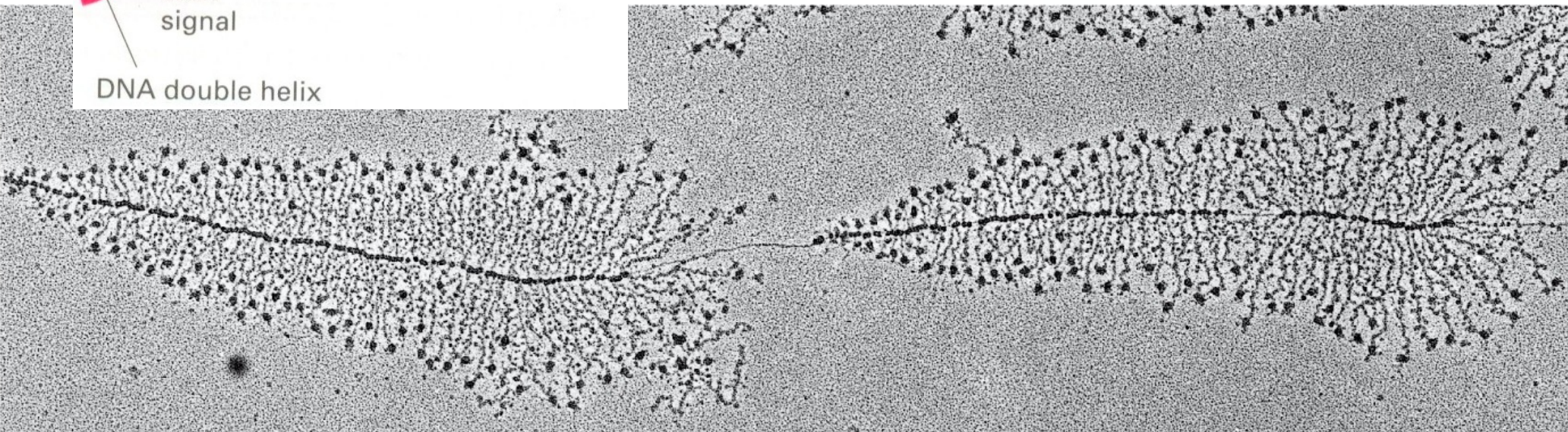
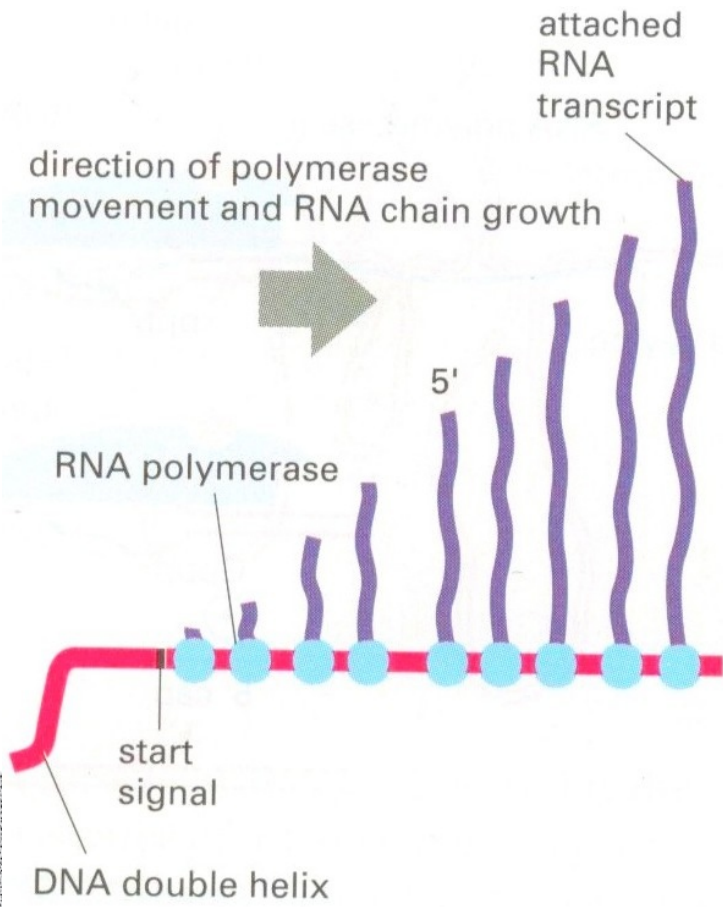
- is *transcribed* from DNA to RNA
- the mRNA is *processed* in various ways
- is *exported* from the nucleus (eukaryotes)
- is *translated* into protein

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)



1 μm

Regulation

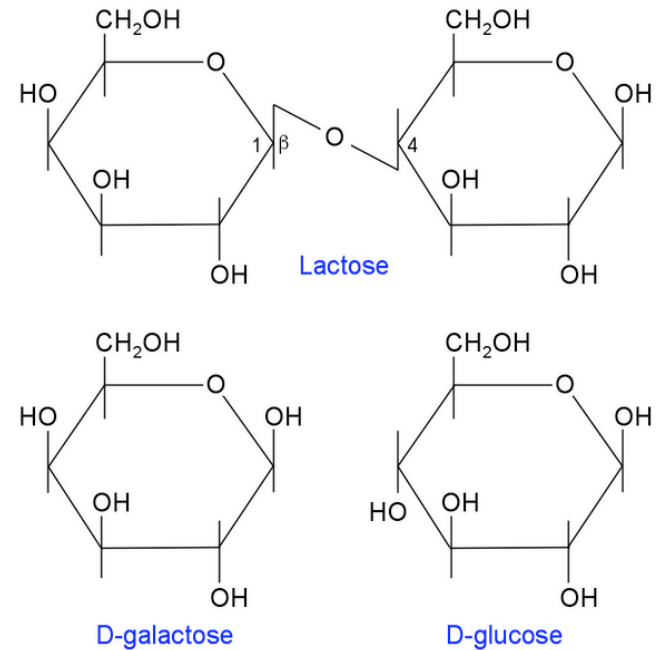
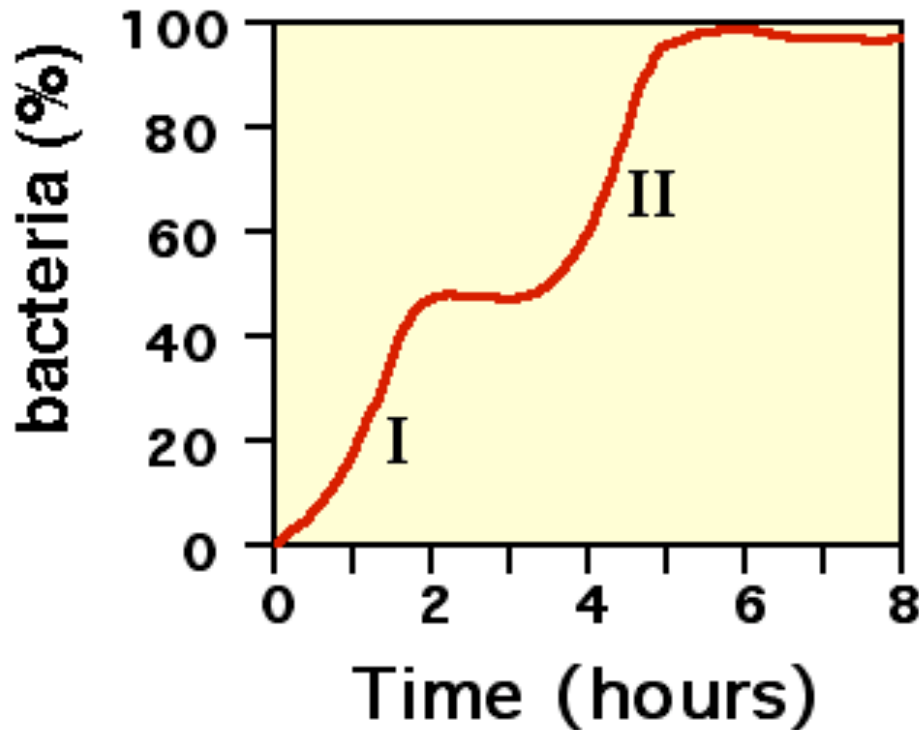
In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes

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, ...

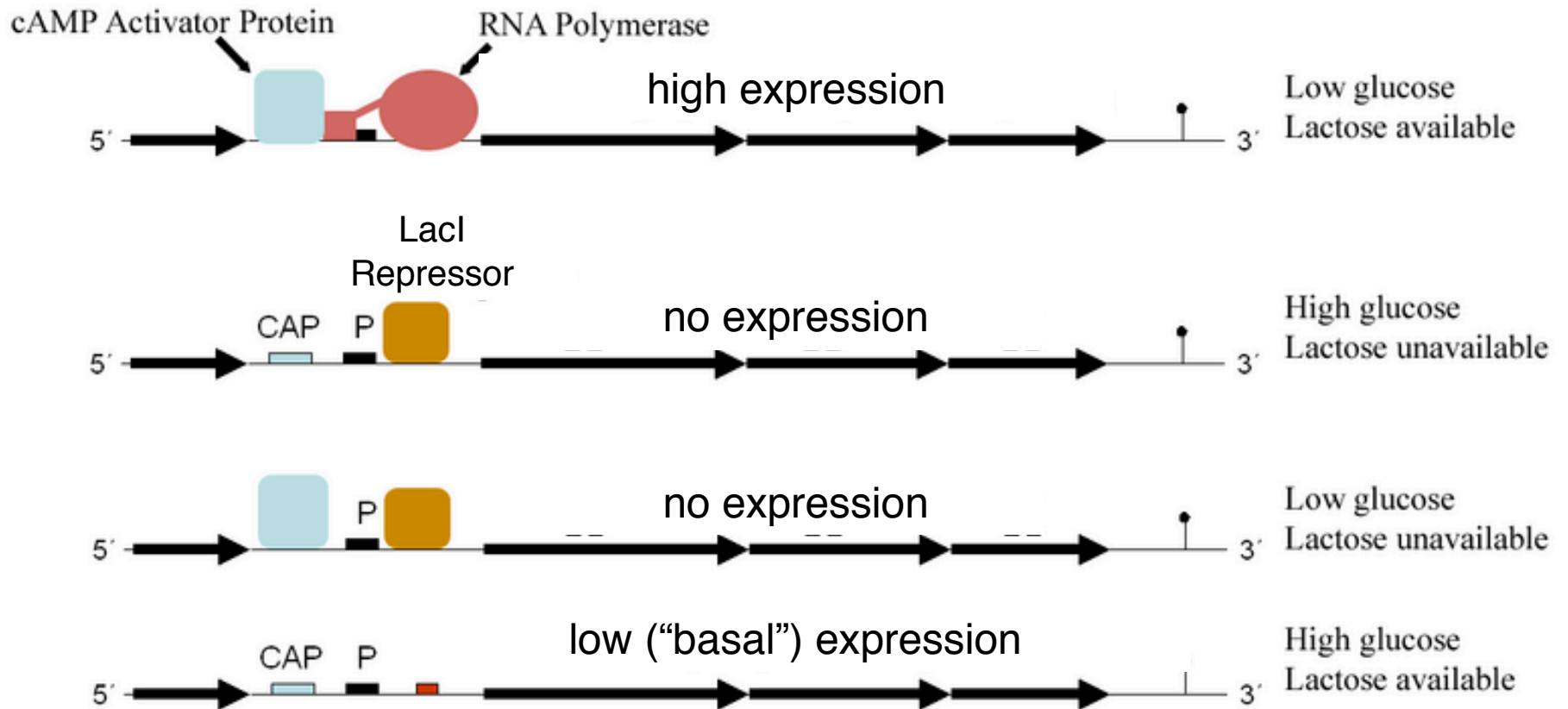
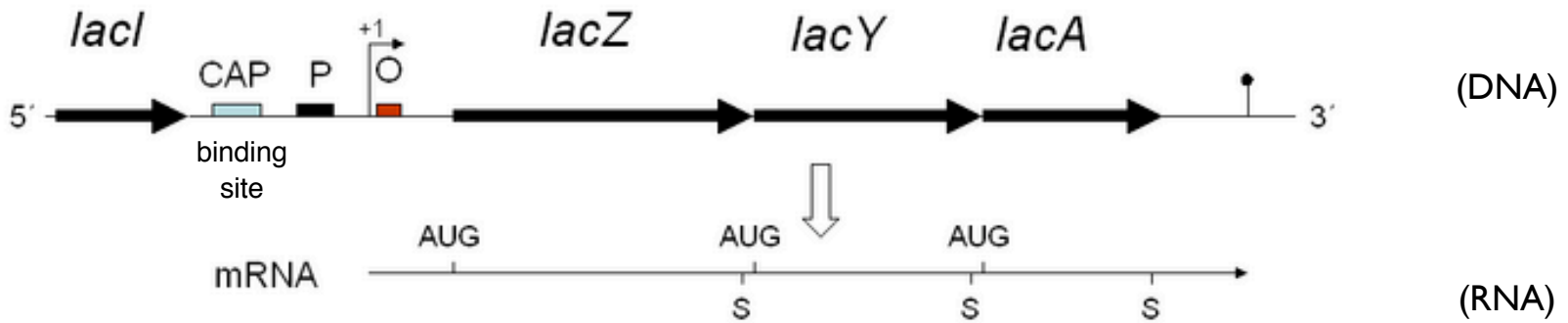
All are important, but below, focus on 1st step only:

- ✦ transcriptional regulation

E. coli growth on glucose + lactose



The *lac* Operon and its Control Elements



1965 Nobel Prize

Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff

1920-2013

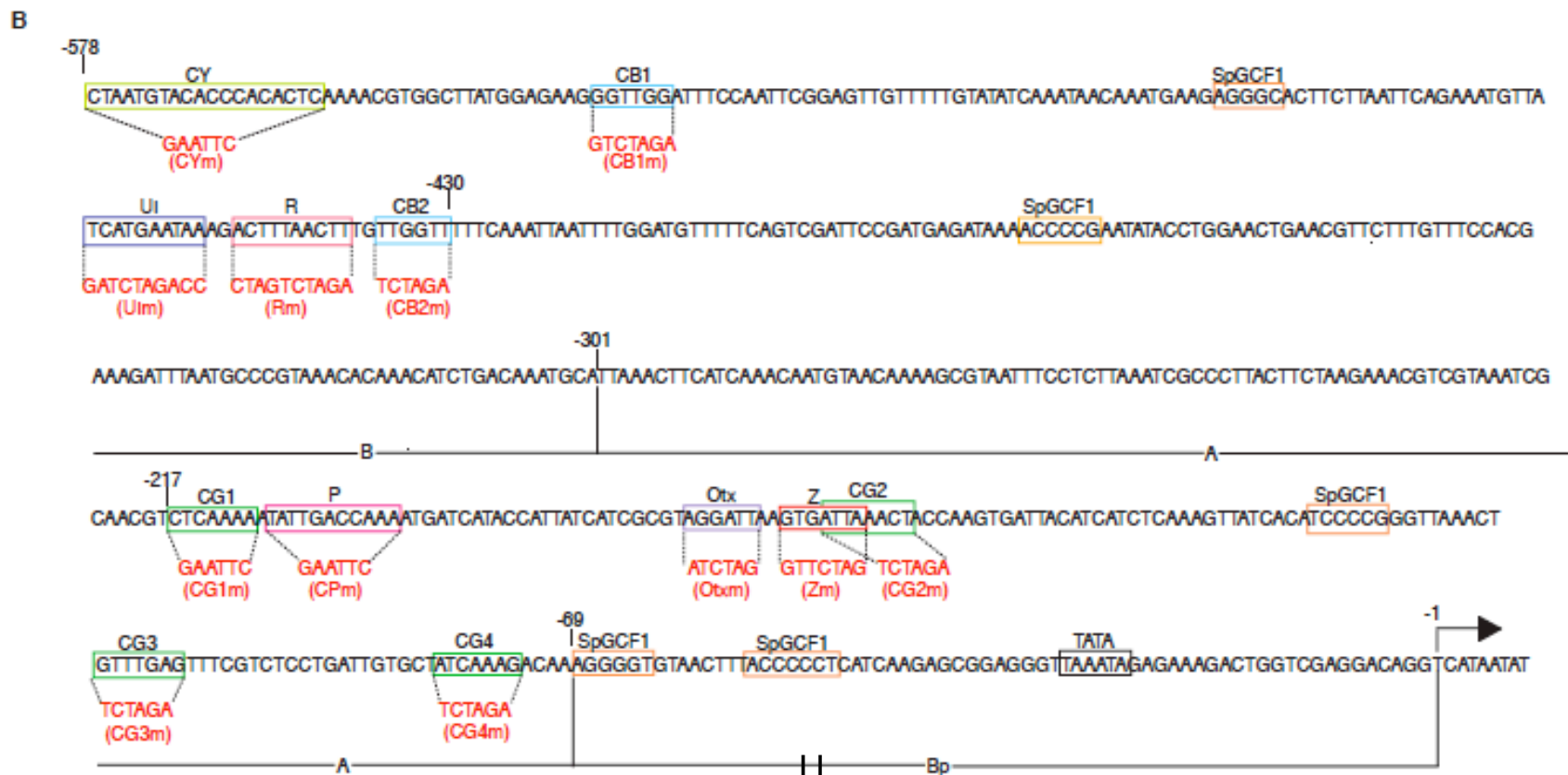
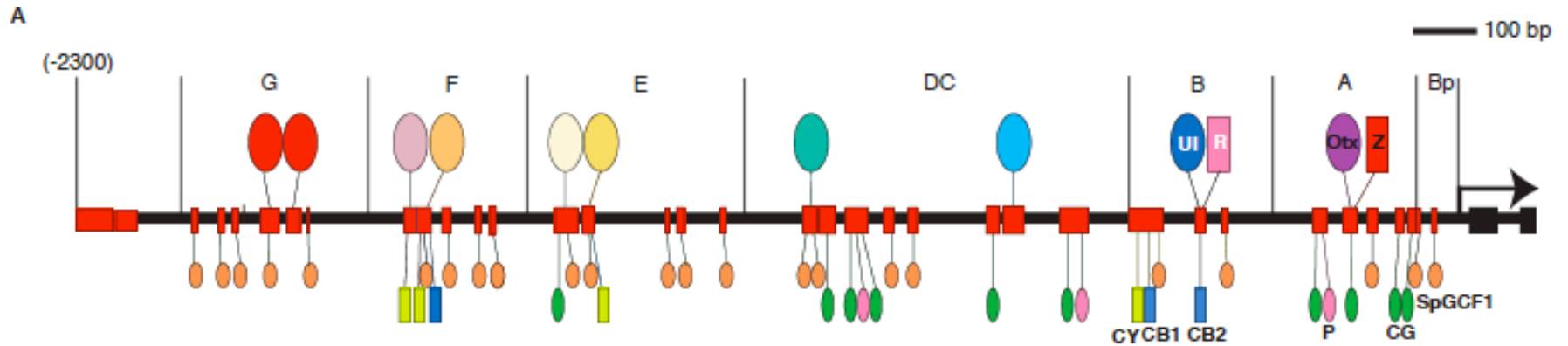
1910-1976

