CSE 527

Autumn 2009

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)

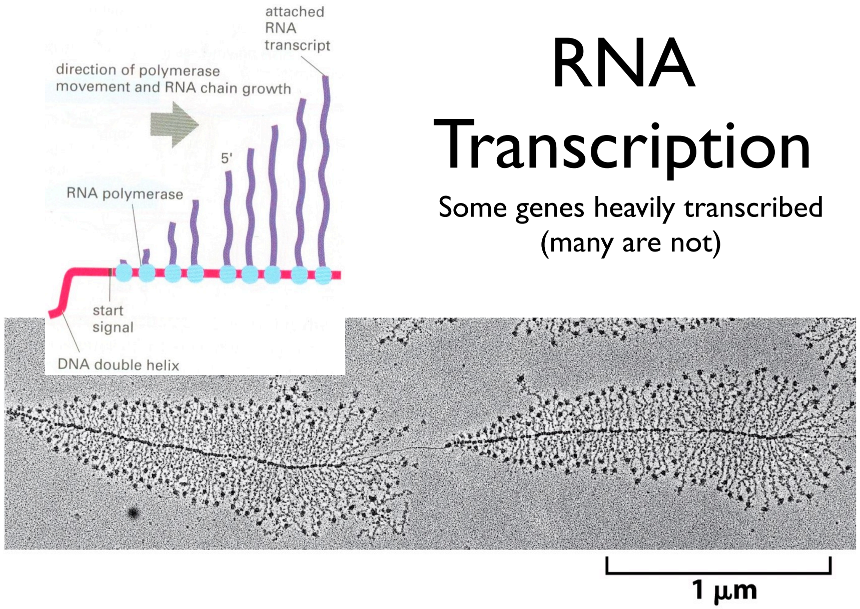


Figure 6-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

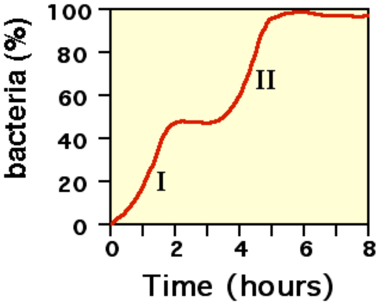
Regulation

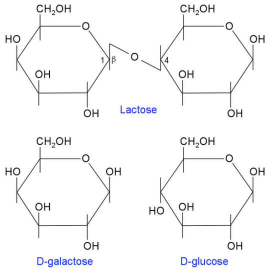
In most cells, pro- or eukaryote, easily a 10,000-fold difference between least- and most-highly expressed genes

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, ...

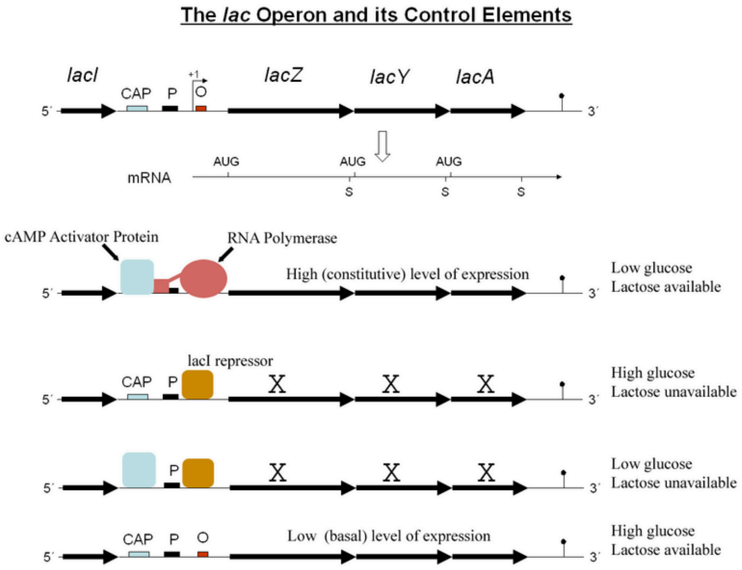
Below, focus on 1st step only:
transcriptional regulation

