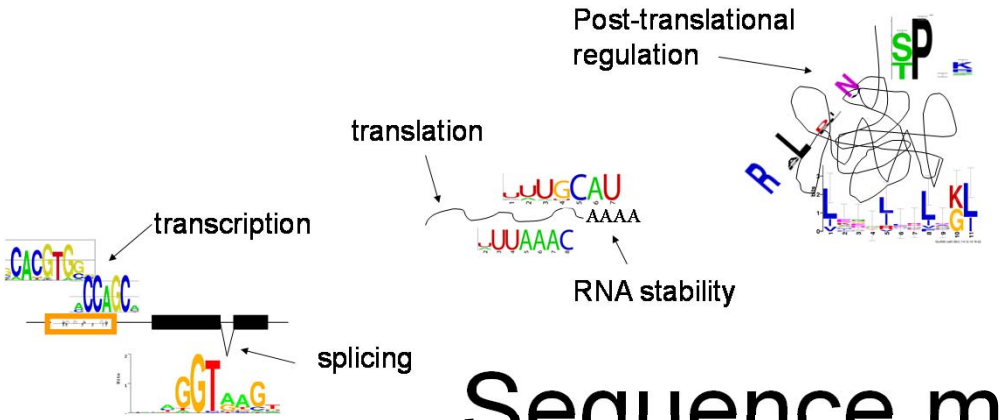


Sequence motif analysis

Alan Moses

Associate Professor and Canada Research Chair in Computational Biology

Departments of Cell & Systems Biology, Computer Science, and Ecology & Evolutionary Biology

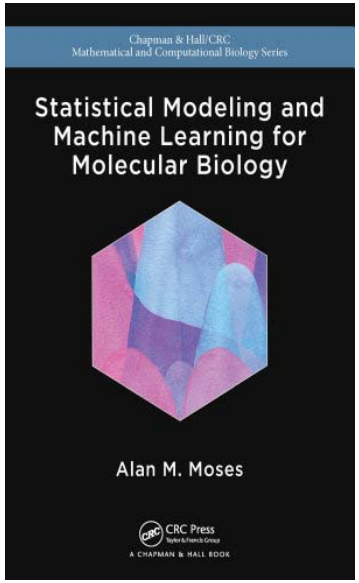
Director, Collaborative Graduate Program in Genome Biology and Bioinformatics

Center for Analysis of Genome Evolution and Function

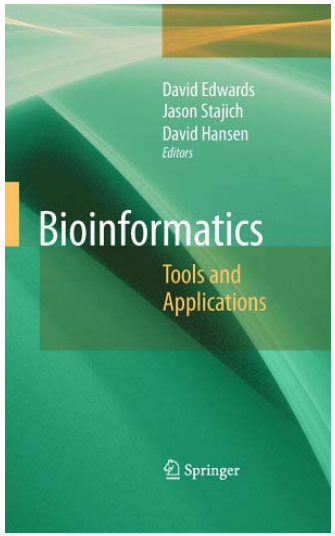
University of Toronto

Classic Bioinformatics topics!

- consensus sequences
- regular expressions
- PWMs / PSSMs / matrix models
  - Motif scanning / matrix matching
  - *De novo* motif finding
- profile HMMs



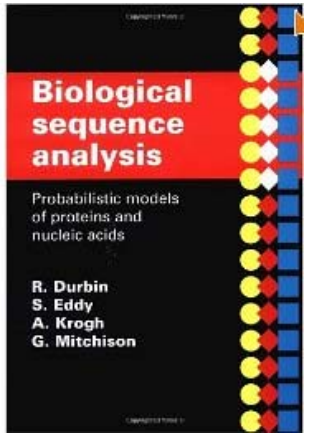
Beginner textbook



Chapter 7
Regulatory Motif Analysis

Alan Moses and Saurabh Sinha

Advanced book chapter



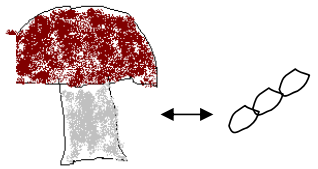
Intermediate textbook

Biology of *cis*-regulation

- Gene regulation at the level of transcription is a major factor in the
  1) response to cellular environment