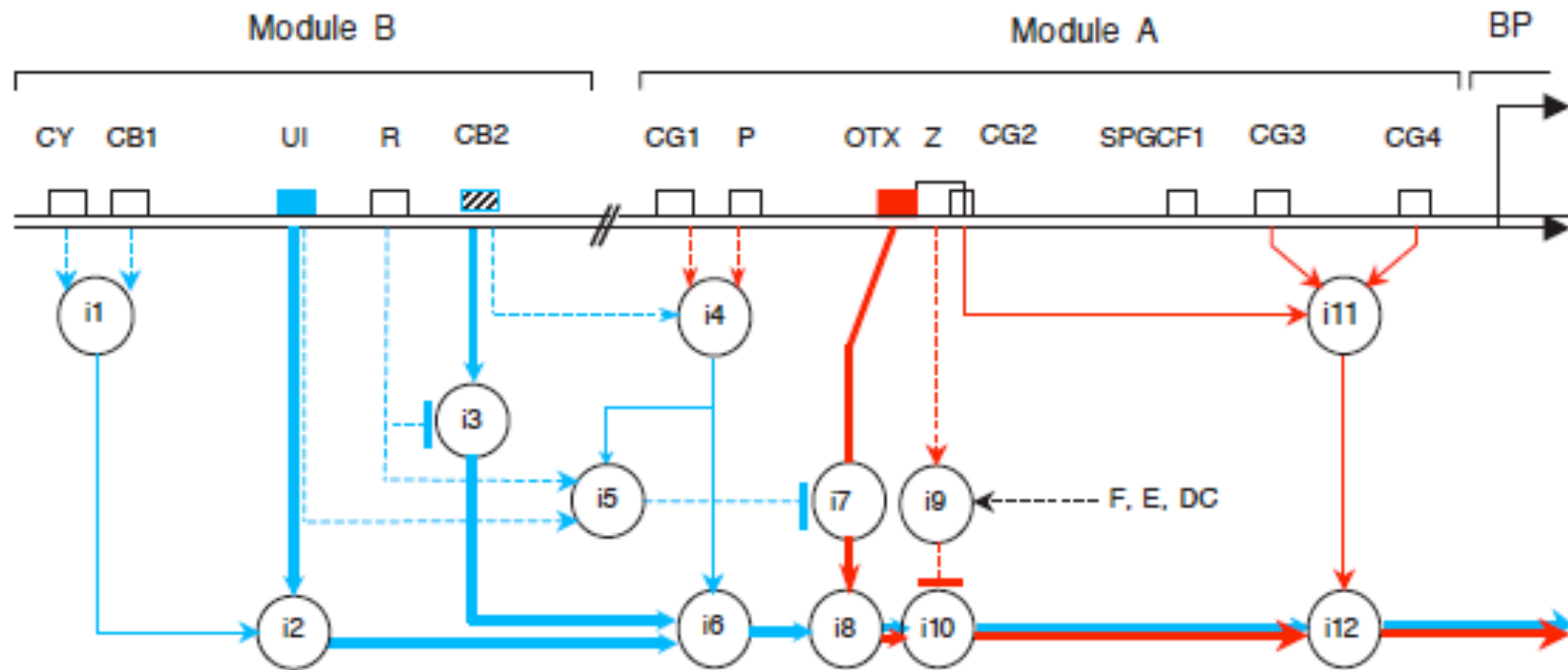
1902-1994



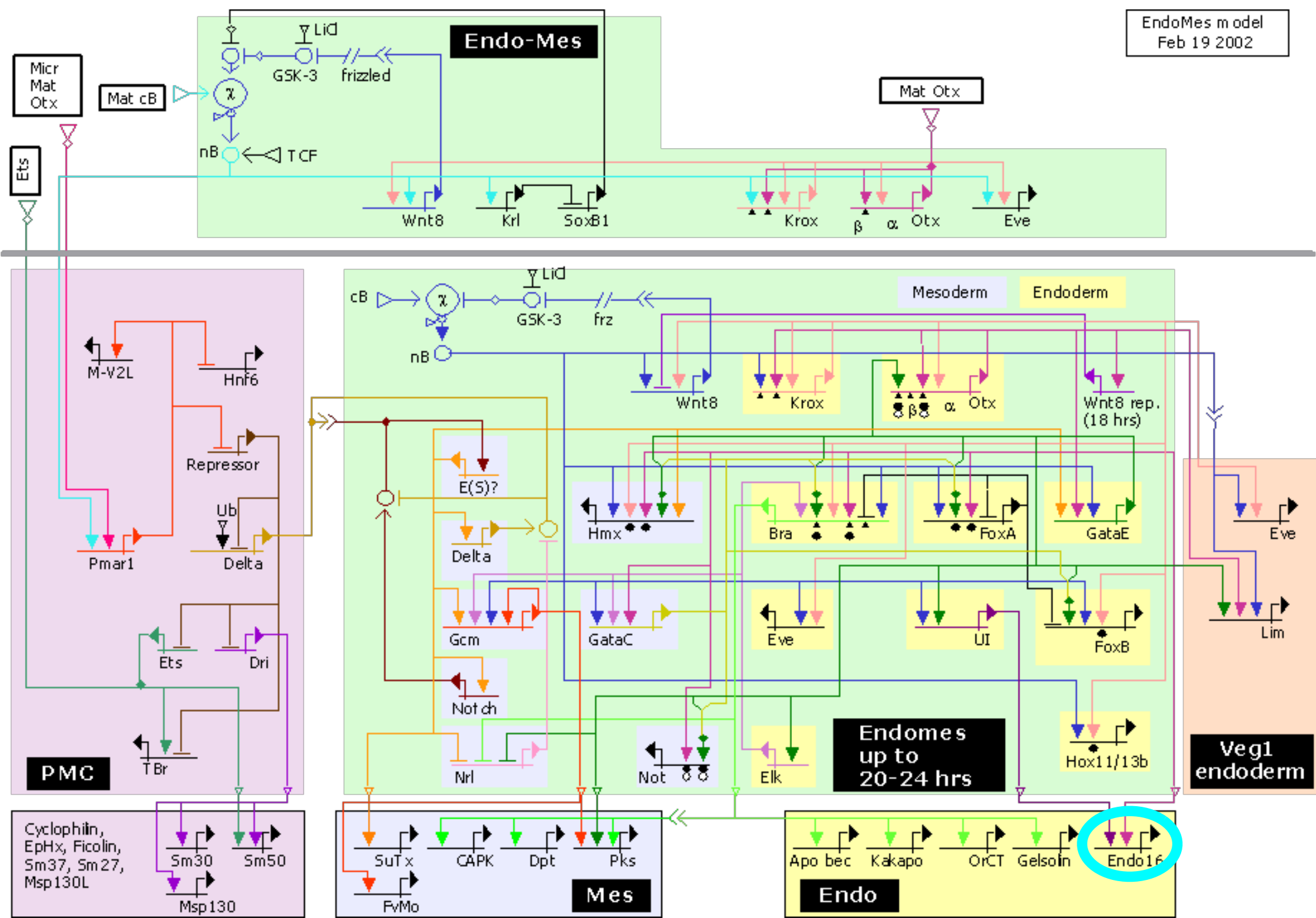
The sea urchin *Strongylocentrotus purpuratus*

Sea Urchin - Endo 16





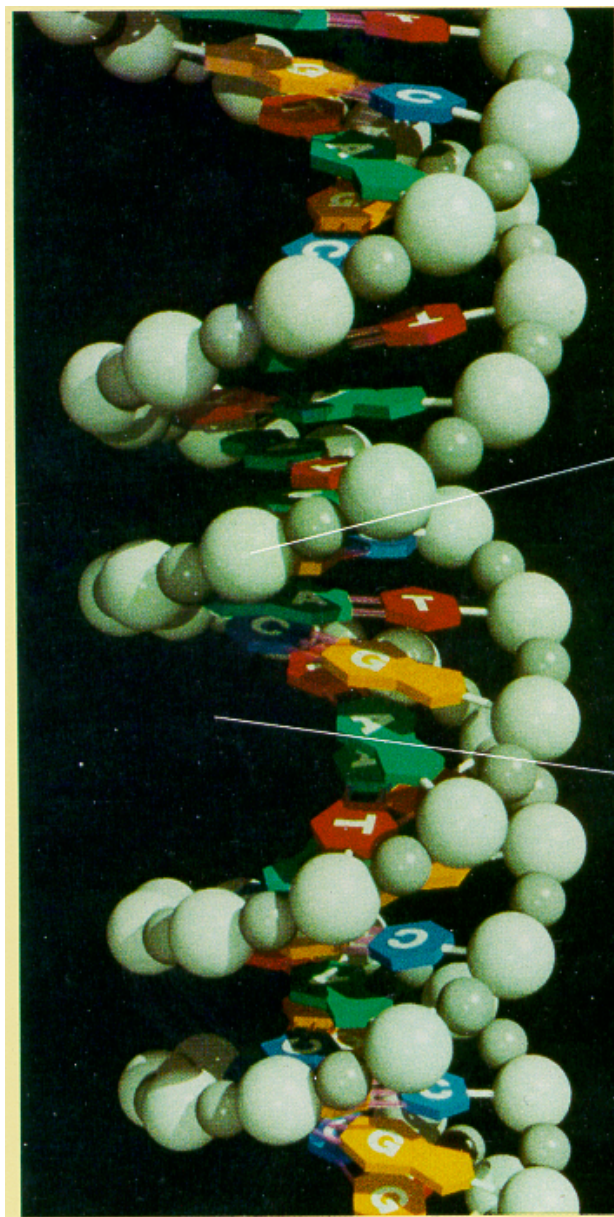
if CY & CB1 else	$i1 = 1$ $i1 = 0.5$	if $i5 = 0$ else	$i7 = OTX(t)$ $i7 = 0$
	$i2 = i1 \cdot UI(t)$		$i8 = i6 + i7$
if R else	$i3 = CB2(t)$ $i3 = k \cdot CB2(t)$ ($1 < k < 2$)	if (F or E or DC) & Z else	$i9 = 1$ $i9 = 0$
if P & CG1 & CB2 else	$i4 = 2$ $i4 = 0$	if $i9 = 1$ else	$i10 = 0$ $i10 = i8$
if $UI(t) > \text{threshold} \& R \& i4 \neq 0$ else	$i5 = 1$ $i5 = 0$	if (CG2 & CG3 & CG4) else	$i11 = 2$ $i11 = 1$
	$i6 = i4 \cdot (i2 + i3)$		$i12 = i11 \cdot i10$



DNA Binding Proteins

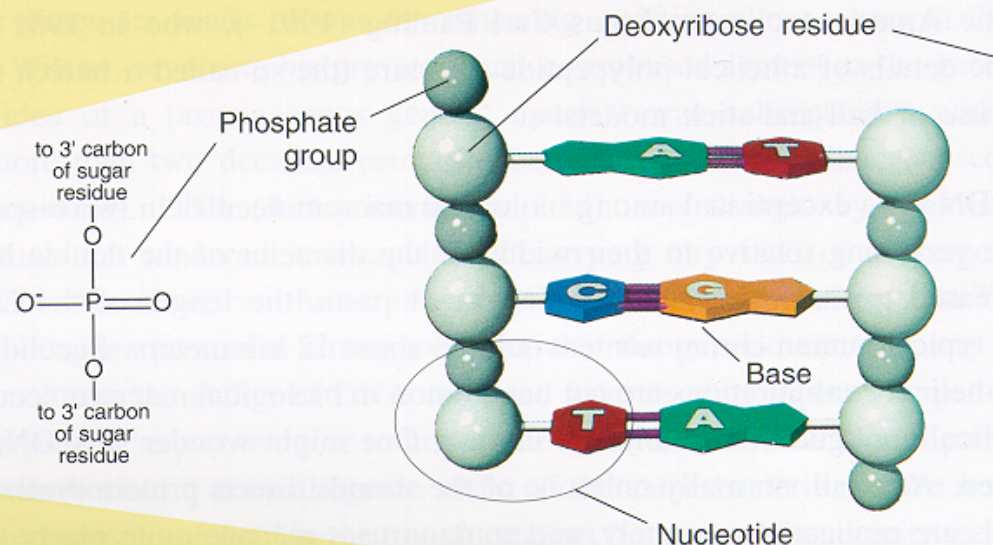
A variety of DNA binding proteins (so-called “transcription factors”; a significant fraction, perhaps 5-10%, of all human proteins) modulate transcription of protein coding genes

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 pair chemist's viewpoint, each strand a polymer made up of four repeating units called deoxyribonucleotides

In the groove

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

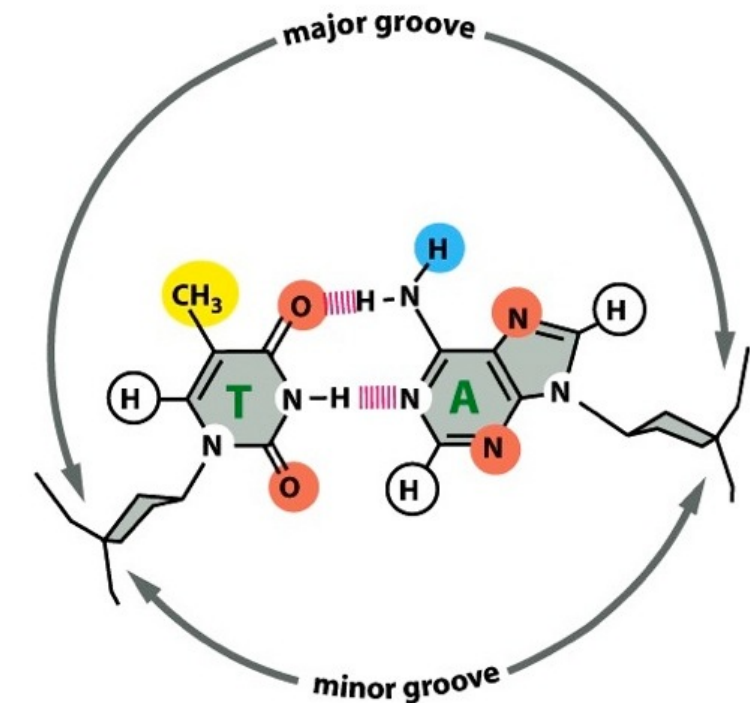
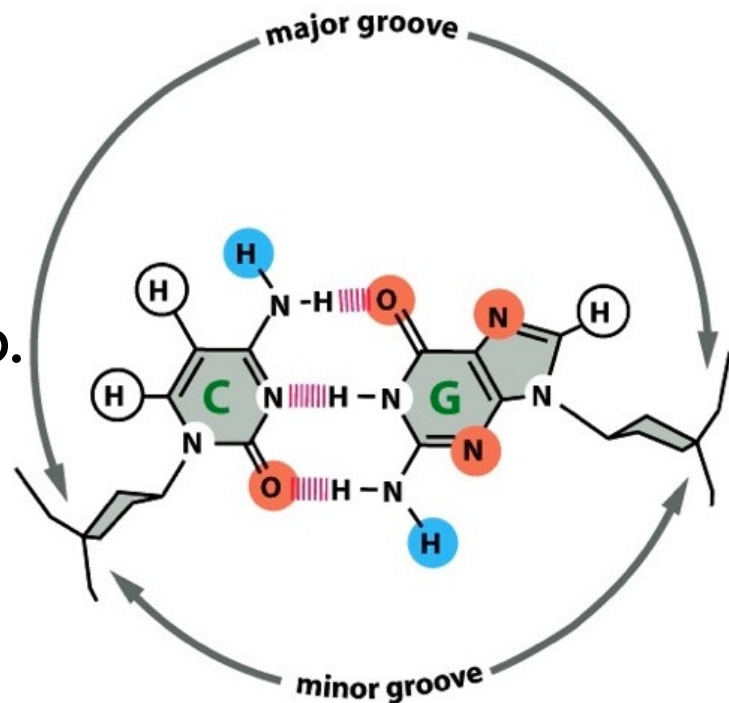
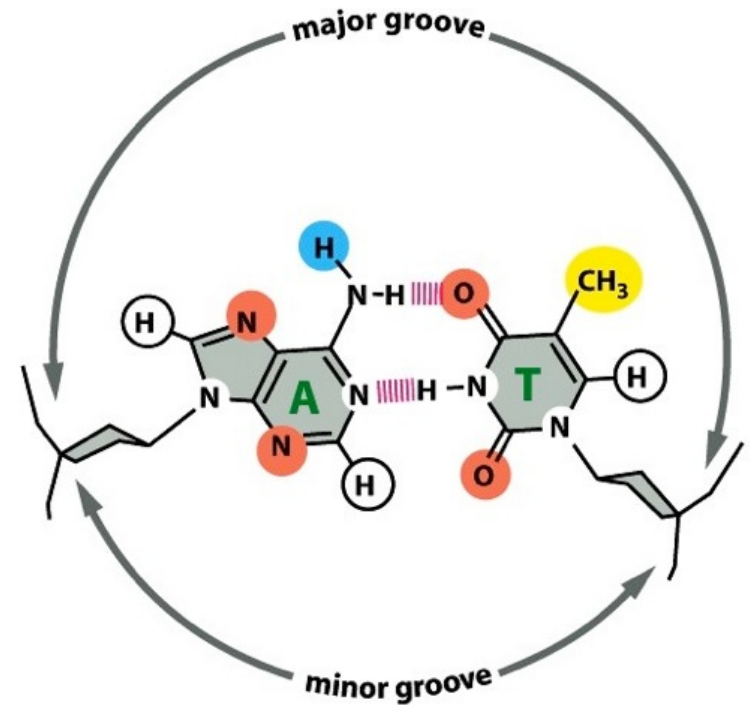
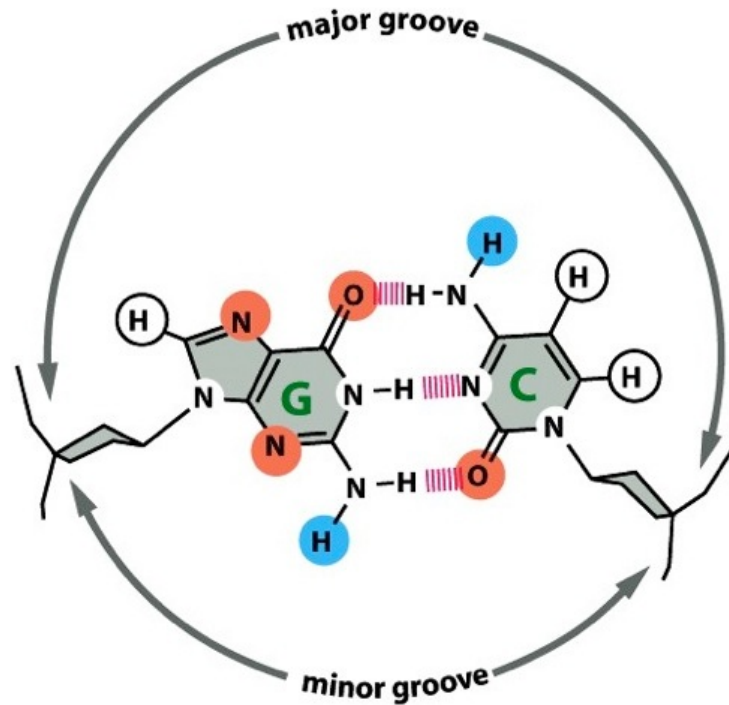