E. coli growth
on glucose + lactose





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

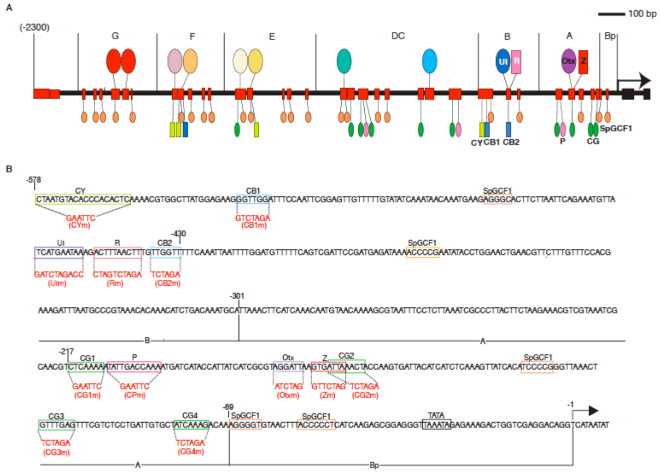


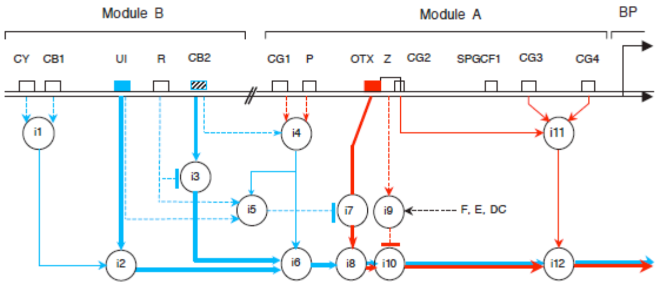
1965 Nobel Prize

Physiology or Medicine

François Jacob, Jacques Monod, André Lwoff

Sea Urchin - Endo16





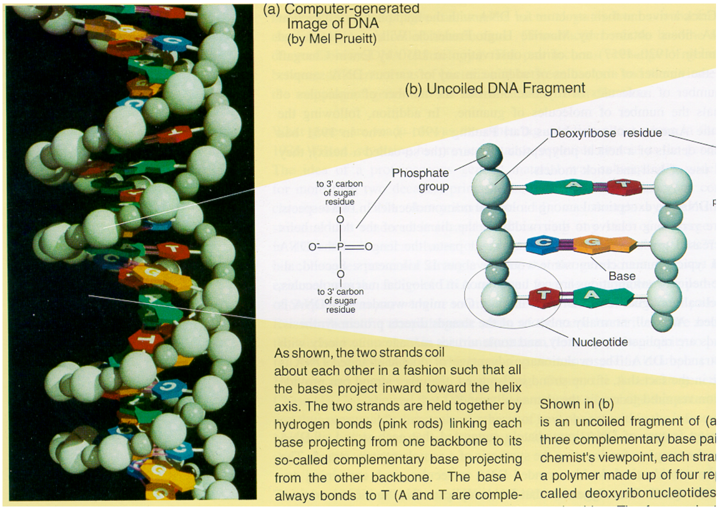
```
if CY & CB1                         i1 = 1
else                                i1 = 0.5

                                    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
```

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
Deoxyribose residue
Phosphate group
to 3' carbon of sugar residue
O
O⁻—P═O
O
to 3' carbon of sugar residue
A
T
C
G
Base
T
A
Nucleotide

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

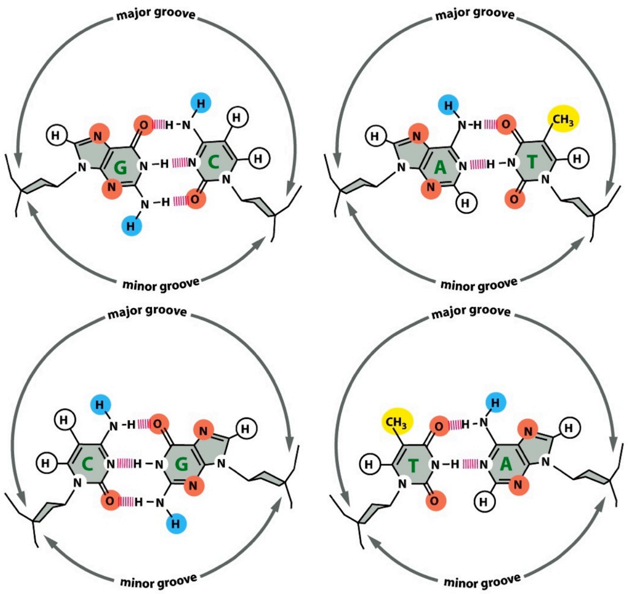


Figure 7-7 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Helix-Turn-Helix DNA Binding Motif

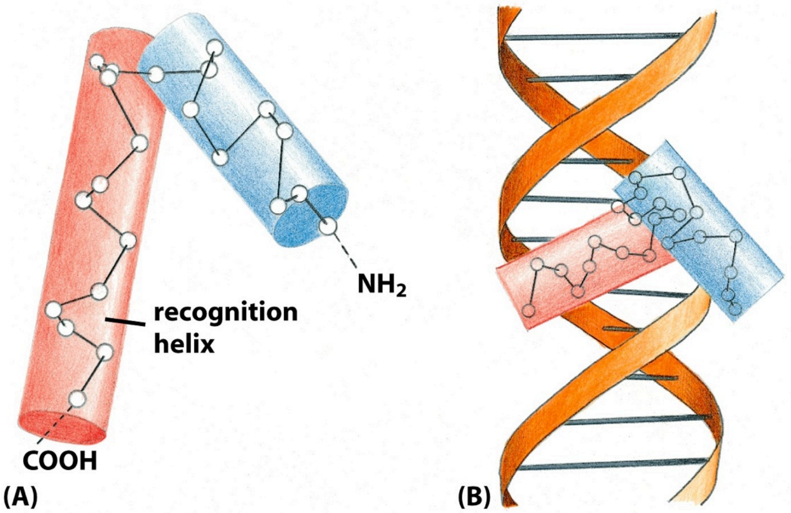


Figure 7-10 Molecular Biology of the Cell 5/e (© Garland Science 2008)