Figure 7-7 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Helix-Turn-Helix DNA Binding Motif

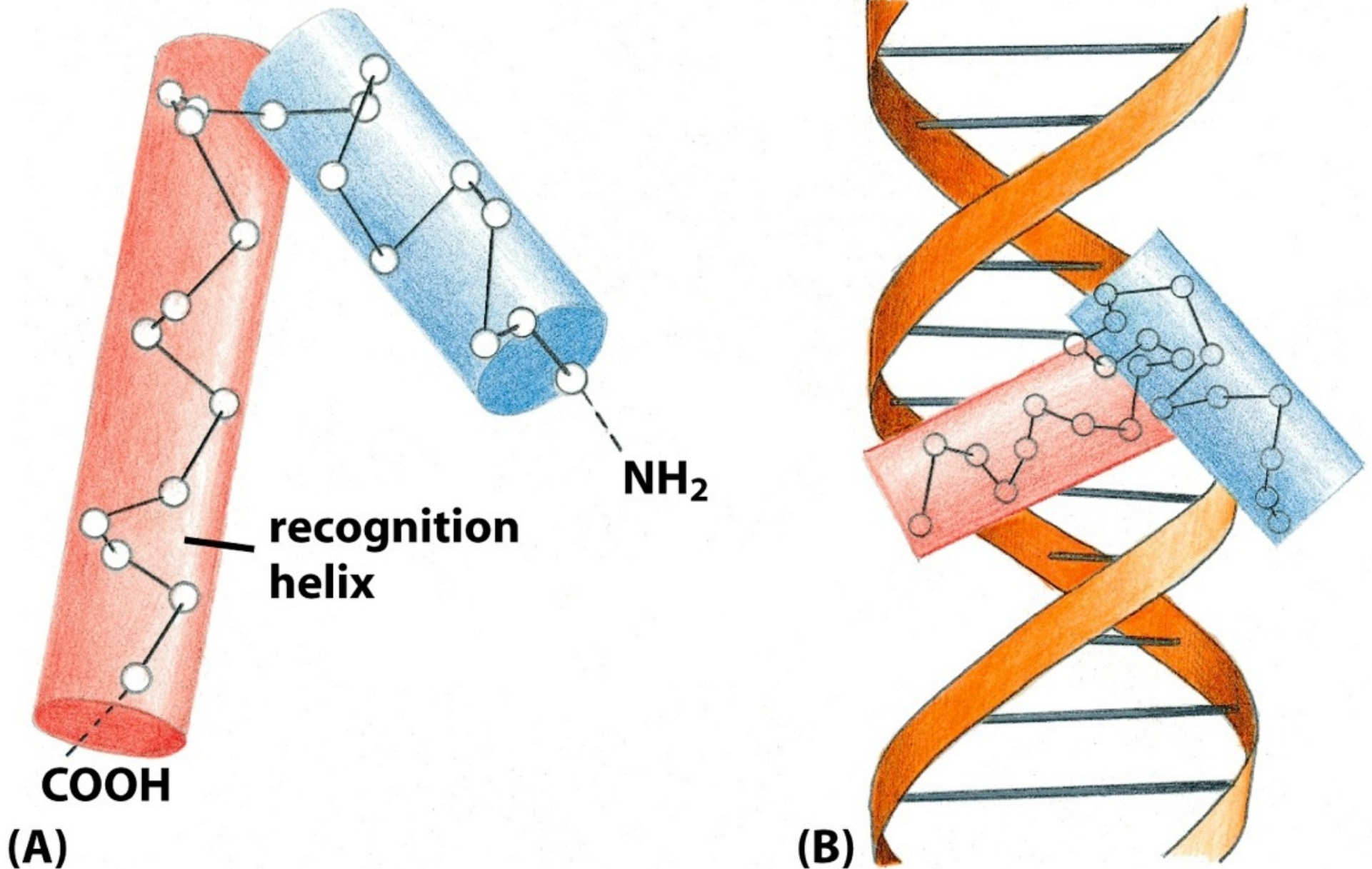


Figure 7-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)

H-T-H Dimers

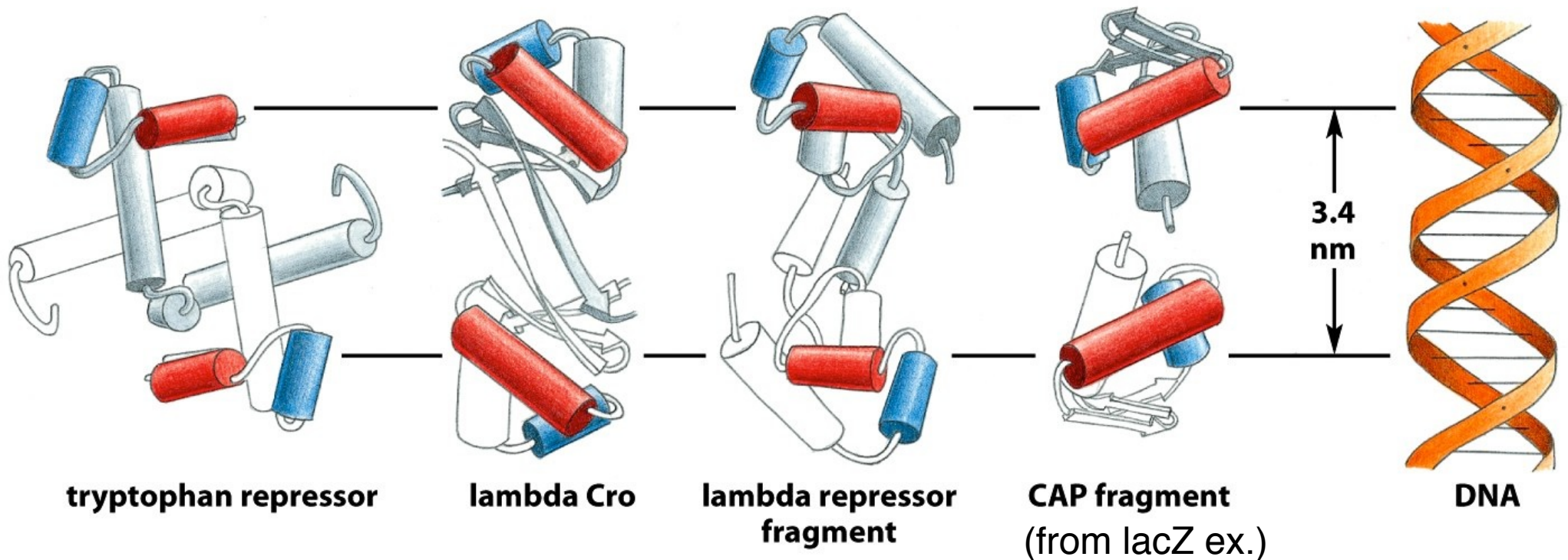
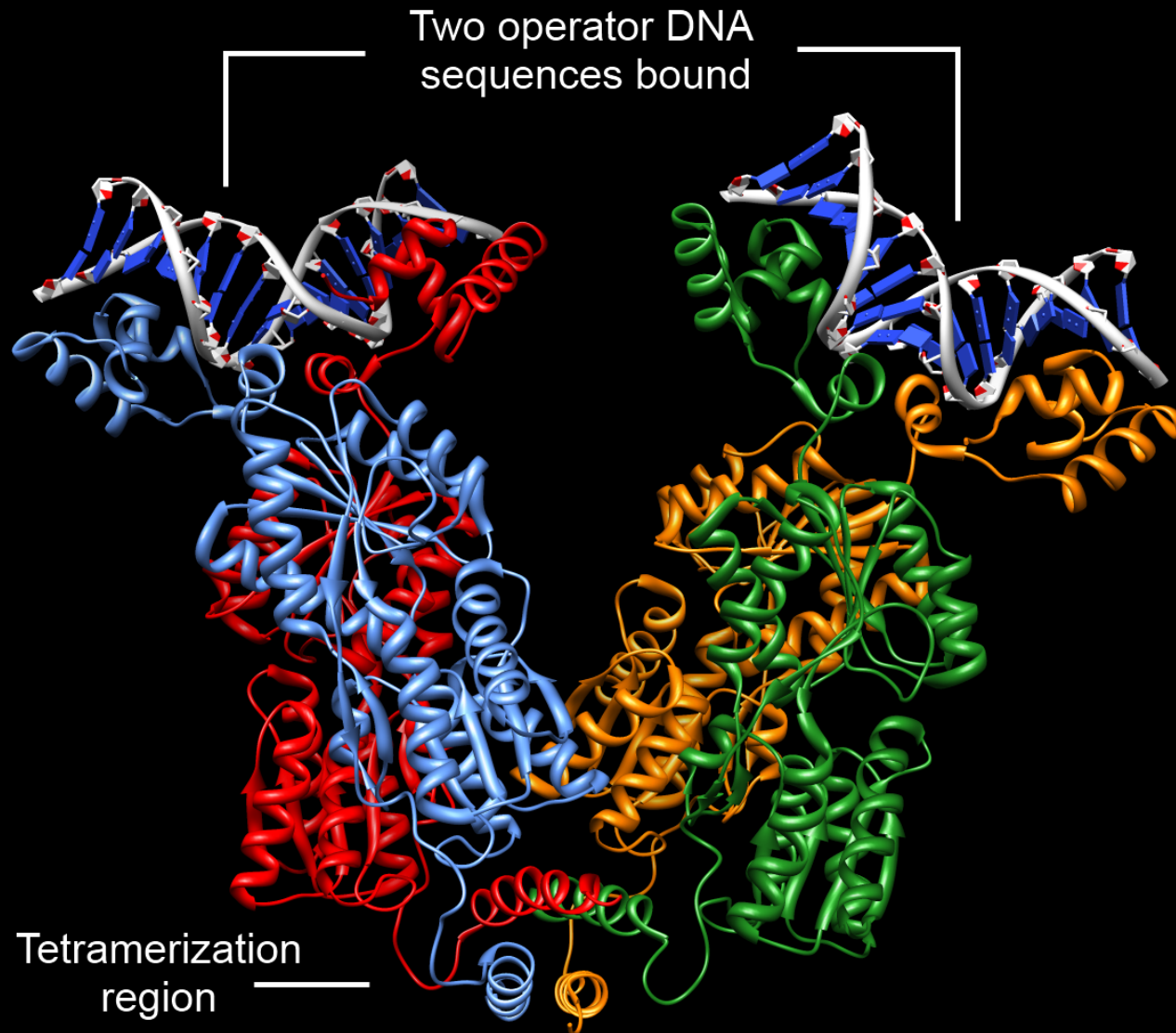


Figure 7-11 Molecular Biology of the Cell 5/e (© Garland Science 2008)

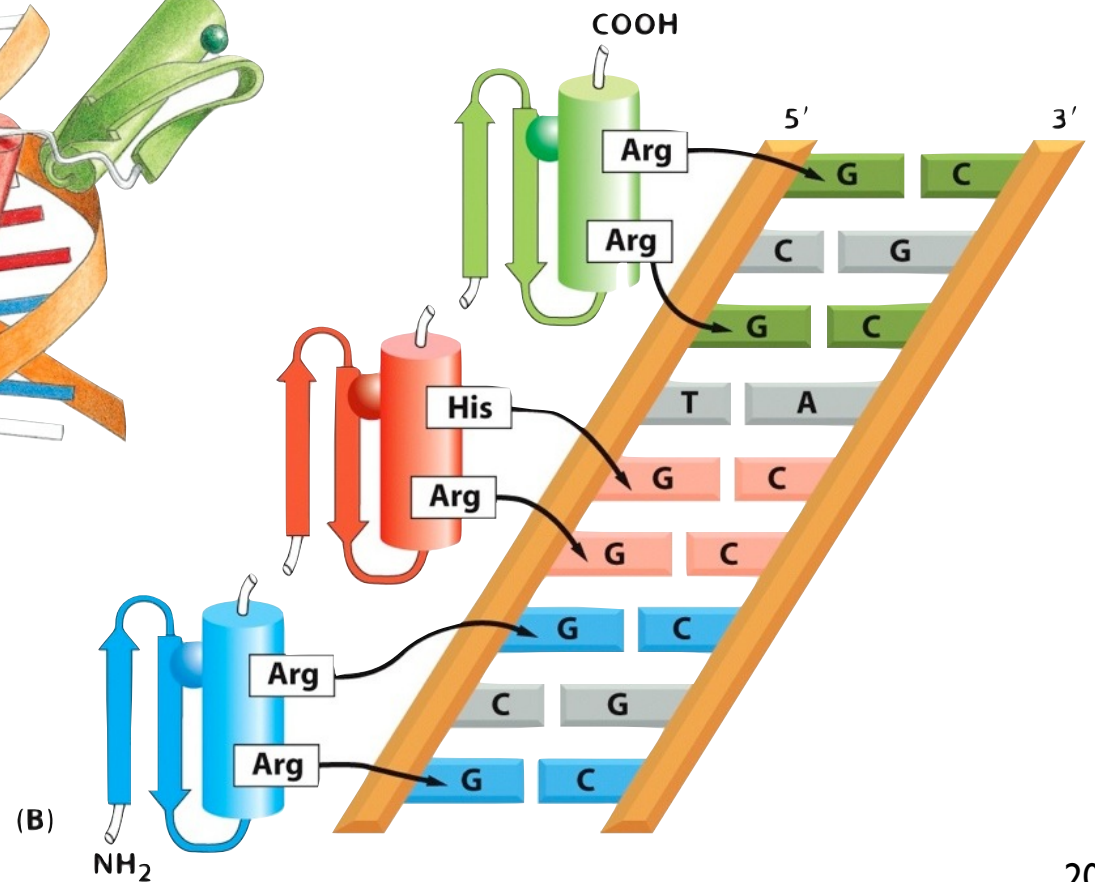
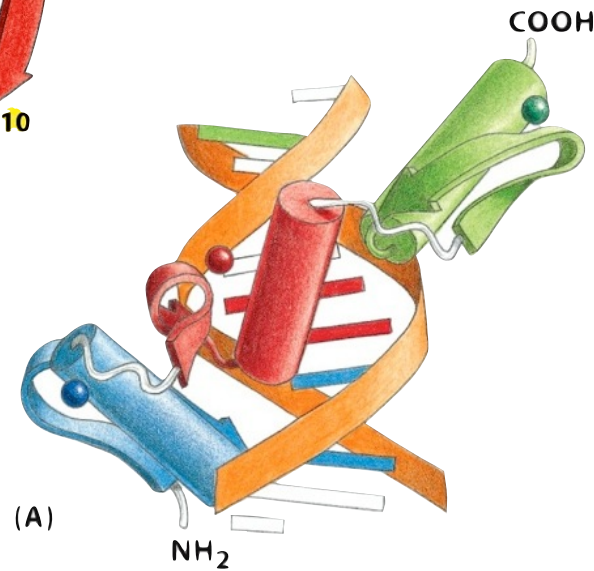
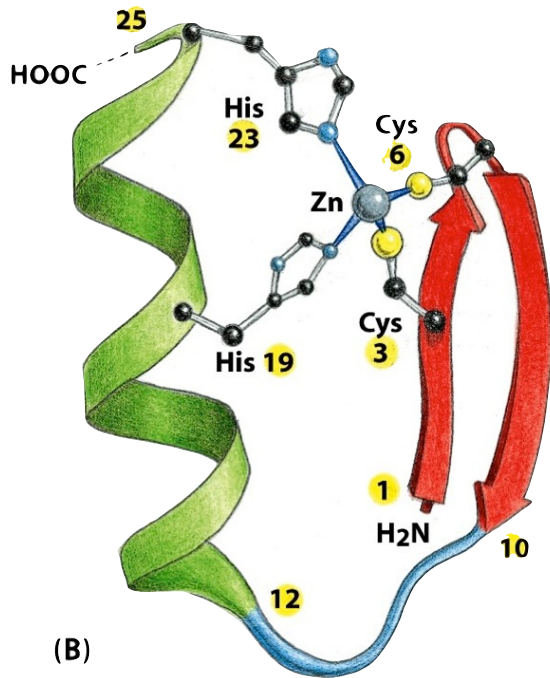
Bind 2 DNA patches, ~ 1 turn apart
Increases both specificity and affinity

LacI Repressor + DNA

(a tetrameric HTH protein)



Zinc Finger Motif



Overheard at the Halloween Party



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

Leucine Zipper Motif

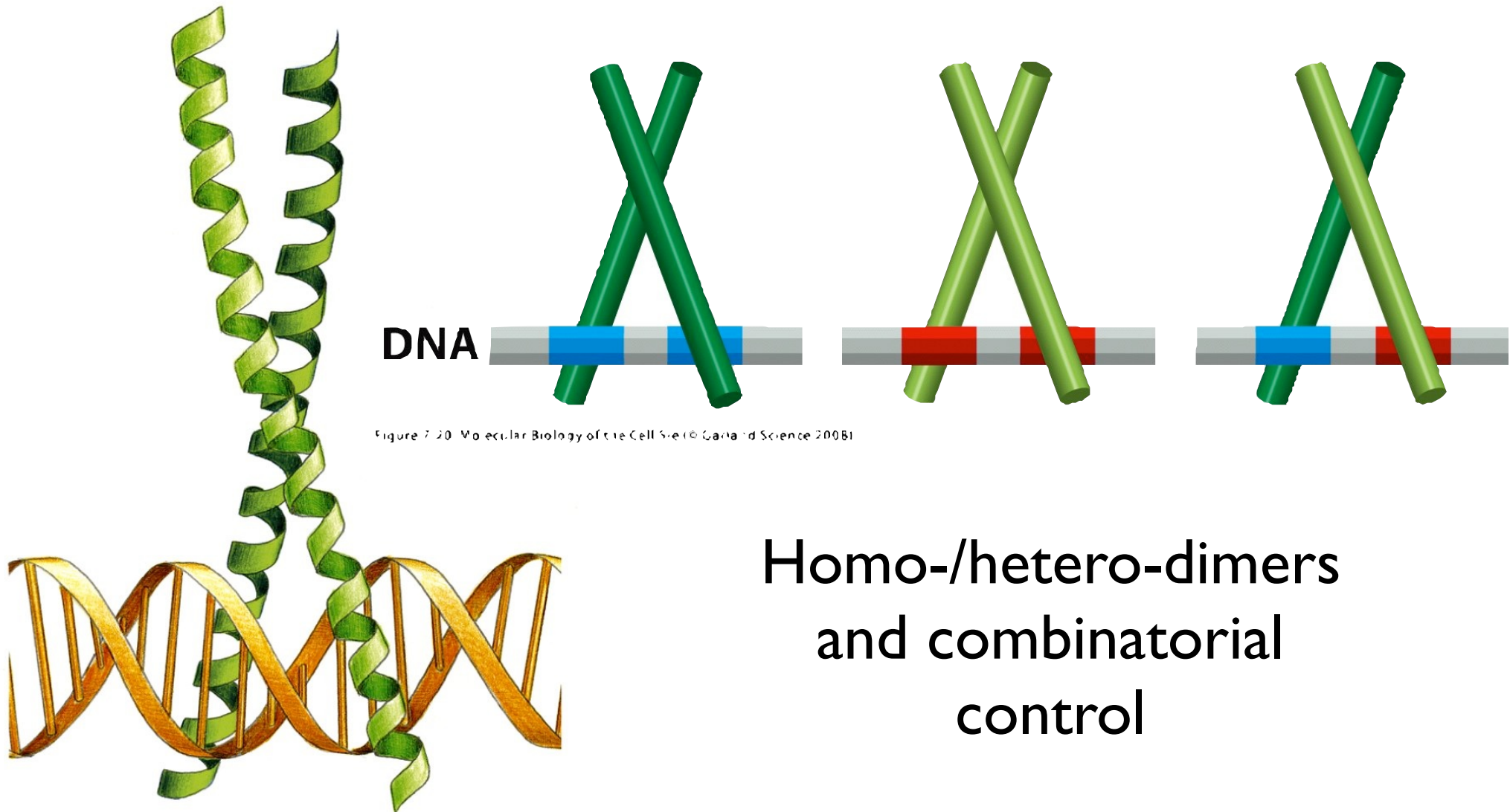


Figure 7-20 Molecular Biology of the Cell 5/e © Garland Science 2008

Homo-/hetero-dimers
and combinatorial
control

MyoD



jmol S

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

We understand some Protein/DNA interactions

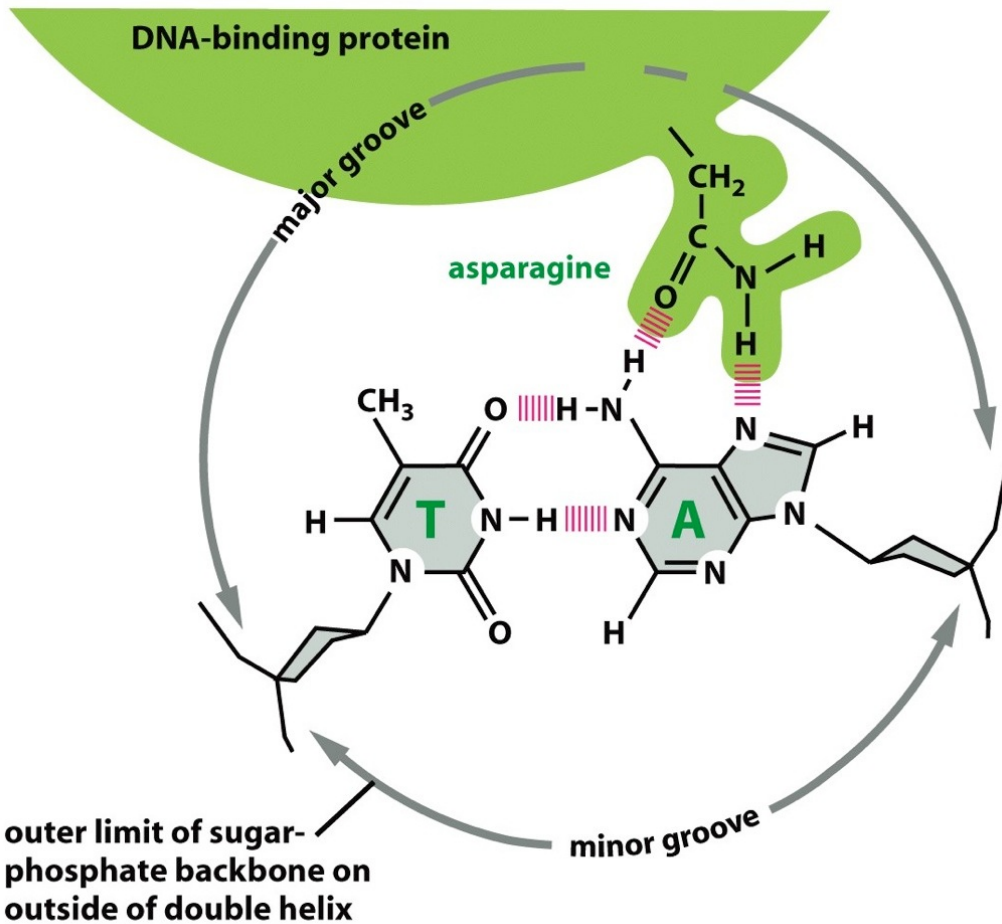


Figure 7-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

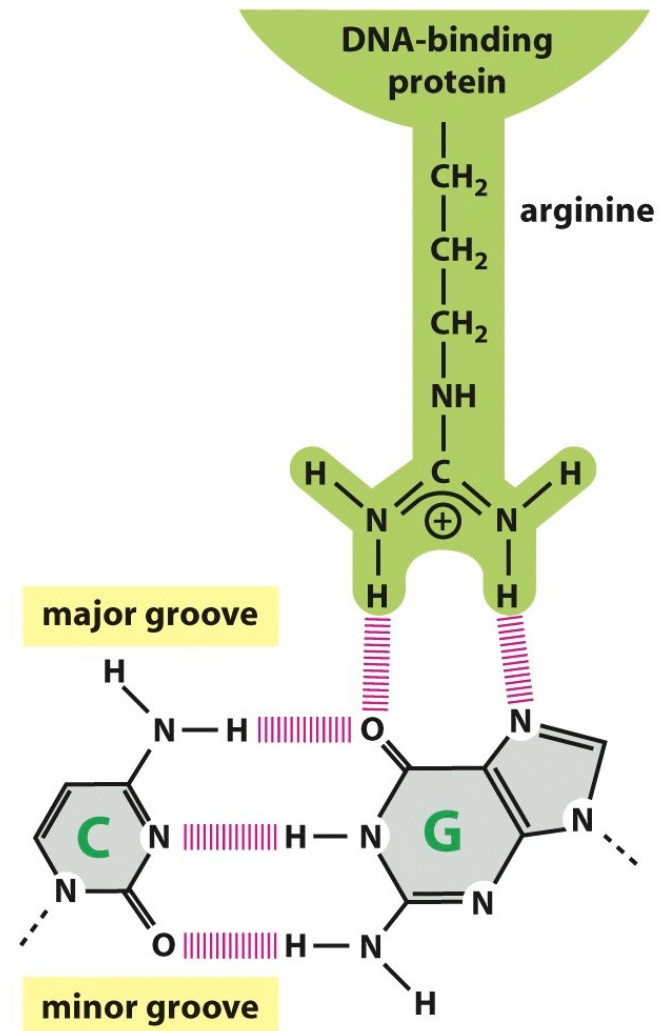


Figure 7-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)

But the overall DNA binding “code” still defies prediction

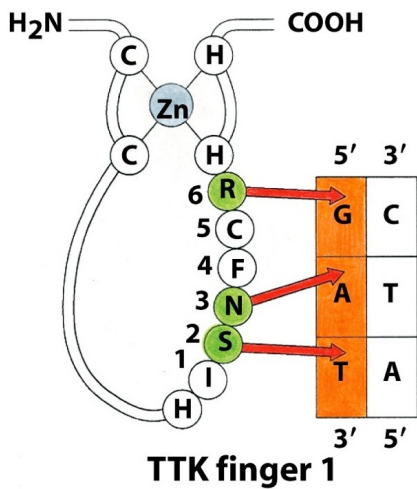


Figure 7-26 part 1 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

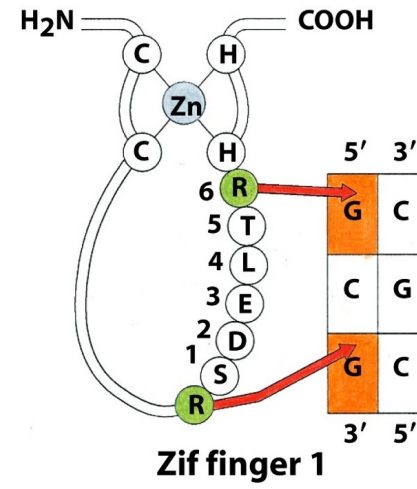
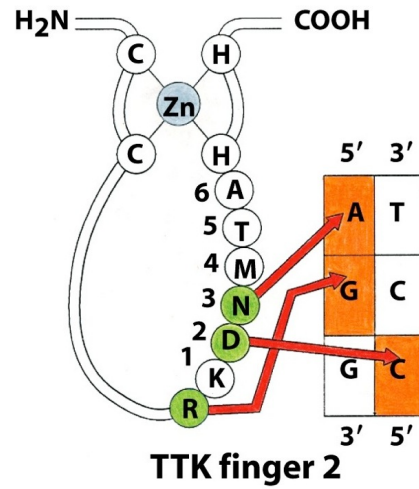


Figure 7-26 part 2 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

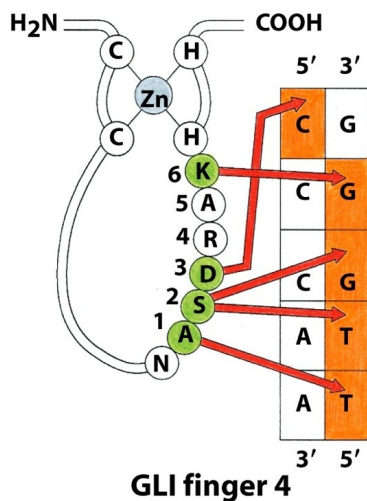
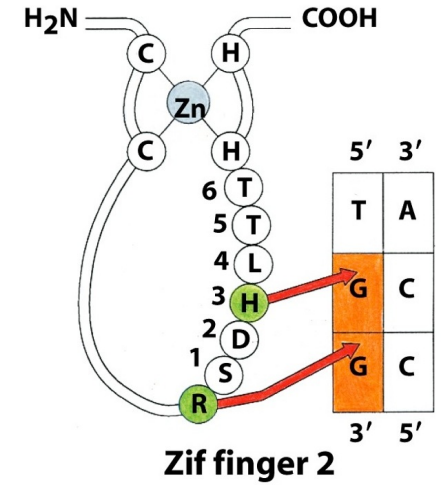
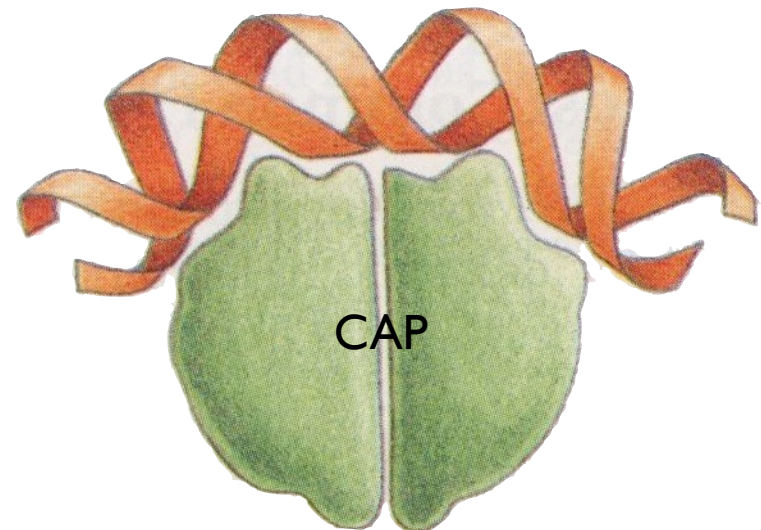
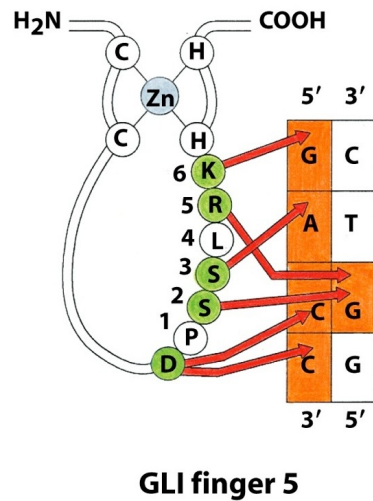


Figure 7-26 part 3 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



CAP, LacI repressor, and others bend DNA

Summary

Proteins can “bind” DNA to regulate gene expression (i.e., production of proteins, including themselves)

This is widespread

Complex, combinatorial control is both possible and commonplace

Sequence Motifs

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			
A	C	T	A	T	A	A	T	C	G

= -90

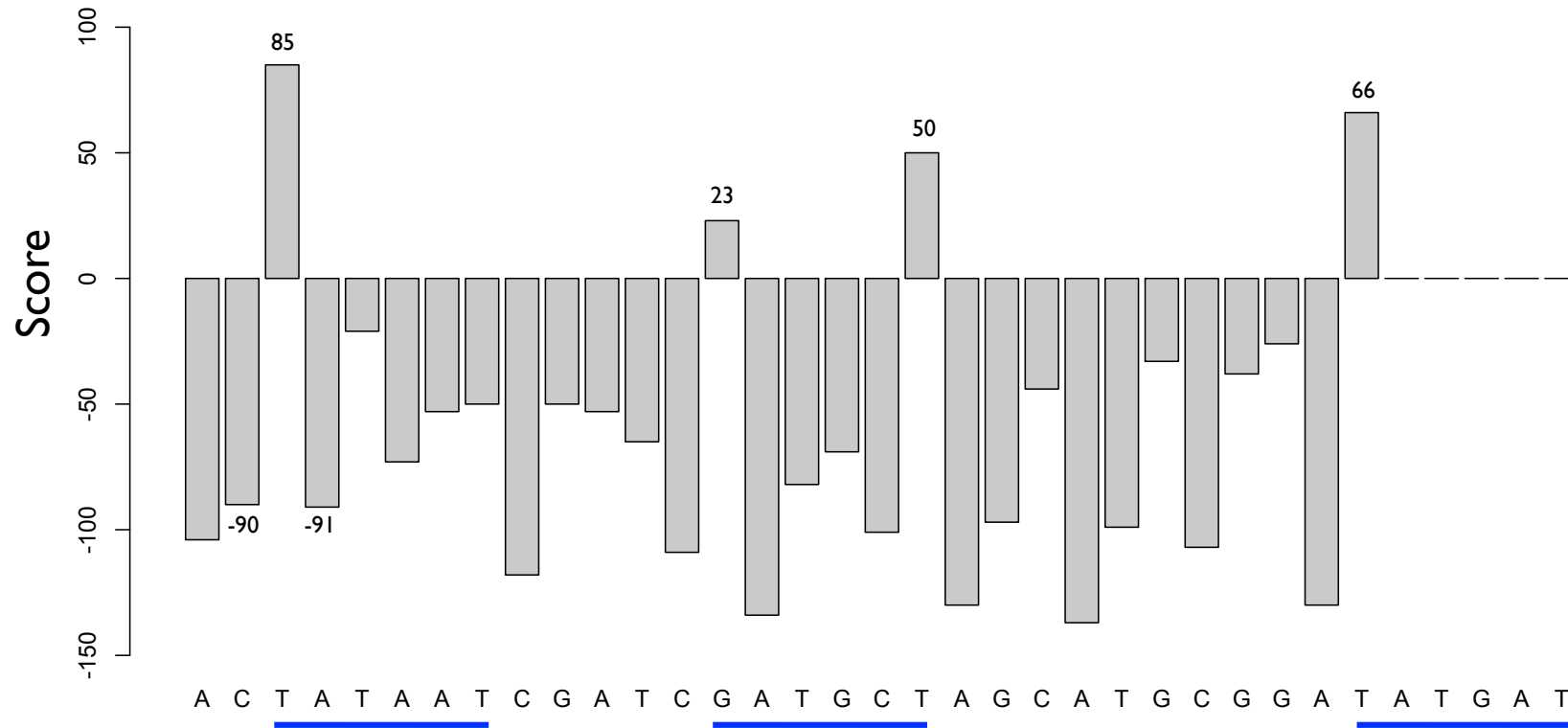
A	-36	19	1	12	10	-46			
C	-15	-36	-8	-9	-3	-31			
G	-13	-46	-6	-7	-9	-46			
T	17	-31	8	-9	-6	19			
A	C	T	A	T	A	A	T	C	G