H-T-H Dimers

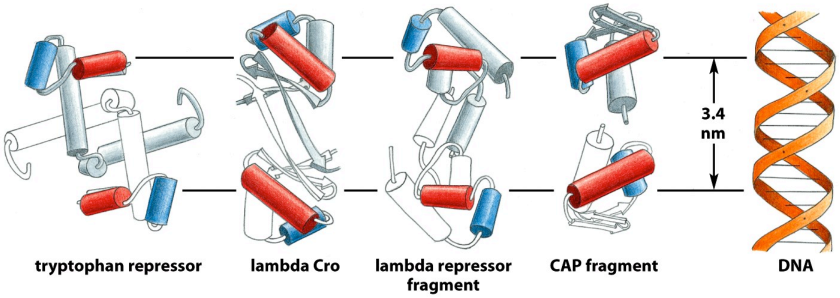
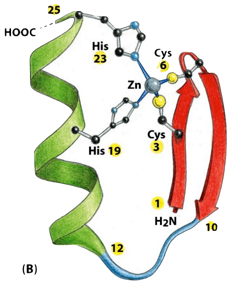


Figure 7-11 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Bind 2 DNA patches, ~ 1 turn apart

Increases both specificity and affinity

Zinc Finger Motif



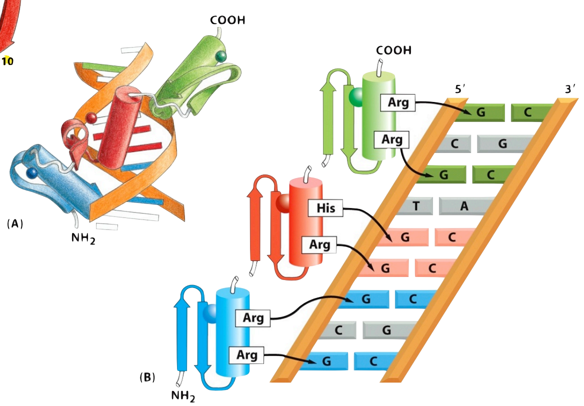
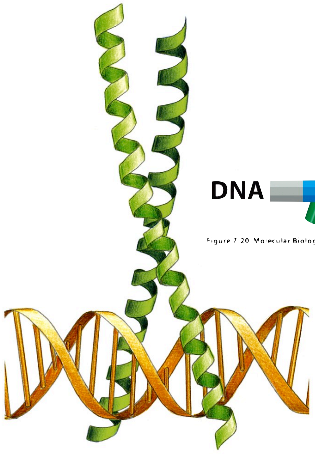


Figure 7-15 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Leucine Zipper Motif



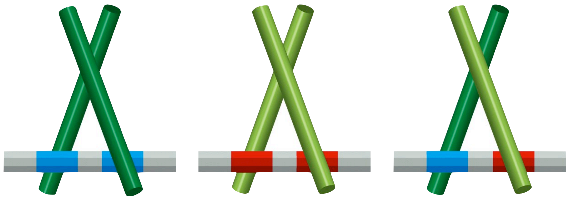
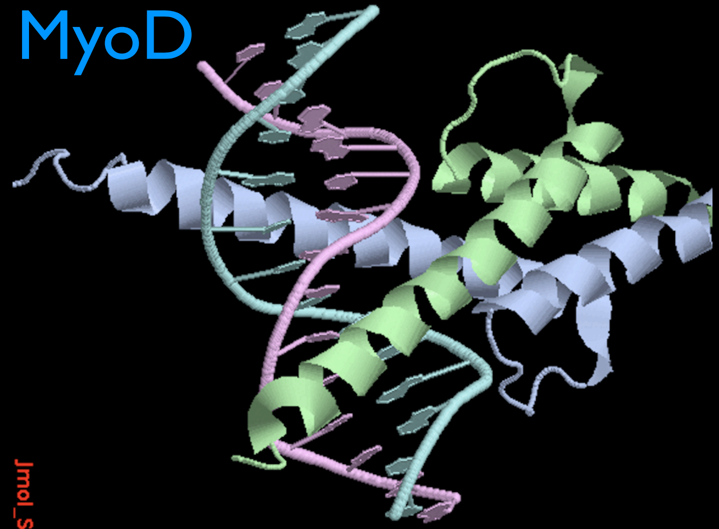


Figure 7-20 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Homo-/hetero-dimers
and combinatorial
control

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

MyoD



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

Some Protein/DNA interactions well-understood

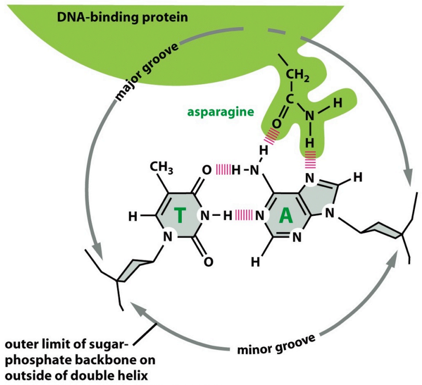


Figure 7-9 Molecular Biology of the Cell 5/e (© Garland Science 2008)