= 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

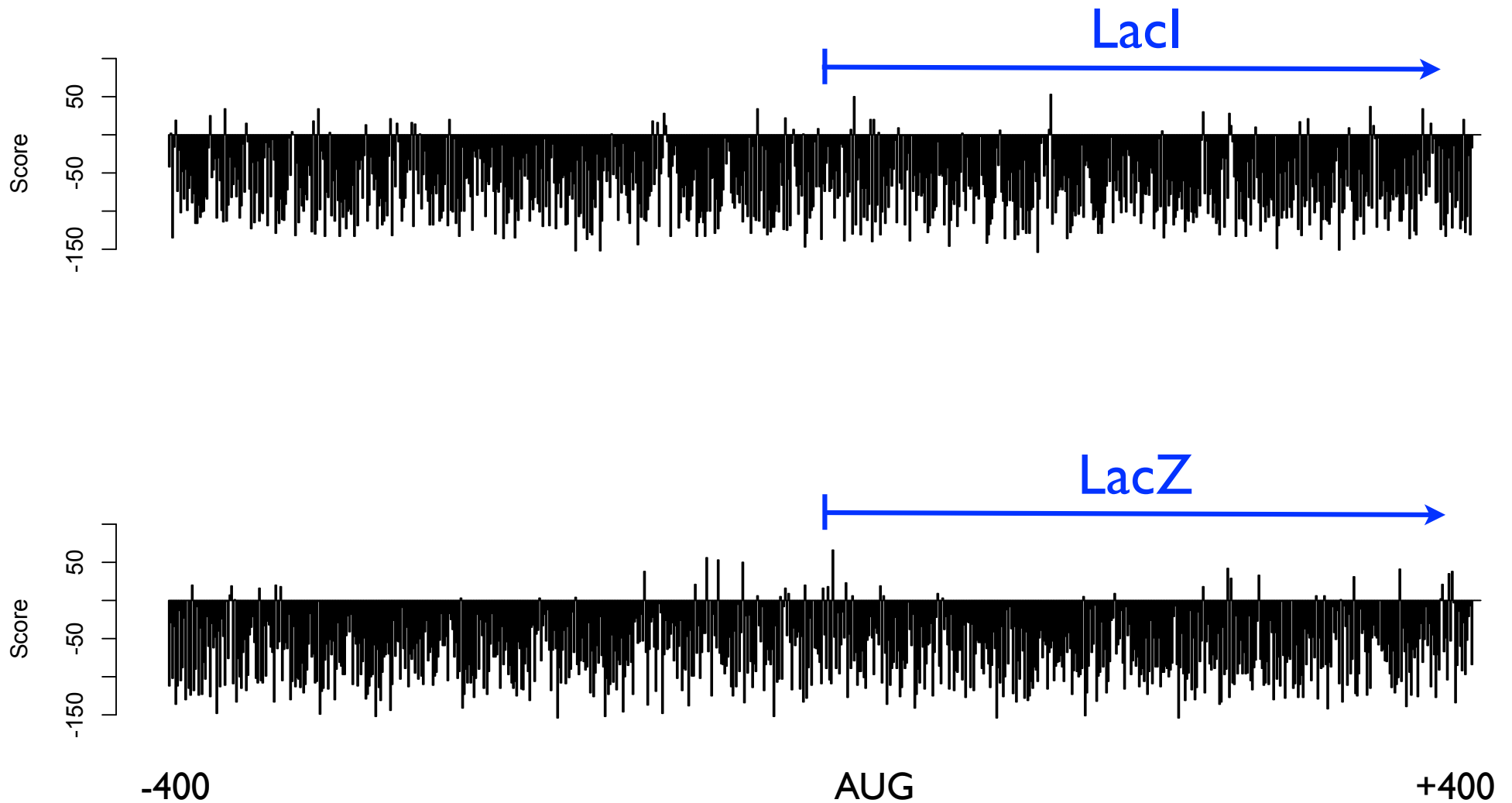
= -91

Scanning for TATA



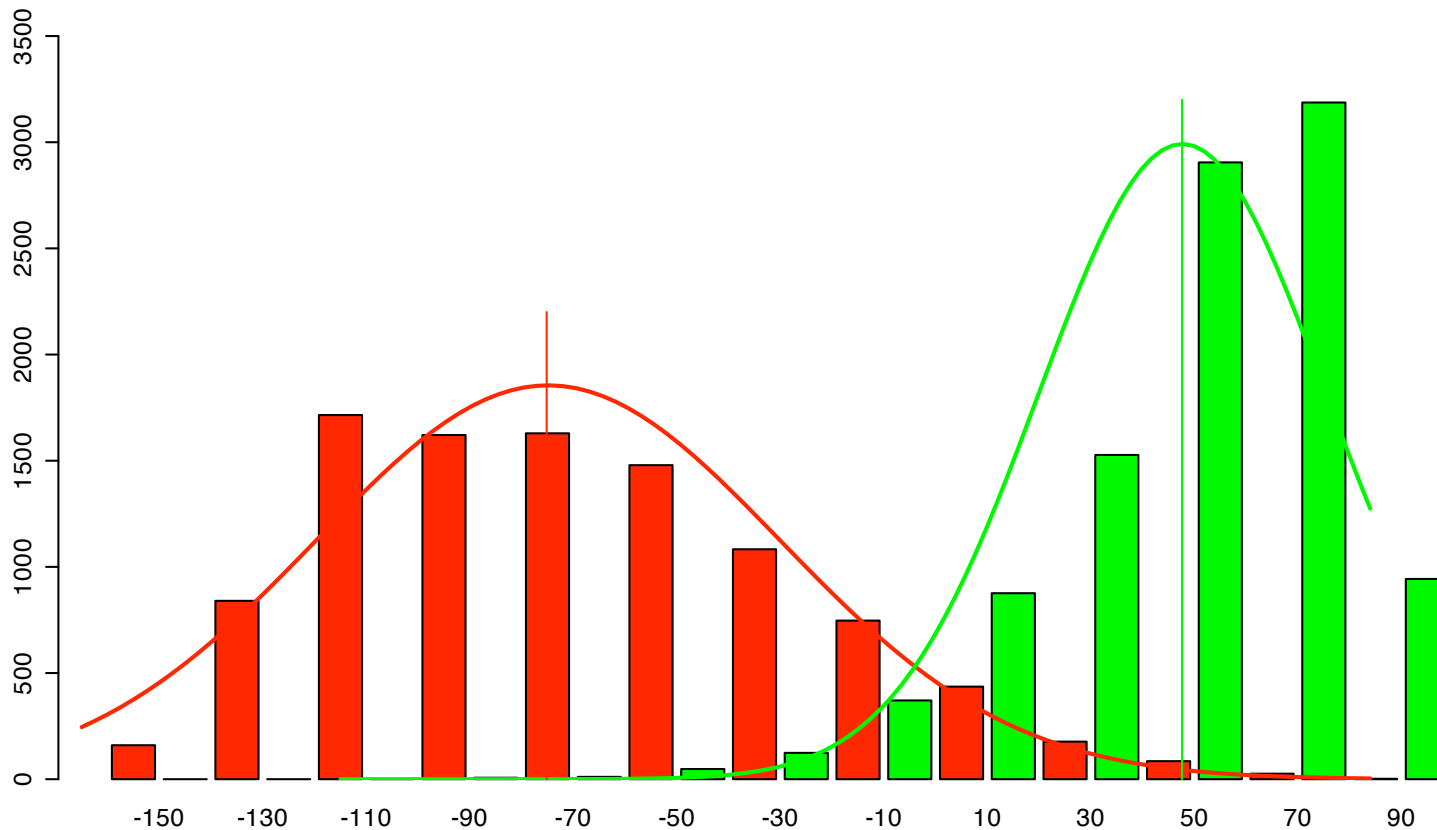
[See also slide 63](#)

TATA Scan at 2 genes



Score Distribution

(Simulated)



10^4 random 6-mers from foreground (green) or uniform background (red)₃₇

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 $S = B_1B_2...B_6$:

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

Assumes independence

Neyman-Pearson

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

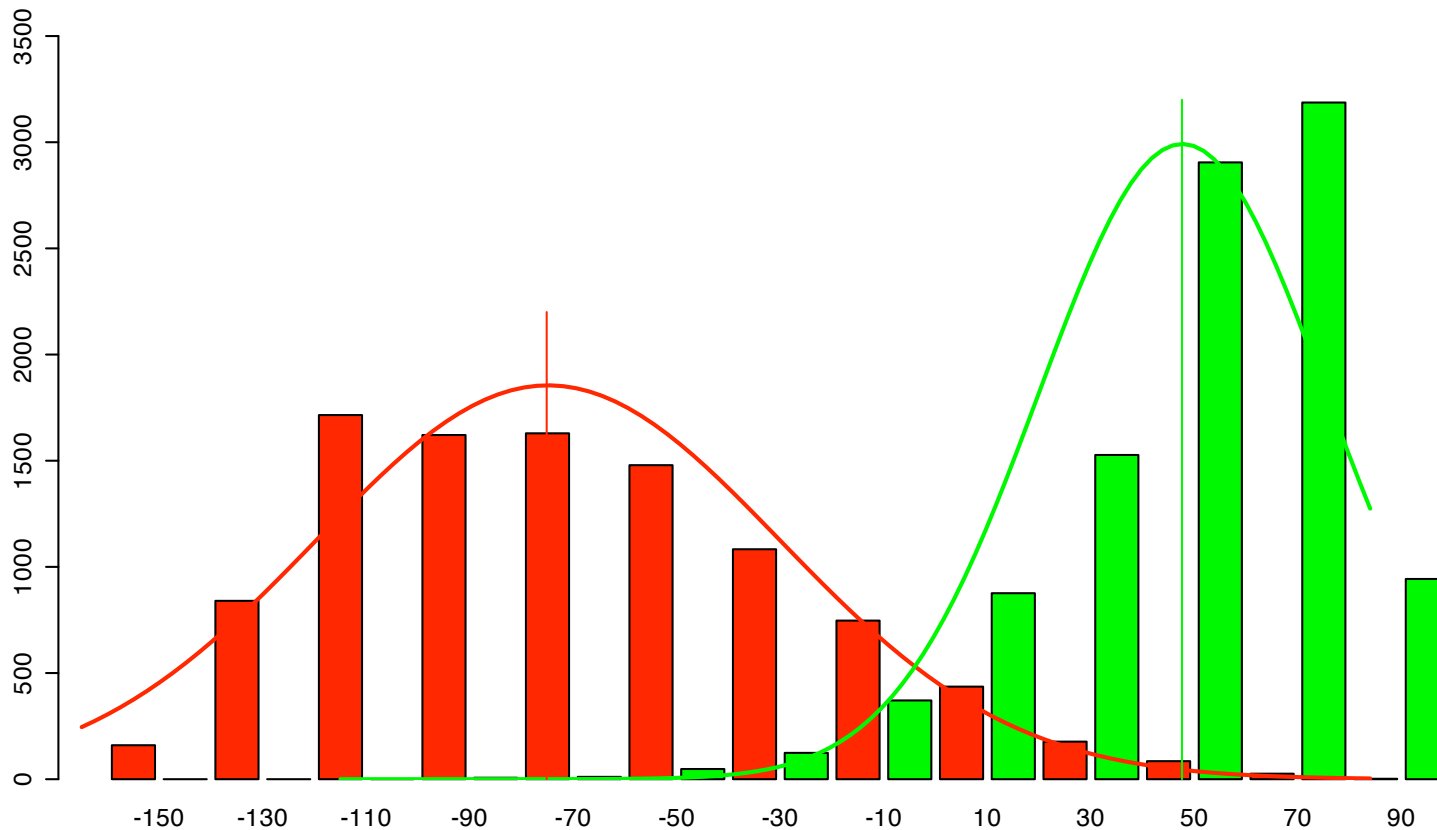
Might as well look at *likelihood ratio*:

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

(or *log likelihood ratio*)

Score Distribution

(Simulated)



10^4 random 6-mers from foreground (green) or uniform background (red)₄₀

What's best WMM?

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

E.g., what's MLE for θ given data s_1, s_2, \dots, s_k ?

Answer: like coin flips or dice rolls, count frequencies per position. (Possible HW?)

Weight Matrices: Biophysics

Experiments show ~80% correlation of log likelihood weight matrix scores to measured binding energies [Fields & Stormo, 1994]

i.e.,

- $\log \text{prob} \propto \text{energy}$
- “independence assumption” \Rightarrow
 - probabilities multiply & energies are additive

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	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	0	$-\infty$	2
T	-1	2	$-\infty$

Log-Likelihood Ratio:

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

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	$-\infty$	$-\infty$
C	$-\infty$	$-\infty$	$-\infty$
G	1	$-\infty$	3
T	-1.58	1.42	$-\infty$

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

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

Notes:

The “sample space”

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

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

WMM: How “Informative”?

Mean score of site vs bkg?

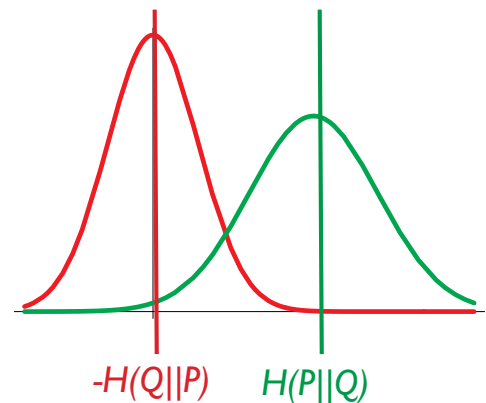
For any fixed length sequence x , let

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

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

Relative Entropy:

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

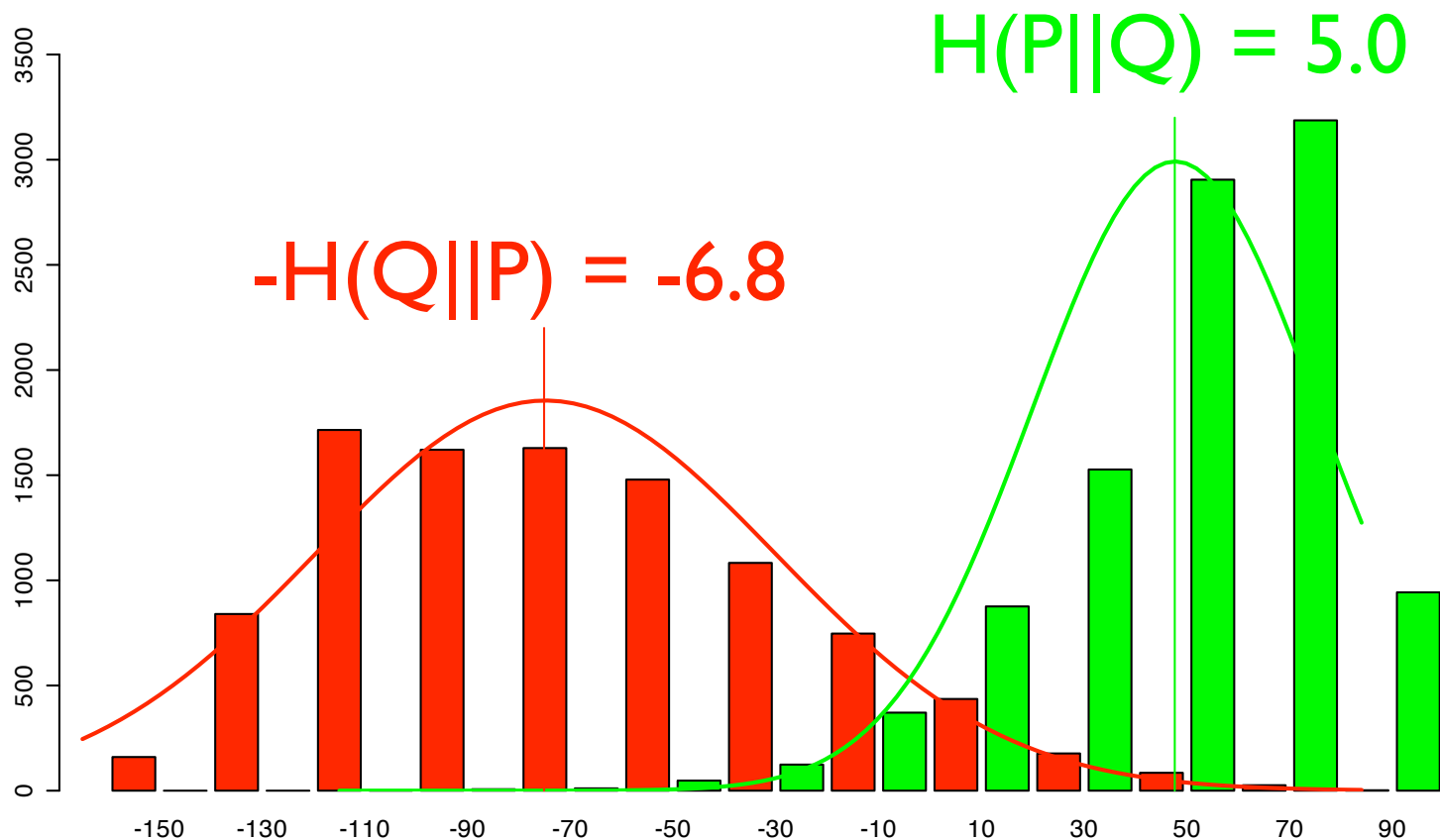


$H(P||Q)$ is *expected log likelihood score* of a sequence randomly chosen from **WMM** (wrt background);

$-H(Q||P)$ 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).
 $2^{11.8} \approx 3566$, which is good, since *many* more non-TATA than TATA

For a WMM:

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

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	$-\infty$	$-\infty$	
C	$-\infty$	$-\infty$	$-\infty$	
G	0	$-\infty$	2	
T	-1	2	$-\infty$	
RelEnt	0.7	2	2	4.7

Non-uniform

LLR	Col 1	Col 2	Col 3	
A	0.74	$-\infty$	$-\infty$	
C	$-\infty$	$-\infty$	$-\infty$	
G	1	$-\infty$	3	
T	-1.58	1.42	$-\infty$	
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 (often 1.0; but needn't be)

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

Motif Discovery

Based on the above, a natural approach to motif discovery, given, say, unaligned upstream sequences of genes thought to be co-regulated, is to find a set of subsequences of *max relative entropy*

```
cgatcTACGATaca...  
tagTAAAATtttc...  
ccgaTATACTcc...  
ggGATAATgagg...  
gactTATGATaa...  
ccTATGTTtgcc...
```

Unfortunately, this 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, \dots, s_k (all of length $n+L-1$, say), each with one instance of an unknown motif

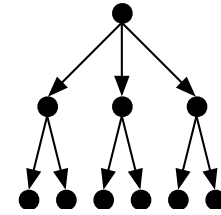
Algorithm:

Build all k -tuples of length L subsequences, one from each of s_1, s_2, \dots, 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, \dots, 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, \dots, s_k (nk sets)

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

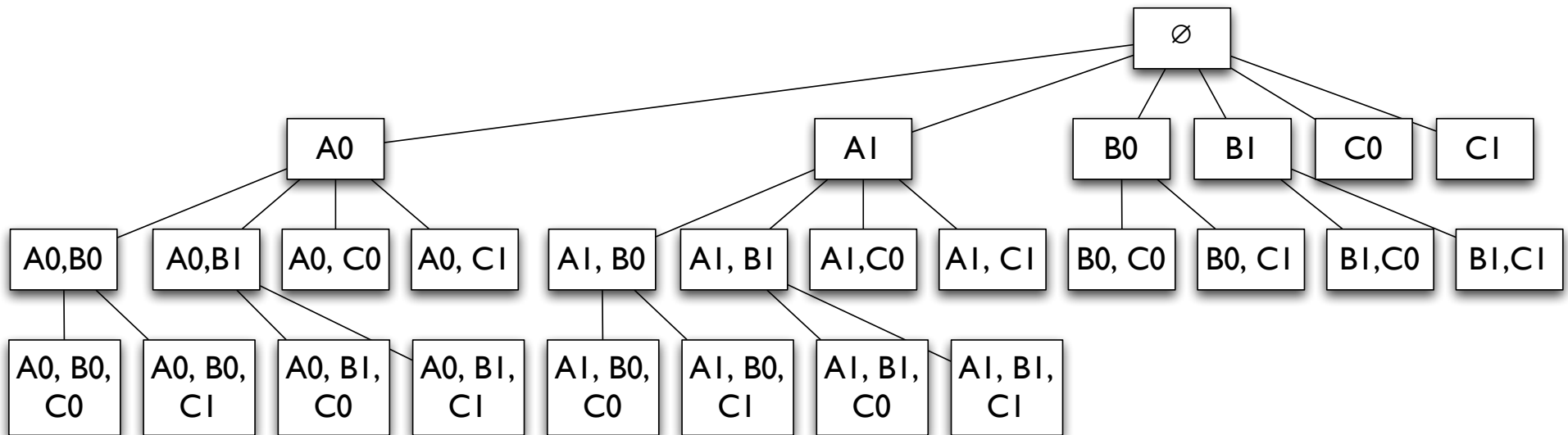
Then triples: len L subseqs of each triple of seqs ($n^3 \binom{k}{3}$ sets)

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

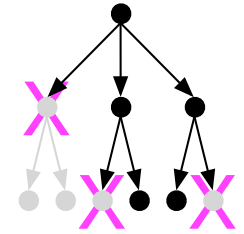
Example



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

Greedy Best-First

[Hertz, Hartzell & Stormo, 1989, 1990]



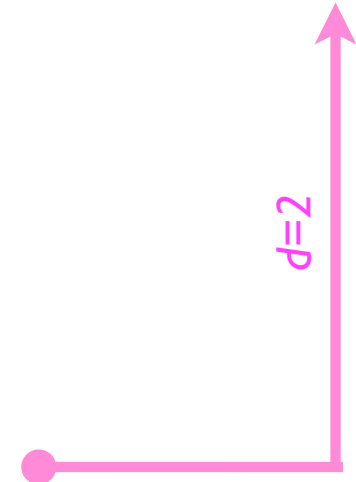
Input:

Sequences s_1, s_2, \dots, 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, \dots, 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 $\theta^{(t)}$ 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 $\theta^{(t+1)}$

Repeat

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

Cartoon Example

xATAyz

CATGACTAGCATAATCCGAT
 TATAATTTCCCAGGGATAACA
 TACAATAGGACCATAGAATGCGC

CATAAT
 CATGAC
 GATAAC
 TATAAT
 CATAGA
 TAGAAT
AATAGG

xATAAz

CATGACTAGCATAATCCGAT
 TATAATTTCCCAGGGATAACA
 TACAATAGGACCATAGAATGCGC

CATAAT
 GATAAC
 TATAAT
 TAGAAT
TACAAT
TAtAAT

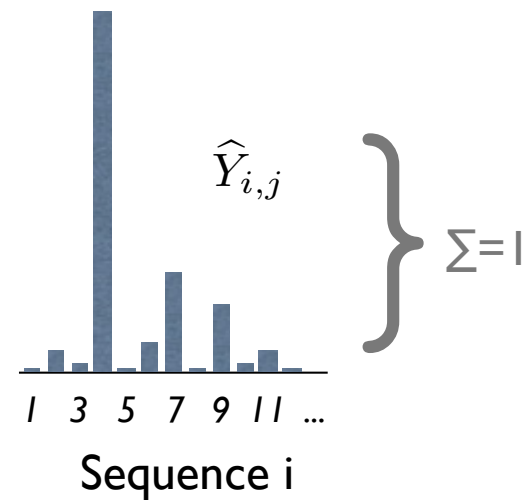
TAtAAT

CATGACTAGCATAATCCGAT
 TATAATTTCCCAGGGATAACA
 TACAATAGGACCATAGAATGCGC

Expectation Step

(where are the motif instances?)

$$\begin{aligned}
 \hat{Y}_{i,j} &= E(Y_{i,j} \mid s_i, \theta^t) \xrightarrow{\text{Bayes}} E = 0 \cdot P(0) + 1 \cdot P(1) \\
 &= P(Y_{i,j} = 1 \mid s_i, \theta^t) \\
 &= P(s_i \mid Y_{i,j} = 1, \theta^t) \frac{P(Y_{i,j}=1|\theta^t)}{P(s_i|\theta^t)} \\
 &= cP(s_i \mid Y_{i,j} = 1, \theta^t) \\
 &= c' \prod_{k=1}^l P(s_{i,j+k-1} \mid \theta^t)
 \end{aligned}$$



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

[Recall slide 35](#)

Maximization Step

(what is the motif?)

Find θ maximizing expected log likelihood:

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

From E-Step

M-Step (cont.)

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

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

s_1 : A**CGG**ATT...

...
 s_k : GC...T**CGG**AC

$\hat{Y}_{1,1}$	ACGG
$\hat{Y}_{1,2}$	CGGA
$\hat{Y}_{1,3}$	GGAT
\vdots	\vdots
$\hat{Y}_{k,l-1}$	CGGA
$\hat{Y}_{k,l}$	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-suite.org>

Another Motif Discovery Approach The Gibbs Sampler

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

Sigma-37	223	IIDLTYIQNK	SQKETGDILGISQMHVSR	LQRKAVKKLR	240	A25944	
SpoIIC	94	RFGLDLKKEK	TQREIAKELGISRSYVSR	IEKRALMKMF	111	A28627	
NahR	22	VVFNQLLVDR	RVSITAENLGLTQPAVSN	ALKRLRTSLQ	39	A32837	
Antennapedia	326	FHFNRYLTRR	RRIEIAHALCLTERQIKI	WFQNRMRKWK	343	A23450	
NtrC (Brady.)	449	LTAALAATRG	NQIRAADLLGLNRNTRK	KIRDLDIQVY	466	B26499	
DicA	22	IRYRRKNLKH	TQRLAKALKISHVSVSQ	WERGDSEPTG	39	B24328	(BVECDA)
MerD	5	MNAY	TVSRLALDAGVSVHIVRD	YLLRGLLRPV	22	C29010	
Fis	73	LDMVMQYTRG	NQTRALMMGINRGTLRK	KLKKYGMN	90	A32142	(DNECF5)
MAT a1	99	FRRKQSLNSK	EKEEVAKKCGITPLQVRV	WFINKRMRK	116	A90983	(JEBY1)
Lambda cII	25	SALLNKIAML	GTEKTAEAVGVDSQISR	WKRDWIPKFS	42	A03579	(QCBP2L)
Crp (CAP)	169	THPDGMQIKI	TRQEIGQIVGCSRETVGR	ILKMLEQNL	186	A03553	(QRECC)
Lambda Cro	15	ITLKDYAMRF	GQTKTAKDLGVYQSAINK	AIHAGRKIFL	32	A03577	(RCBPL)
P22 Cro	12	YKKDVIDHFG	TQRAVAKALGISDAVSVQ	WKÉVIPEKDA	29	A25867	(RGBP22)
AraC	196	ISDHLADSNF	DIASVAQHVCLSPSRLSH	LFRQQLGISV	213	A03554	(RGECA)
Fnr	196	FSPREFRLTM	TRGDIGNYLGLTVETISR	LLGRFQKSGM	213	A03552	(RGECE)
HtpR	252	ARWLDEDNKS	TLQELADRYGVSAERVQR	LEKNAMKKLR	269	A00700	(RGECH)
NtrC (K.a.)	444	LTTALRHHTQG	HKQEAARLLGWGRNTR	KLKELGME	461	A03564	(RGKBCP)
Cytr	11	MKAKKQETAA	TMKDVALKAKVSTATVSR	ALMNPDKVSQ	28	A24963	(RPECCT)
DeoR	23	LQELKRSDKL	HLKDAALLVSEMTIRR	DLNNHSAPVV	40	A24076	(RPECDO)
GalR	3	MA	TIKDVARLAGVSVATVSR	VINNSPKASE	20	A03559	(RPECG)
LacI	5	MKPV	TLYDVAEYAGVSYQTVSR	VVNQASHVSA	22	A03558	(RPECL)
TetR	26	LLNEVGIEGL	TTRKLAQKLGVEQPTLYW	HVKNKRALLD	43	A03576	(RPECTN)
TrpR	67	IVEELLRGEM	SQRELKNELGAGIATITR	GSNSLKAAPV	84	A03568	(RPECW)
NifA	495	LIAALEKAGW	VQAKAARLLGMTPRQVAY	RIQIMDITMP	512	S02513	
SpoIIG	205	RFGLVGEEEK	TQKDVADMMGISQSYISR	LEKRIIKRLR	222	S07337	
Pin	160	QAGRLIAAGT	PRQKVAIIYDVGSTLYK	TFPAGDK	177	S07958	
PurR	3	MA	TIKDVAKRANVSTTVSH	VINKTRFVAE	20	S08477	
EbgR	3	MA	TLKDIAIEAGVSLATVSR	VLNDDPTLNV	20	S09205	
LexA	27	DHISQTMPP	TRAEIAQRLGFRSPNAAE	EHLKALARKG	44	S11945	
P22 cI	25	SSILNRIAIR	GQRKVAADALGINESQISR	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

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

How to Average

An old problem:

n random variables:

$$x_1, x_2, \dots, x_k$$

Joint distribution (p.d.f.):

$$P(x_1, x_2, \dots, x_k)$$

Some function:

$$f(x_1, x_2, \dots, x_k)$$

Want Expected Value:

$$E(f(x_1, x_2, \dots, x_k))$$

How to Average

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

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 $\vec{x}^{(1)}, \vec{x}^{(2)}, \dots, \vec{x}^{(n)} \sim P(\vec{x})$ 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 $\vec{X}_{t+1} \sim P(\vec{X}_{t+1} | \vec{X}_t)$ w/ stationary dist = P

- Simplest & most common: Gibbs Sampling

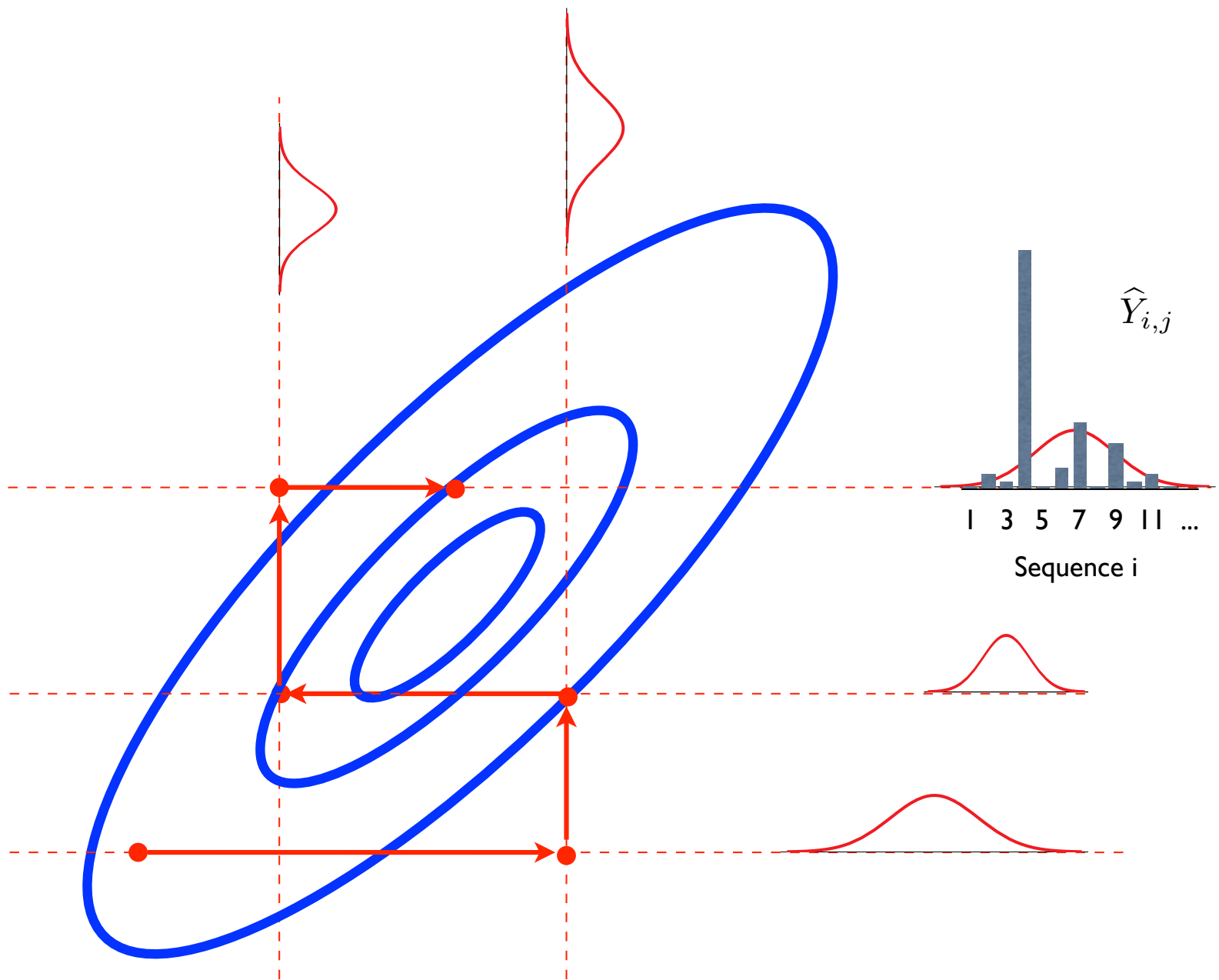
$$P(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_k)$$

- Algorithm

for $t = 1$ to ∞

for $i = 1$ to k do :

$$x_{t+1,i} \sim P(x_{t+1,i} | x_{t+1,1}, x_{t+1,2}, \dots, x_{t+1,i-1}, x_{t,i+1}, \dots, x_{t,k})$$



Input: again assume sequences s_1, s_2, \dots, 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

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

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 $j = 1 \dots |s_i| - w + 1$

calculate prob that i^{th} motif is at j :

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

pick new x_i according to that distribution

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

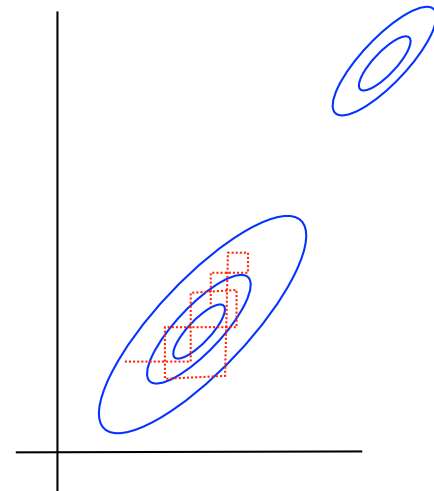
Issues

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

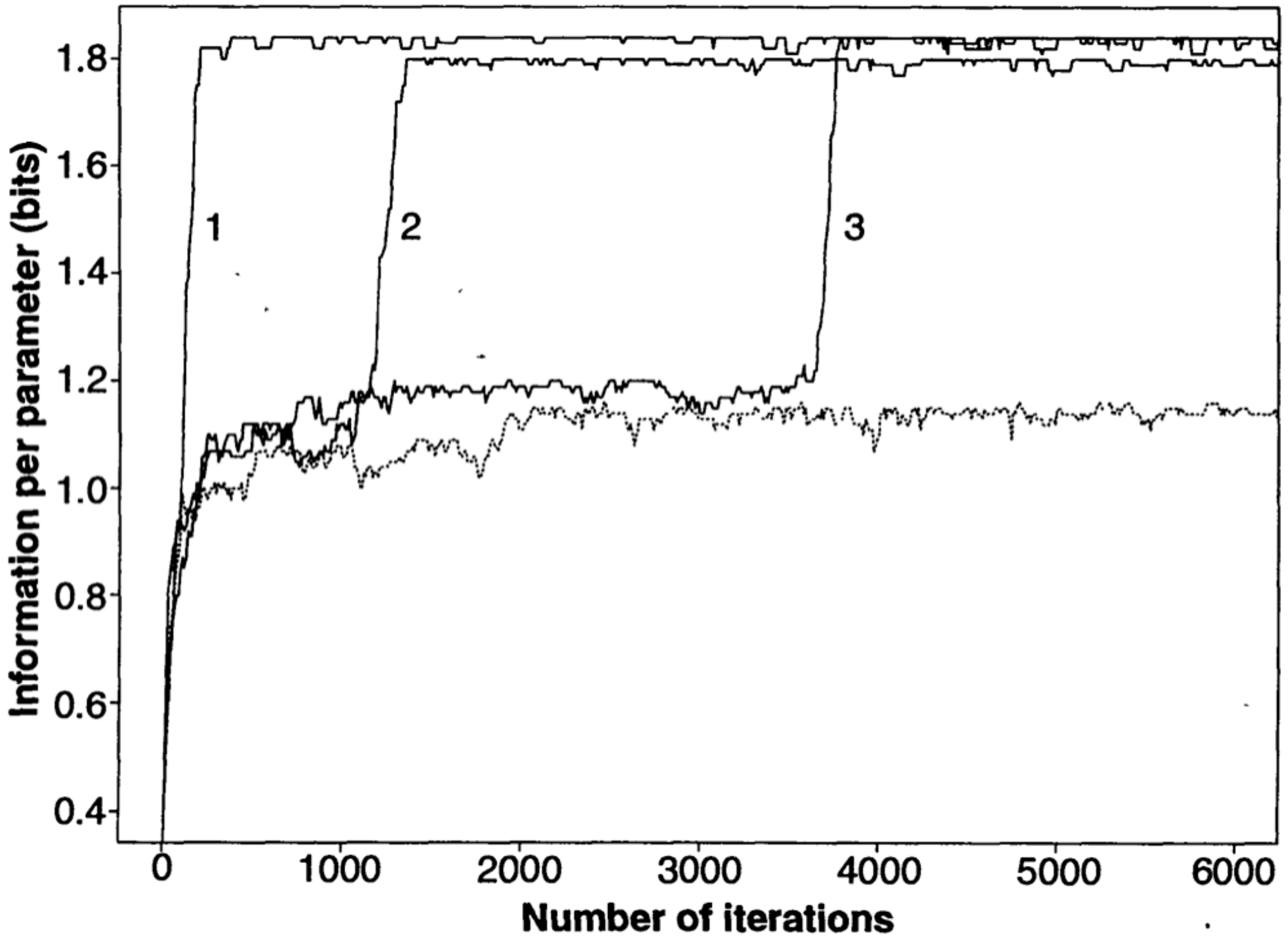


Variants & Extensions

“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



Assessing computational tools for the discovery of transcription factor binding sites

Martin Tompa^{1,2}, Nan Li¹, Timothy L Bailey³, George M Church⁴, Bart De Moor⁵, Eleazar Eskin⁶, Alexander V Favorov^{7,8}, Martin C Frith⁹, Yutao Fu⁹, W James Kent¹⁰, Vsevolod J Makeev^{7,8}, Andrei A Mironov^{7,11}, William Stafford Noble^{1,2}, Giulio Pavesi¹², Graziano Pesole¹³, Mireille Régnier¹⁴, Nicolas Simonis¹⁵, Saurabh Sinha¹⁶, Gert Thijs⁵, Jacques van Helden¹⁵, Mathias Vandenbogaert¹⁴, Zhiping Weng⁹, Christopher Workman¹⁷, Chun Ye¹⁸ & Zhou Zhu⁴