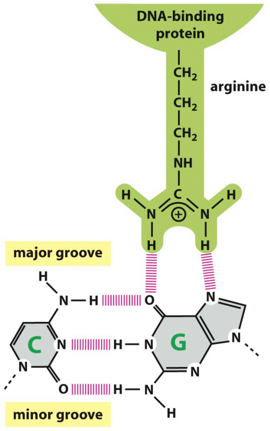


Figure 7-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)

But the overall DNA binding “code” still defies prediction

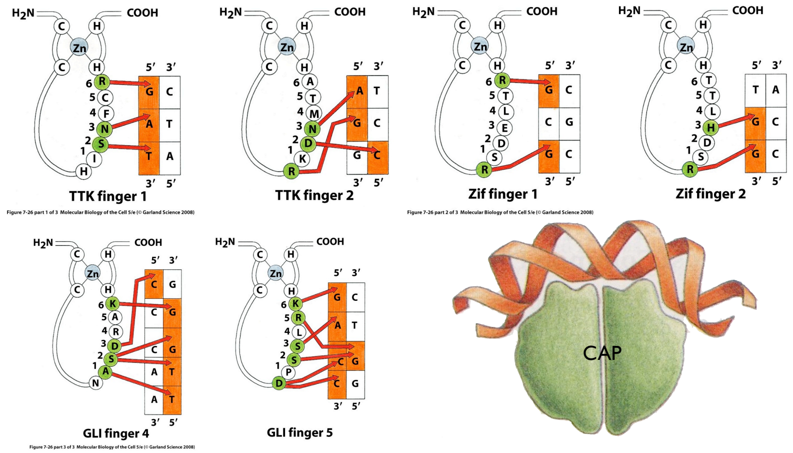


Figure 7-26 part 1 of 3 Molecular Biology of the Cell 5/e (© Garland Science 2008)

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)

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 DNA *sequence* motifs 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

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

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

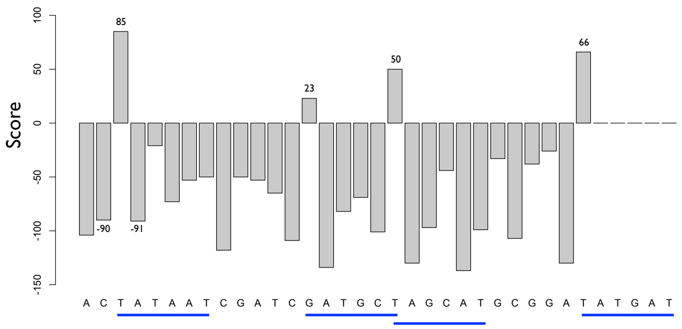
A	-36	19	1	12	10	-46
C	-15	-36	-8	-9	-3	-31
G	-13	-46	-6	-7	-9	-46
T	17	-31	8	-9	-6	19

A C T A T A A T C G

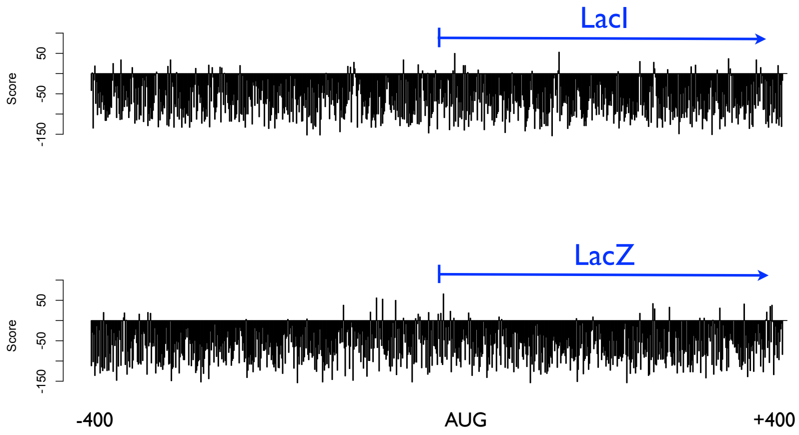
= -91

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

Scanning for TATA

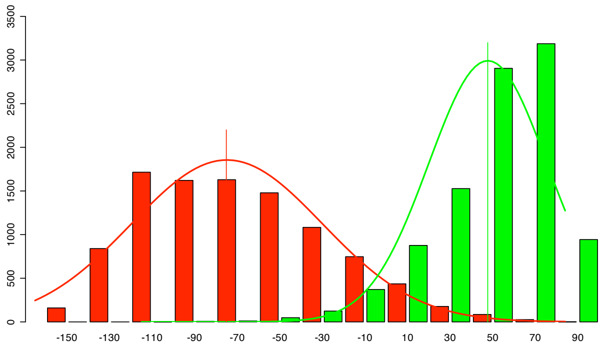


TATA Scan at 2 genes



Score Distribution

(Simulated)



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

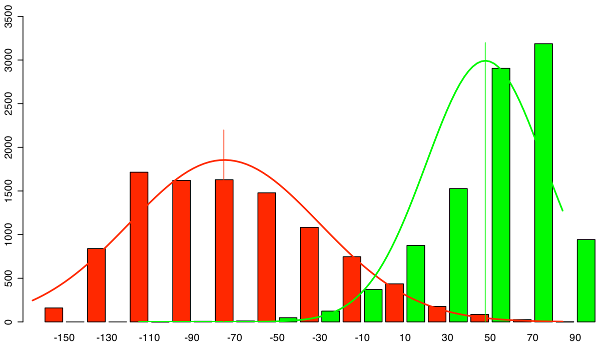
Might as well look at *likelihood ratio:*