Methodology

13 tools

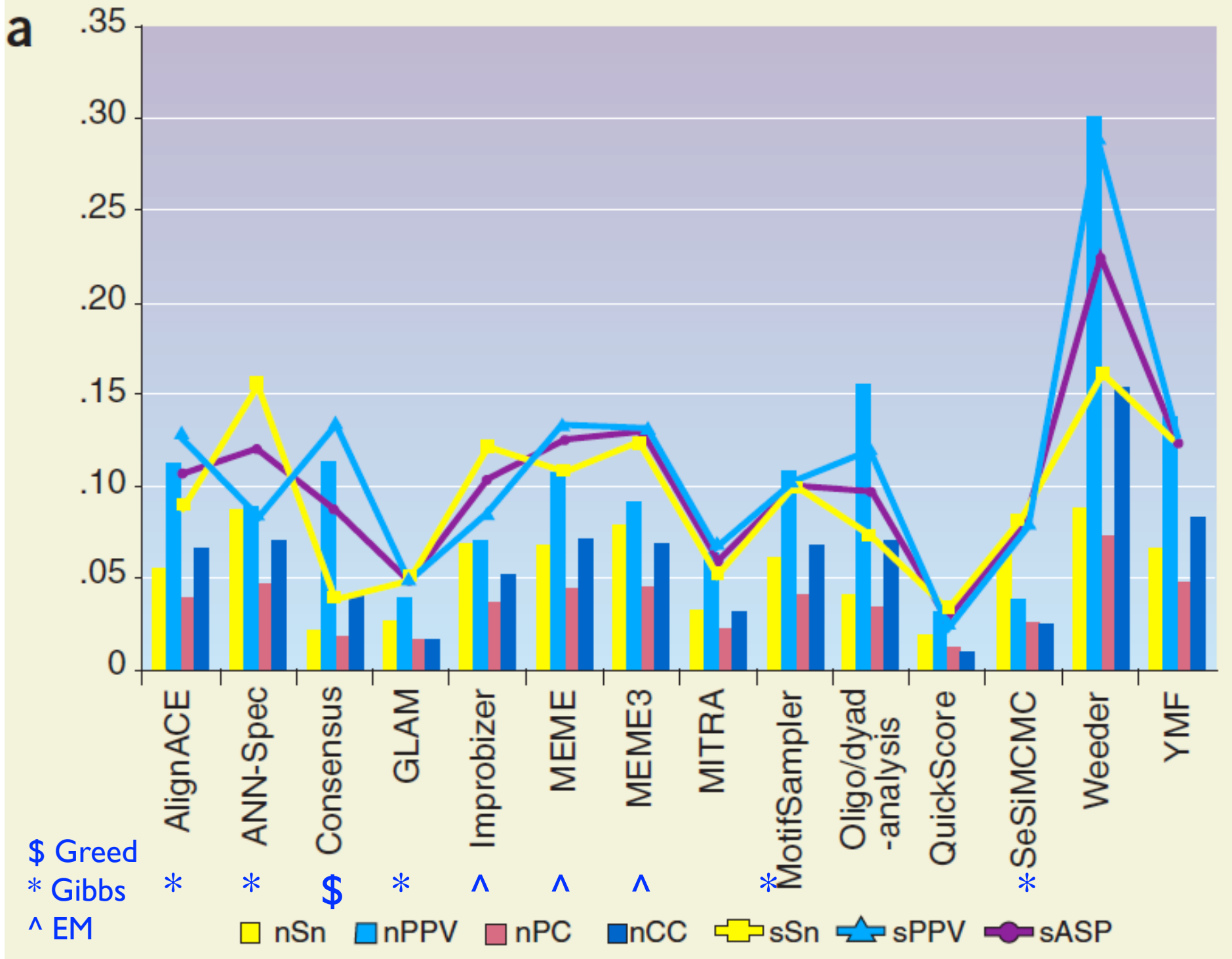
Real 'motifs' (Transfac)

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

'Real', 'generic', 'Markov'

Expert users, top prediction only

"Blind" – sort of



Notation

- nTP is the number of nucleotide positions in both known sites and predicted sites,
- nFN is the number of nucleotide positions in known sites but not in predicted sites,
- nFP is the number of nucleotide positions not in known sites but in predicted sites, and
- nTN is the number of nucleotide positions in neither known sites nor predicted sites.
- sTP be the number of known sites overlapped by predicted sites,
- sFN be the number of known sites not overlapped by predicted sites, and
- sFP be the number of predicted sites not overlapped by known sites.

At either the nucleotide ($x = n$) or site ($x = s$) level, one can then define:

- *Sensitivity*: $xSn = xTP/(xTP + xFN)$, and
- *Positive Predictive Value*: $xPPV = xTP/(xTP + xFP)$.

Specificity: $nSP = nTN/(nTN + nFP)$.

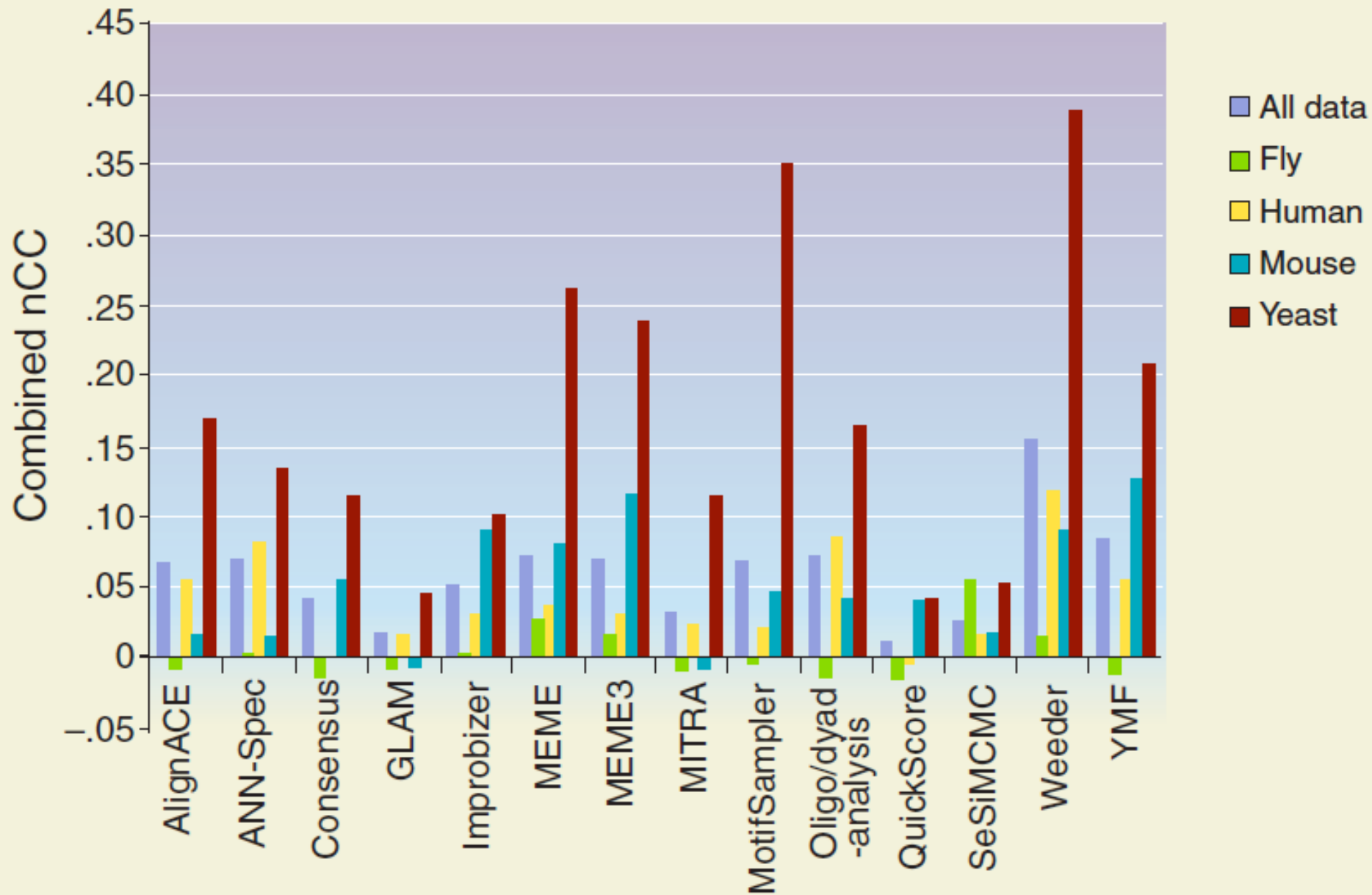
Finally, it is enlightening to consider various single statistics that in some sense average (some of) these quantities. Following Pevzner & Sze¹, define the (nucleotide level) performance coefficient as:

$$nCC = \frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}$$

- $nPC = nTP/(nTP + nFN + nFP)$.

$sASP = (sSn + sPPV)/2$.

The correlation coefficient nCC is the Pearson product-moment coefficient of correlation in the particular case of two binary variables, also called the 'phi coefficient of correlation.' The two binary variables are the characteristic vectors of the known nucleotide positions and

b

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

Bioinformatics Advance Access published November 26, 2013

BIOINFORMATICS ORIGINAL PAPER

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

Sequence analysis

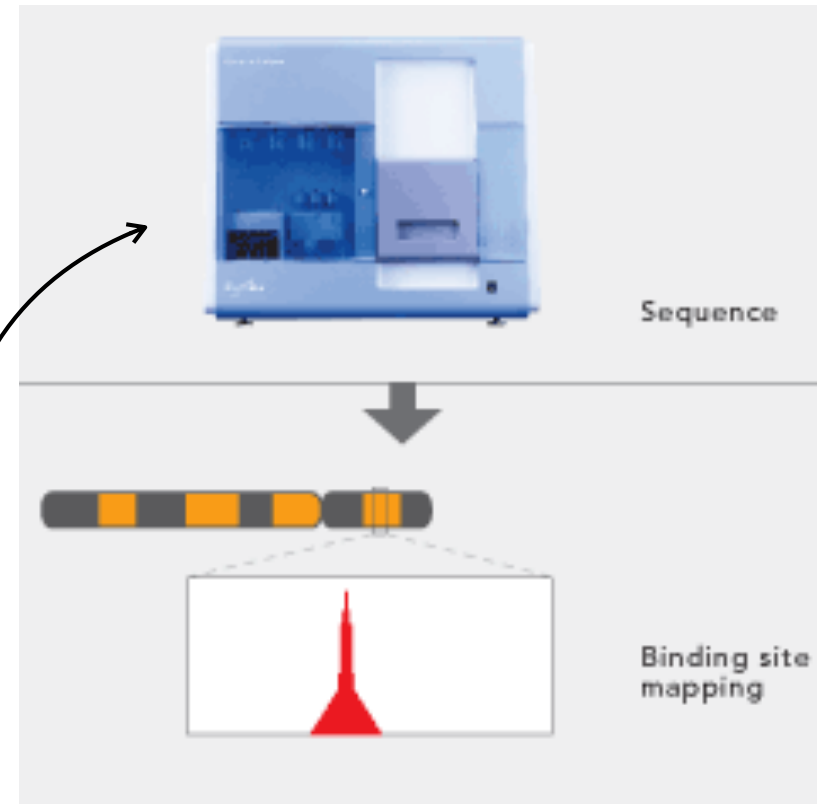
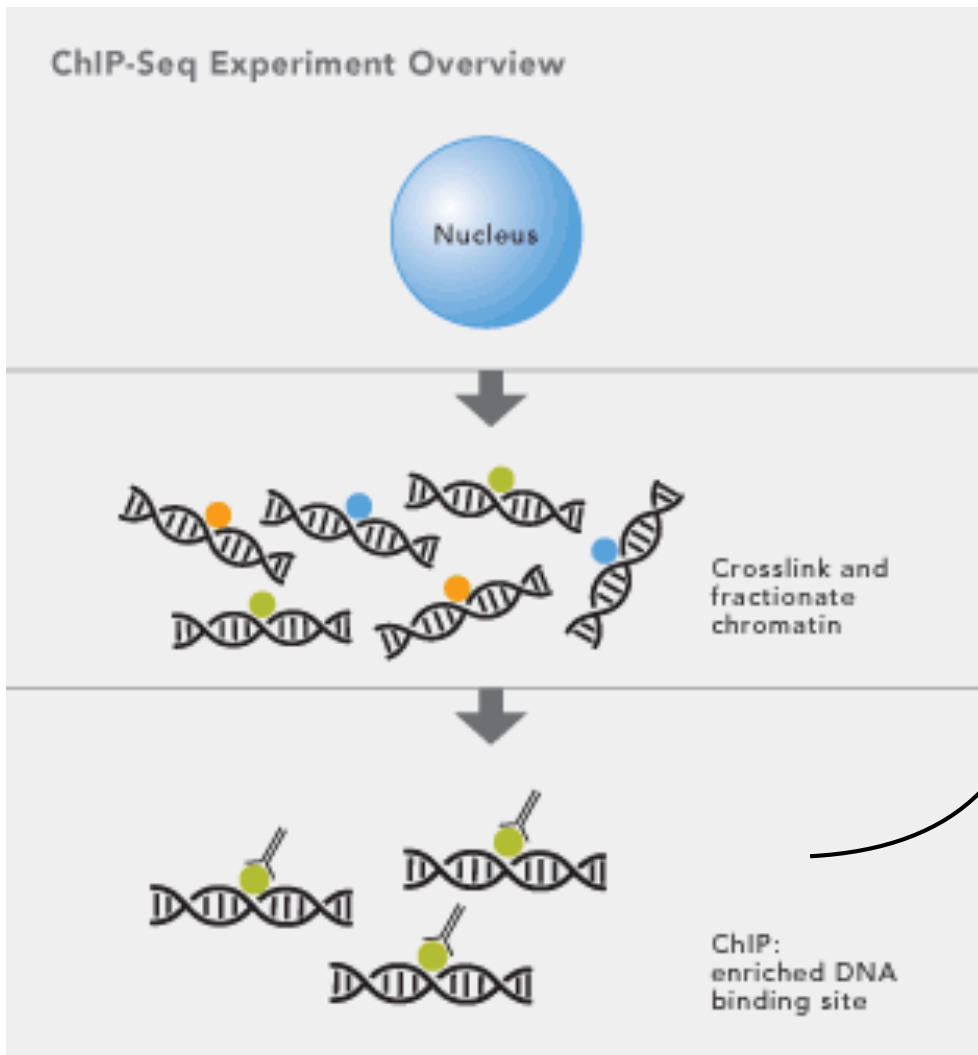
Advance Access publication October 25, 2013