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

(or *log likelihood ratio*)

Score Distribution

(Simulated)



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

Log-Likelihood Ratio:

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

Freq.	Col 1	Col 2	Col 3
A	0.625	0	0
C	0	0	0
G	0.250	0	1
T	0.125	1	0

LLR	Col 1	Col 2	Col 3
A	1.32	-∞	-∞
C	-∞	-∞	-∞
G	0	-∞	2.00
T	-1.00	2.00	-∞

Non-uniform Background

- *E. coli* - DNA approximately 25% A, C, G, T
- *M. jannaschi* - 68% A-T, 32% G-C

LLR from previous example, assuming

$$f_A = f_T = 3/8$$
$$f_C = f_G = 1/8$$

LLR	Col 1	Col 2	Col 3
A	0.74	-∞	-∞
C	-∞	-∞	-∞
G	1.00	-∞	3.00
T	-1.58	1.42	-∞

e.g., G in col 3 is 8 x more likely via WMM than background, so ($\log_2$) score = 3 (bits).

Relative Entropy

AKA Kullback-Liebler Distance/Divergence, AKA Information Content

Given distributions P, Q

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

Notes:

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?

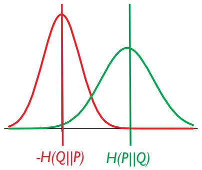
For any fixed length sequence *x*, let

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

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

Relative Entropy:

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

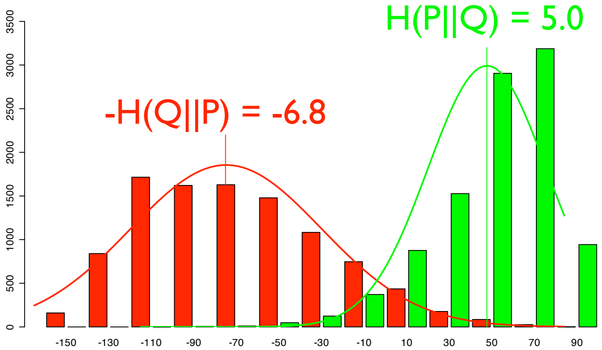


H(P||Q) is *expected log likelihood score* of a sequence randomly chosen from *WMM*;

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

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	0.250	0	1
T	0.125	1	0

Uniform

LLR	Col 1	Col 2	Col 3	
A	1.32	-∞	-∞	
C	-∞	-∞	-∞	
G	0	-∞	2.00	
T	-1.00	2.00	-∞	
RelEnt	0.70	2.00	2.00	4.70

Non-uniform

LLR	Col 1	Col 2	Col 3	
A	0.74	-∞	-∞	
C	-∞	-∞	-∞	
G	1.00	-∞	3.00	
T	-1.58	1.42	-∞	
RelEnt	0.51	1.42	3.00	4.93

Pseudocounts

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

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 $s_1, s_2, ..., s_k$ (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 $s_1, s_2, ..., s_k$ (nk sets)

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

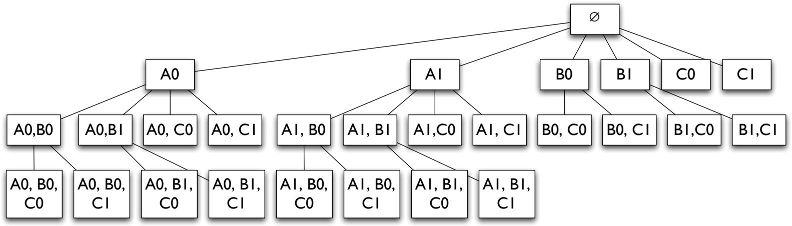
Then triples: len L subseqs of each triple of seqs ($n^3\binom{k}{3}$ sets)

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

Compute relative entropy of each; pick best

problem:
astronomically sloooow

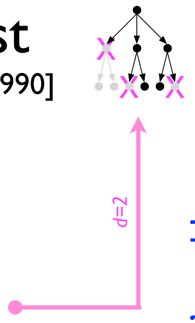
Example



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

Greedy Best-First

[Hertz, Hartzell & Stormo, 1989, 1990]



Input:

  Sequences $s_1, s_2, \ldots, s_k$; 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):

  Sequence $s_1, s_2, \ldots, s_k$; 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

Expectation Step

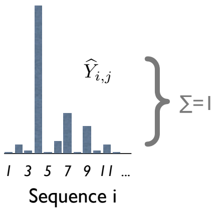
(where are the motif instances?)

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

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

Bayes

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



Maximization Step

(what is the motif?)

Find θ maximizing expected value:

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

M-Step (cont.)

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

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...

...

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

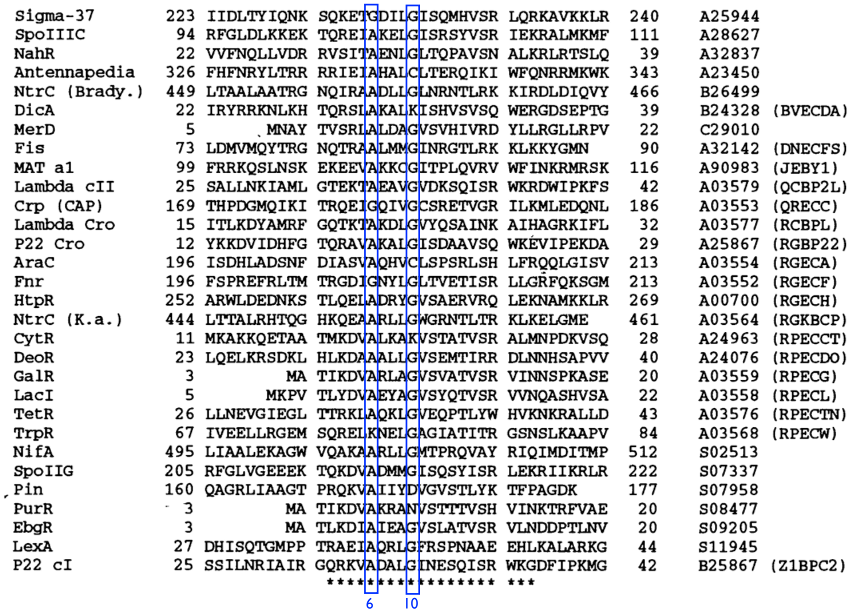
(Having a supercomputer helps):

http://meme.sdsc.edu/

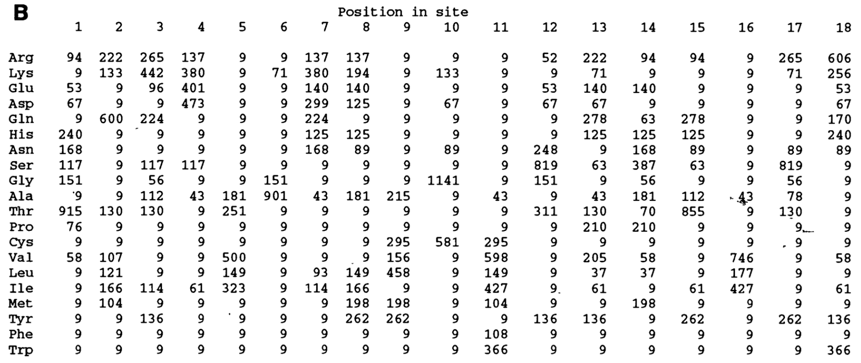
Another Motif Discovery Approach

The Gibbs Sampler

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



```
Sigma-37        223 IIDLTYIQNK SQKETGDILGISQMHVSR LQRKAVKKLR  240    A25944
SpoIIIC          94 RFGLDLKKEK TQREIAKELGISRSYVSR IEKRALMKMF  111    A28627
NahR             22 VVFNQLLVDR RVSITAENLGLTQPAVSN ALKRLRTSLQ   39    A32837
Antennapedia    326 FHFNRYLTRR RRIEIAHALCLTERQIKI WFQNRRMKWK  343    A23450
NtrC (Brady.)   449 LTAALAATRG NQIRAADLLGLNRNTLRK KIRDLDIQVY  466    B26499
DicA             22 IRYRRKNLKH TQRSLAKALKISHVSVSQ WERGDSEPTG   39    B24328 (BVECDA)
MerD              5       MNAY TVSRLALDAGVSVHIVRD YLLRGLLRPV   22    C29010
Fis              73 LDMVMQYTRG NQTRAALMMGINRGTLRK KLKKYGMN     90    A32142 (DNECFS)
MAT a1           99 FRRKQSLNSK EKEEVAKKCGITPLQVRV WFINKRMRSK  116    A90983 (JEBY1)
Lambda cII       25 SALLNKIAML GTEKTAEAVGVDKSQISR WKRDWIPKFS   42    A03579 (QCBP2L)
Crp (CAP)       169 THPDGMQIKI TRQEIGQIVGCSRETVGR ILKMLEDQNL  186    A03553 (QRECC)
Lambda Cro       15 ITLKDYAMRF GQTKTAKDLGVYQSAINK AIHAGRKIFL   32    A03577 (RCBPL)
P22 Cro          12 YKKDVIDHFG TQRAVAKALGISDAAVSQ WKÉVIPEKDA   29    A25867 (RGBP22)
AraC            196 ISDHLADSNF DIASVAQHVCLSPSRLSH LFRQQLGISV  213    A03554 (RGECA)
Fnr             196 FSPREFRLTM TRGDIGNYLGLTVETISR LLGRFQKSGM  213    A03552 (RGECF)
HtpR            252 ARWLDEDNKS TLQELADRYGVSAERVRQ LEKNAMKKLR  269    A00700 (RGECH)