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. Gentleman⁹

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

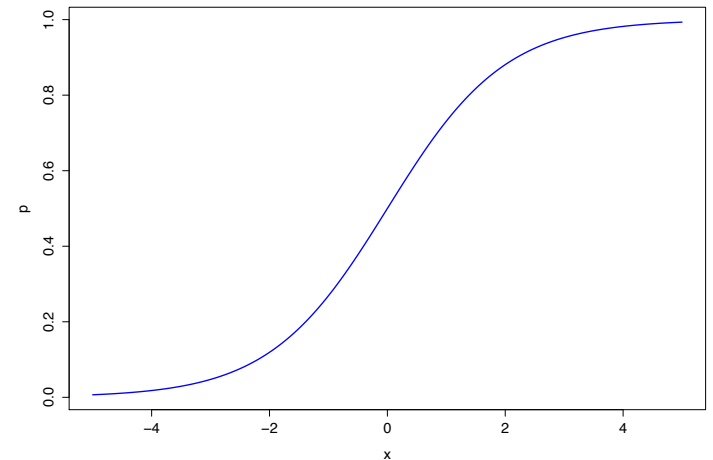
Logistic regression – x = motif count

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

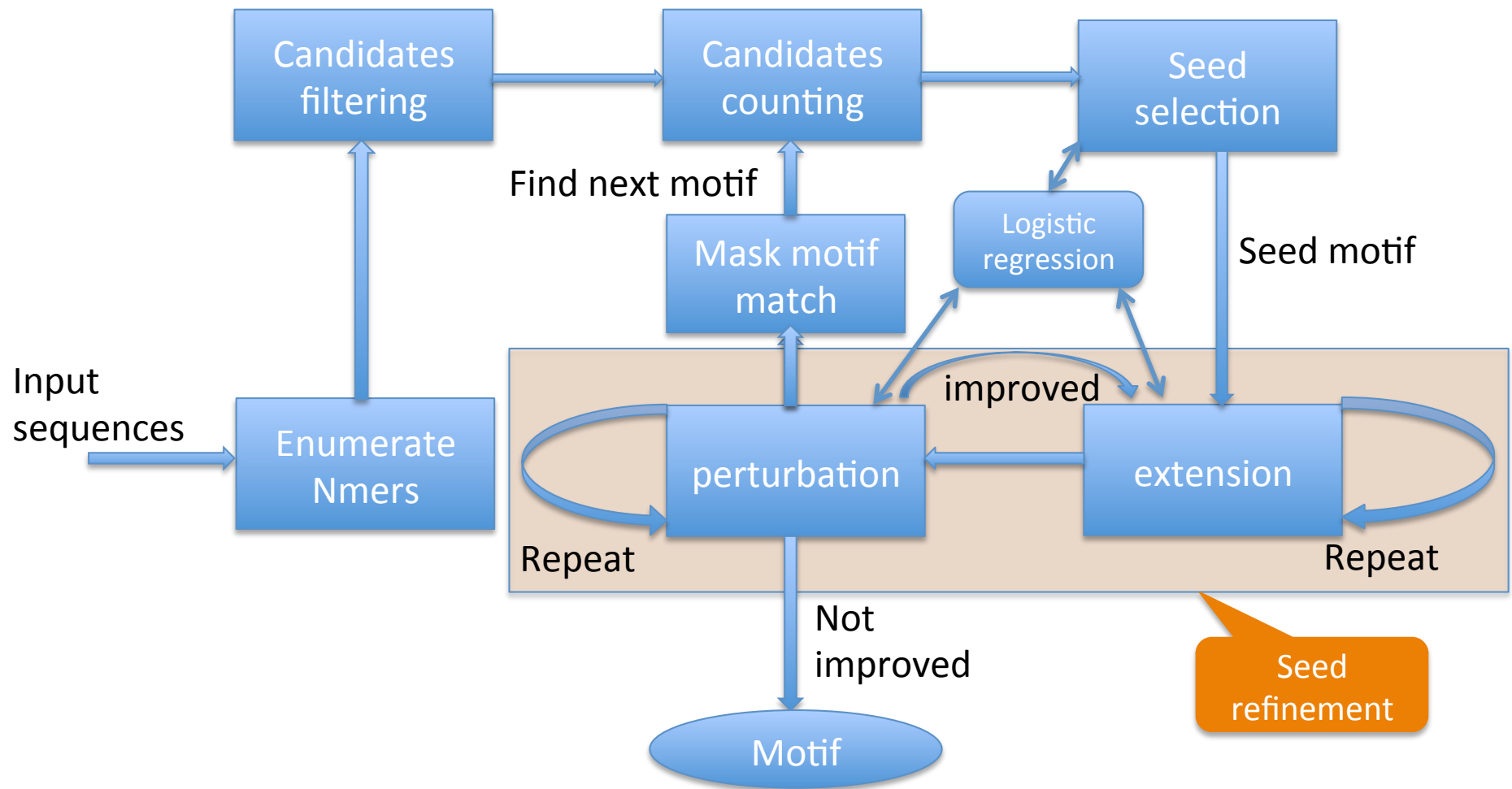
IUPAC patterns – e.g., “R” = A or G

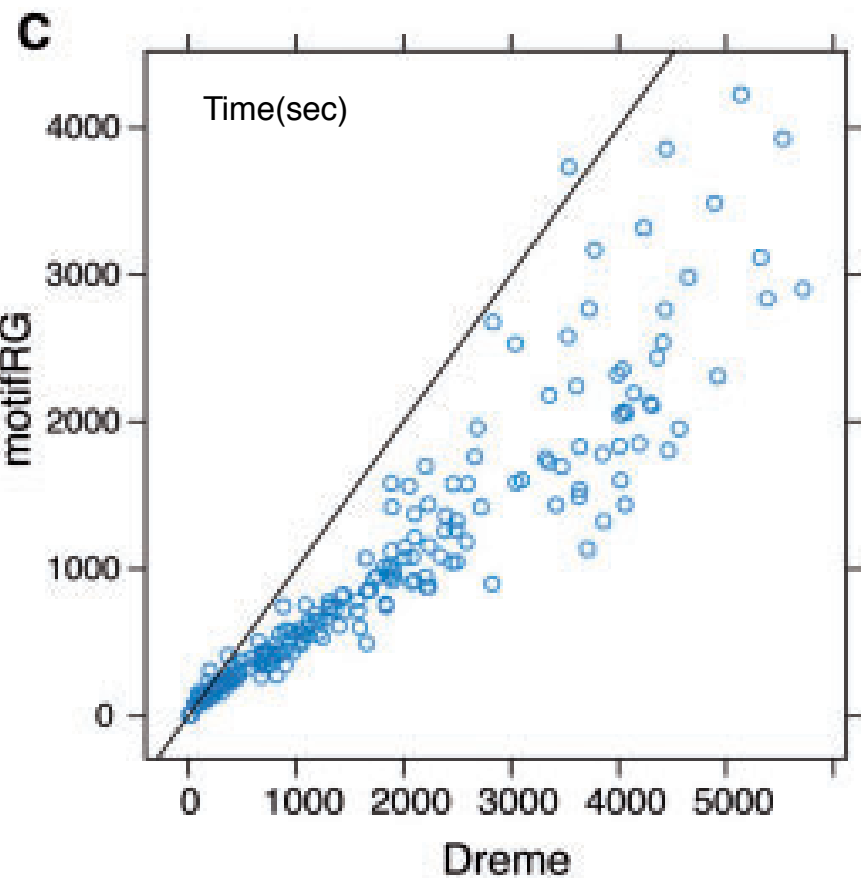
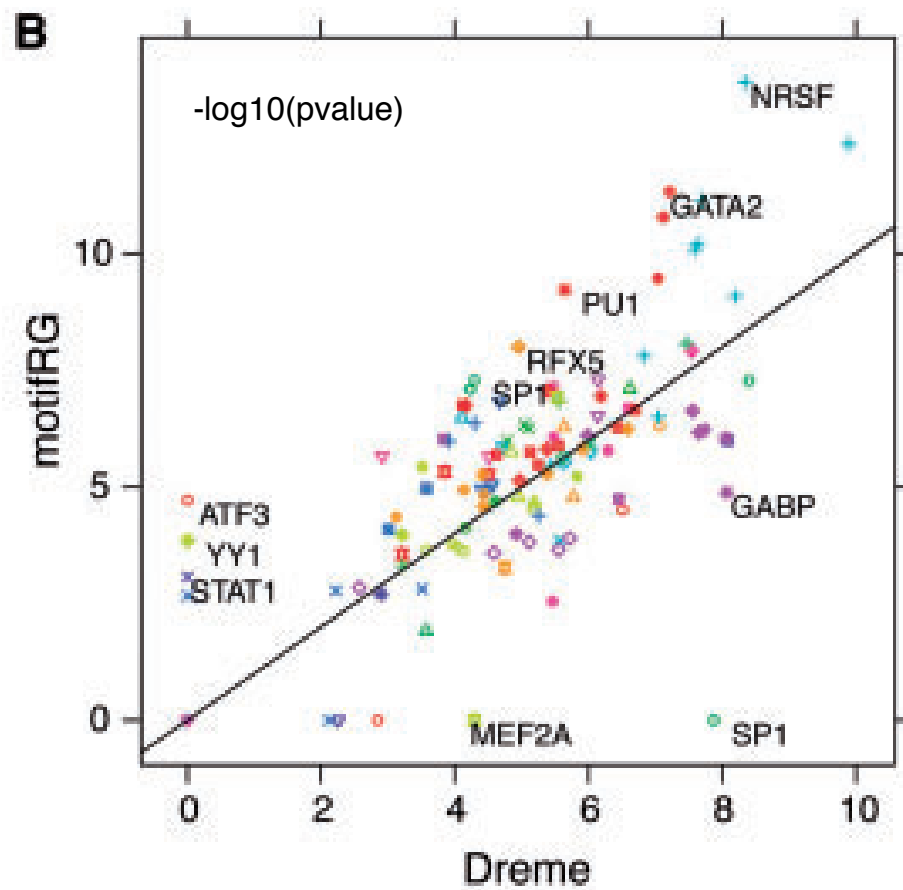
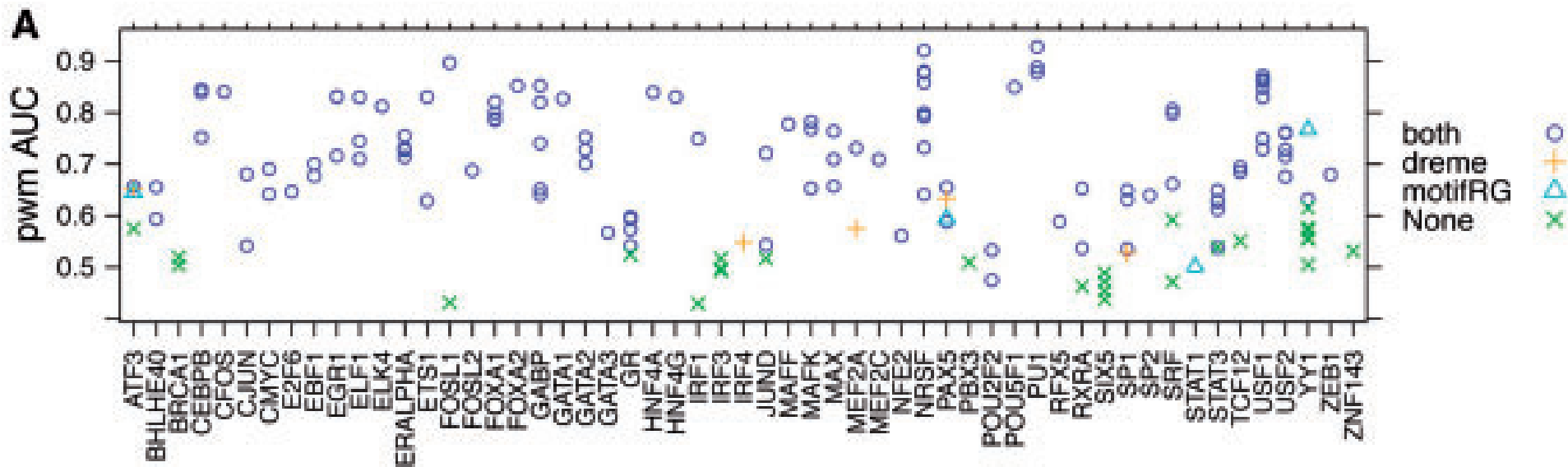
seed/extend/perturb

Z-scores

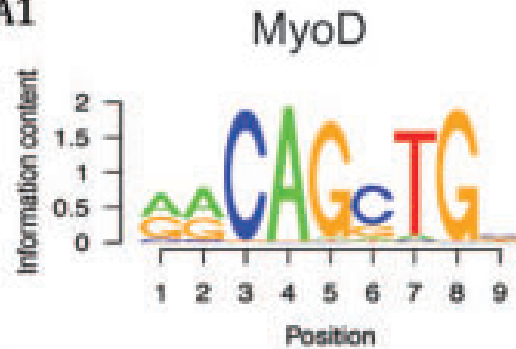


Method Outline





A1

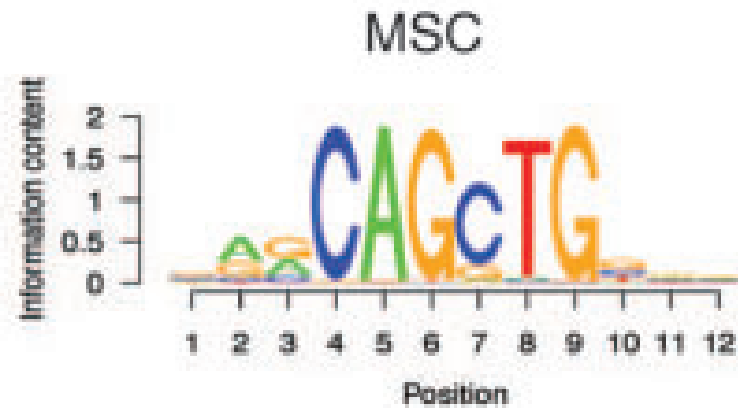
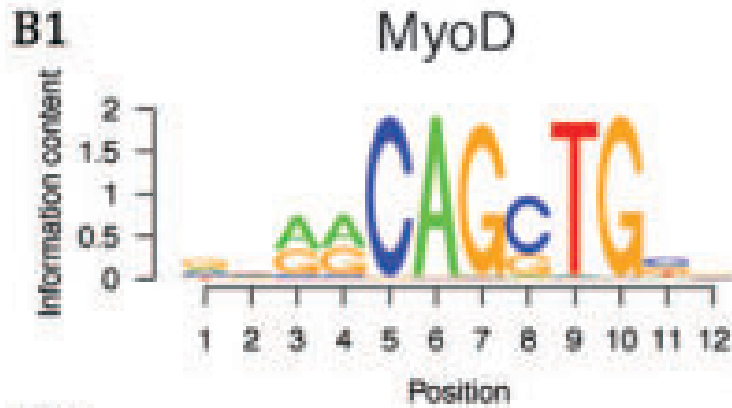


NeuroD



A2

Consensus	scores	ratio	fg.frac	bg.frac	logo	DB match	DB Evaluate	DB logo	
NNVCAGATGGNN	69	5.12	0.68	0.14		Tcfe2a	1.5e-09		} NeuroD
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		} MyoD
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					

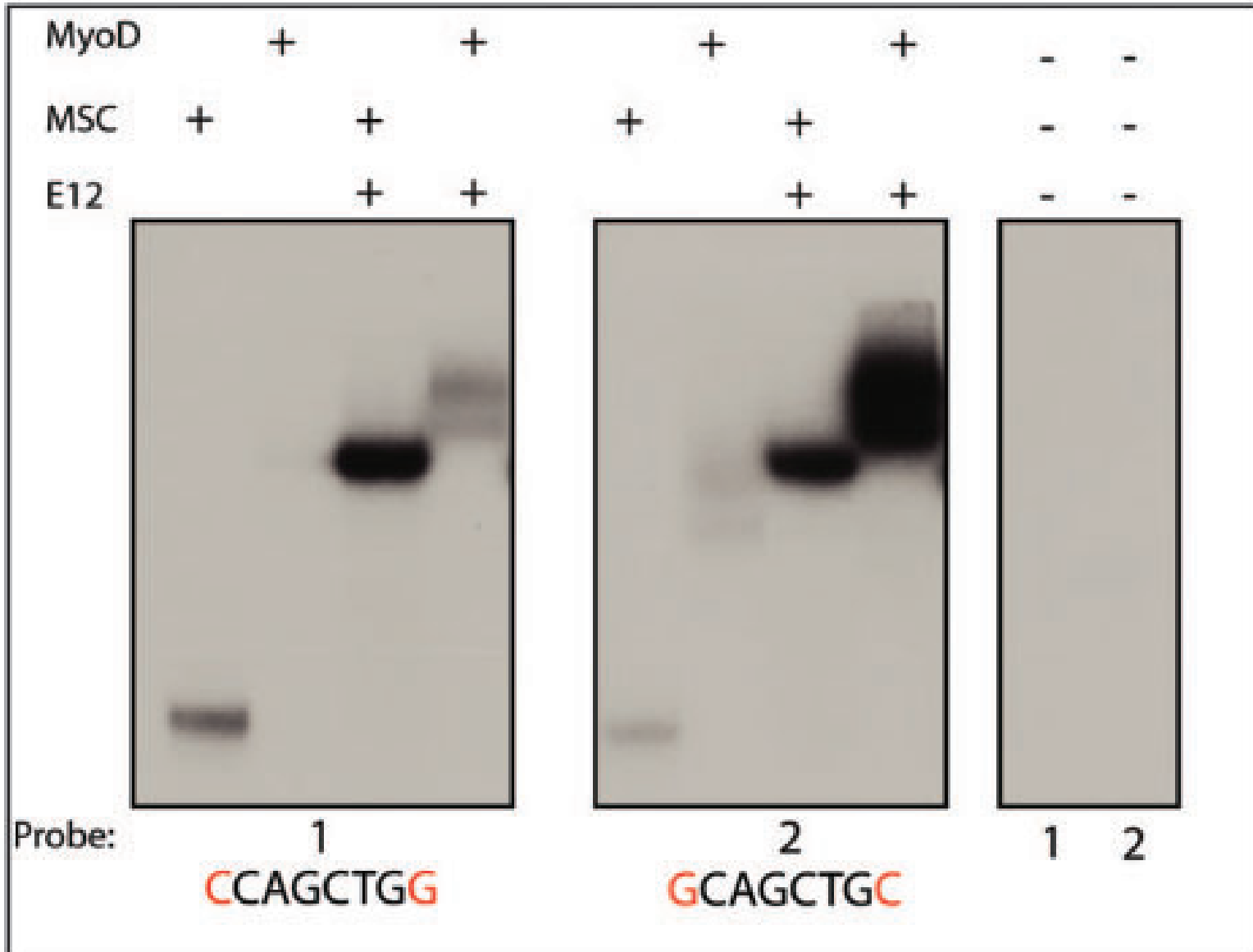


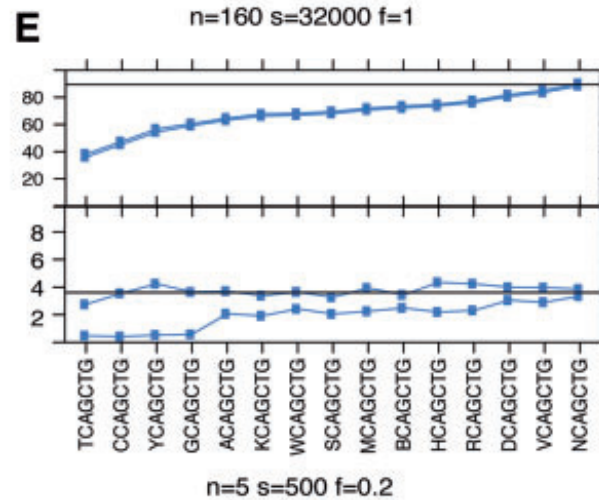
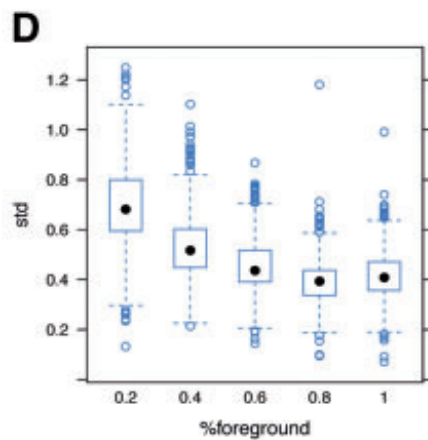
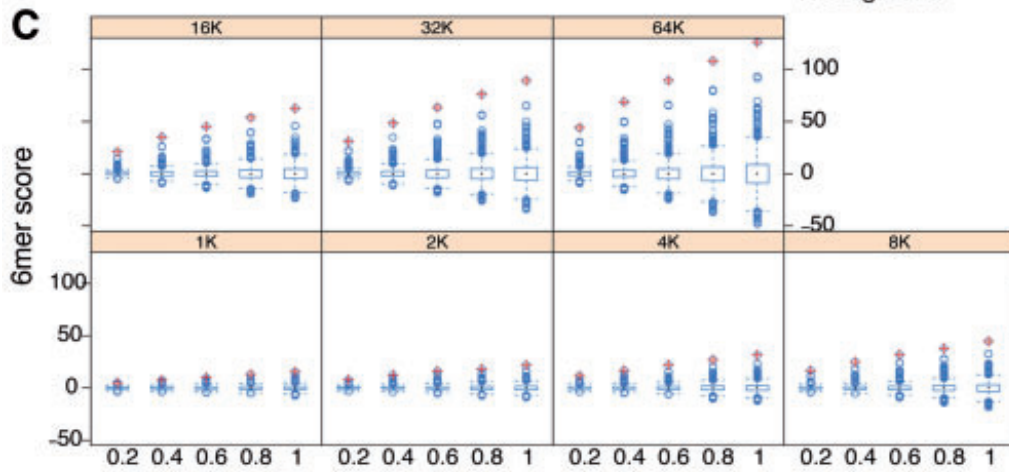
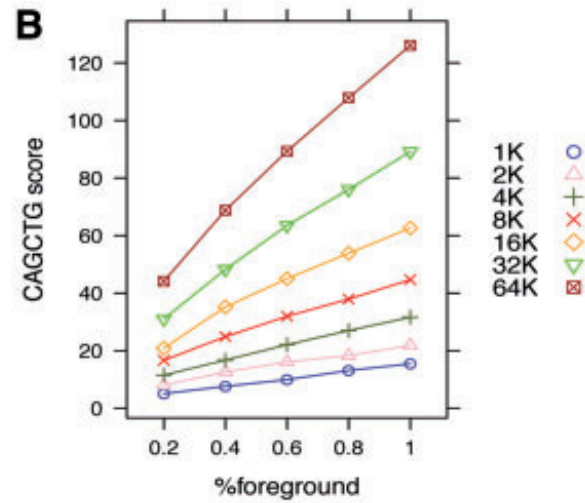
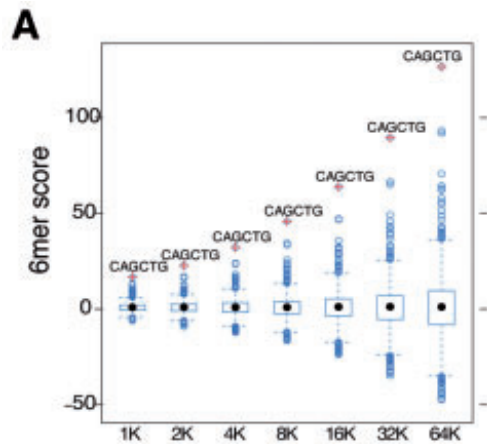
B2

Consensus	scores	ratio	fg.frac	bg.frac	logo
NRRCAGGTGNN	20.5	4.90	0.521	0.12	
NNCAGCTGGNN	-25.6	0.21	0.211	0.76	
NNBCCAGCHN	-20.1	0.43	0.488	0.83	

MyoD

MSC





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:

Weight matrix models for representation & search

Greedy, MEME and Gibbs for discovery

Still room for improvement. E.g., *ChIP-seq* and *Comparative genomics* (cross-species comparison) are very promising.