NtrC (K.a.)     444 LTTALRHTQG HKQEAARLLGWGRNTLTR KLKELGME    461    A03564 (RGKBCP)
CytR             11 MKAKKQETAA TMKDVALKAKVSTATVSR ALMNPDKVSQ   28    A24963 (RPECCT)
DeoR             23 LQELKRSDKL HLKDAAALLGVSEMTIRR 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 PRQKVAIIYDVGVSTLYK TFPAGDK     177    S07958
PurR              3         MA TIKDVAKRANVSTTTVSH VINKTRFVAE   20    S08477
EbgR              3         MA TLKDIAIEAGVSLATVSR VLNDDPTLNV   20    S09205
LexA             27 DHISQTGMPP TRAEIAQRLGFRSPNAAE EHLKALARKG   44    S11945
P22 cI           25 SSILNRIAIR GQRKVADALGINESQISR 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

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:	$x_1, x_2, \ldots, x_k$
Joint distribution (p.d.f.):	$P(x_1, x_2, \ldots, x_k)$
Some function:	$f(x_1, x_2, \ldots, x_k)$
Want Expected Value:	$E(f(x_1, x_2, \ldots, x_k))$

How to Average

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

Approach 1: direct integration
(rarely solvable analytically, esp. in high dim)

Approach 2: numerical integration
(often difficult, e.g., unstable, esp. in high dim)

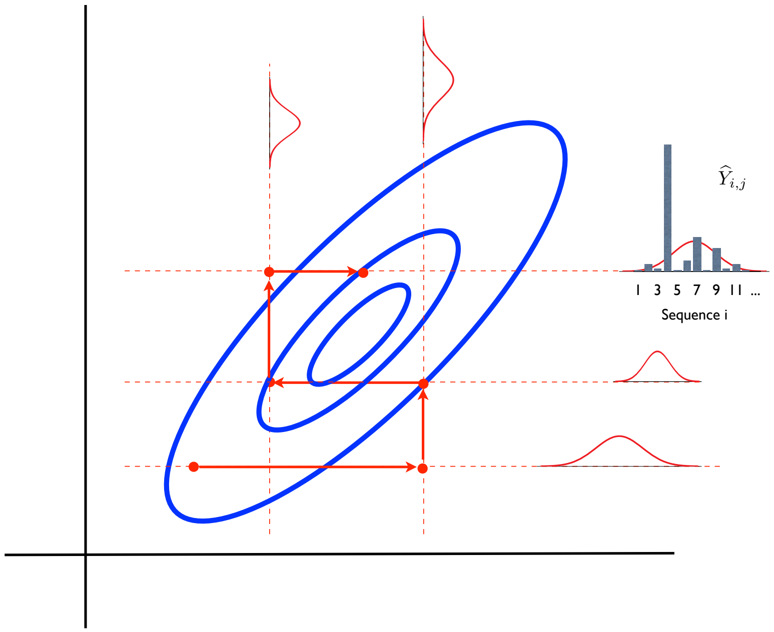
Approach 3: Monte Carlo integration
sample $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots \vec{x}^{(n)} \sim P(\vec{x})$ and average:

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

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



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

Motif model: WMM

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

"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
```

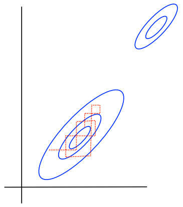
Similar to MEME, but it would *average* over, rather than *sample* from

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

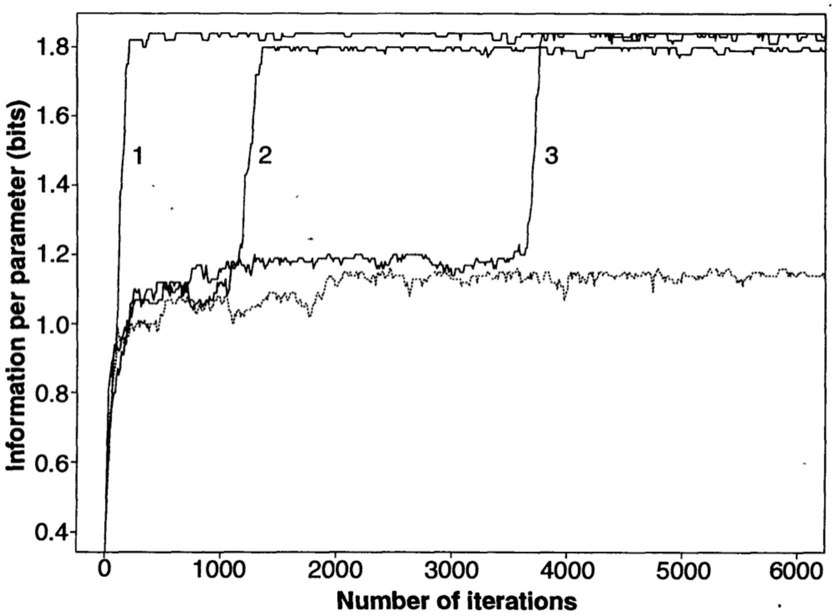


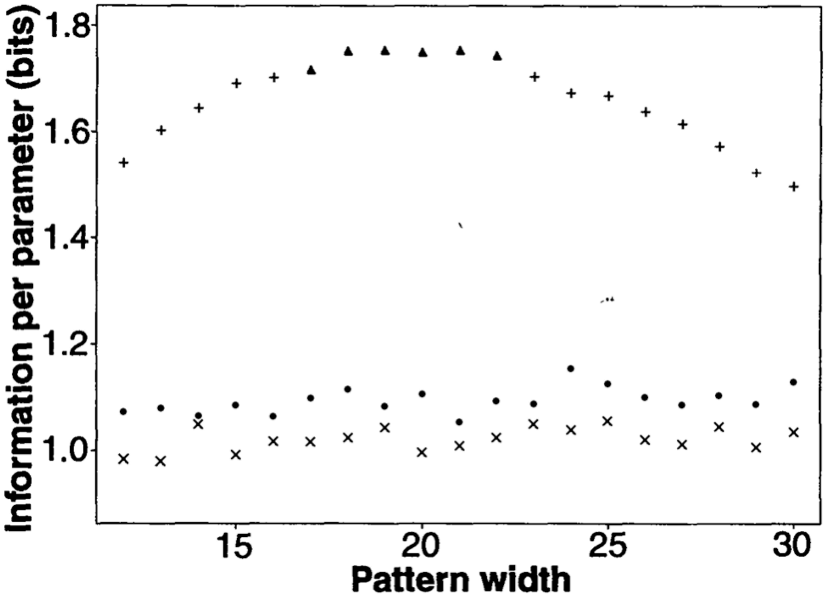
Variants & Extensions

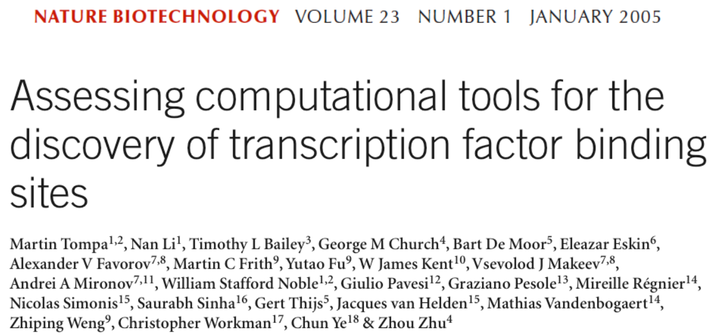
"Phase Shift" - may settle on suboptimal solution that overlaps part of motif. Periodically try moving all motif instances a few spaces left or right.

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

Multiple patterns per string






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

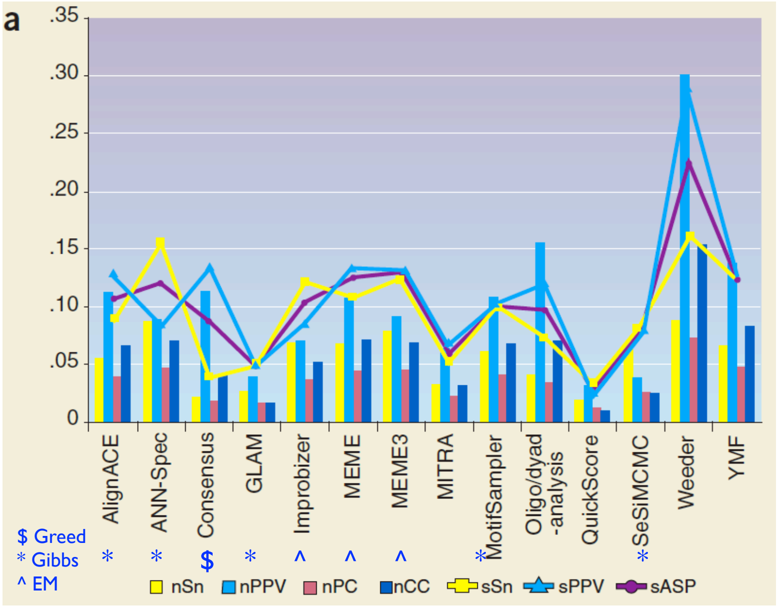
Real 'motifs' (Transfac)

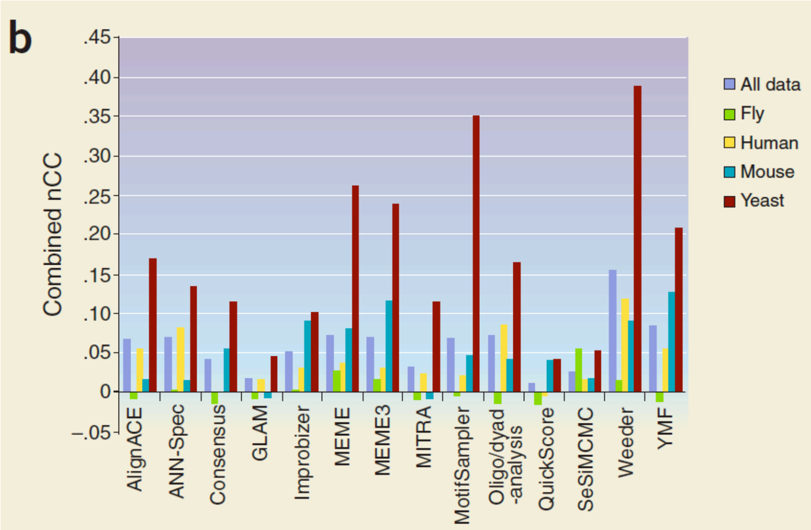
56 data sets (human, mouse, fly, yeast)

'Real', 'generic', 'Markov'

Expert users, top prediction only

"Blind" – sort of





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:

Weight matrix models for representation & search

Greedy, MEME and Gibbs for discovery

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