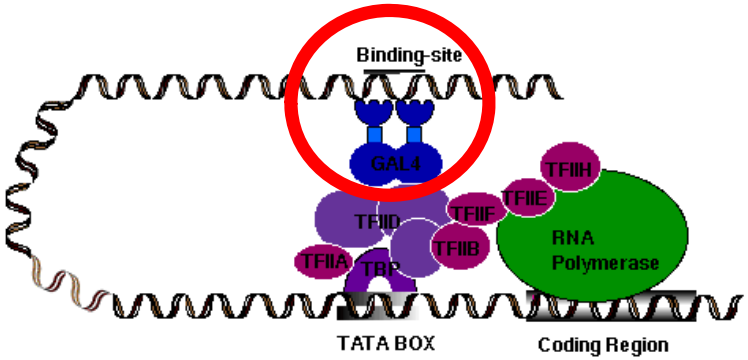
  2) diversity of cell types in complex organisms

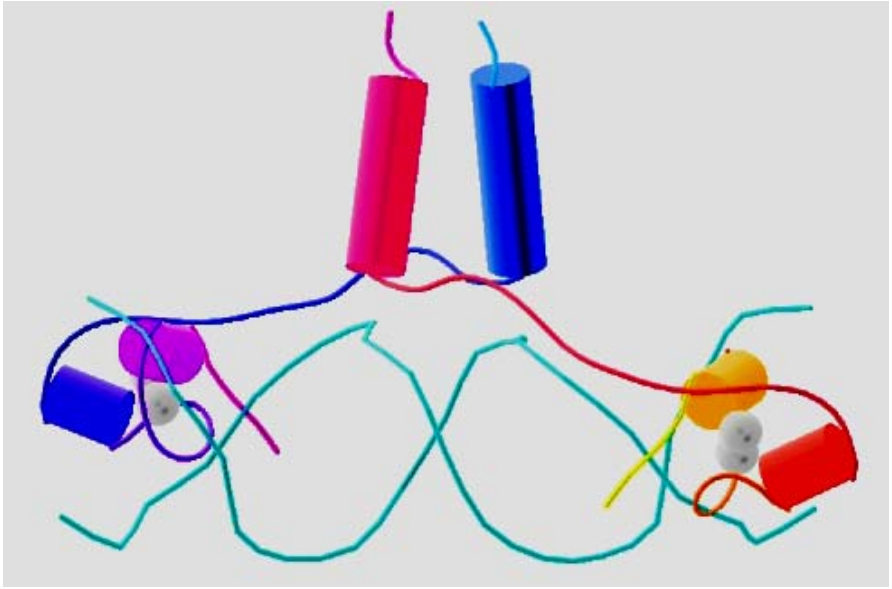
  3) evolutionary diversity of morphology



Major Mechanism:

Sequence specific
DNA binding proteins





Berman et al.: The Protein Data Bank. Nucleic Acids Research, 28 pp. 235-242 (2000).

AG**CGGAAAACTGTCCTCCG**TGCTTAA

What bases does this protein prefer to bind?

Bioinformatics. 2000 Jan;16(1):16-23.
DNA binding sites: representation and discovery.
Stormo GD

YIR032C	GATAAG
YIR032C	GGTAAG
YIR032C	GATAAG
YJL110C	GATAAT
YKR034W	GATAGA
YKR034W	GATAAC
YKR039W	GATAAG
YKR039W	GATAAC

8 GATA sites from SCPD database

"consensus" sequence or motif

string matching with mismatches (a.k.a. alignment)

(i) GATAAG with one mismatch allowed

(ii) GRTARN where R represents A or G and N represents any base

"regular expression" for string matching:

G[AG]TA[AG][ACGT]

Probabilistic models of regulatory motifs (sequence families, consensus patterns)

- Represent the sequence pattern quantitatively using a statistical model

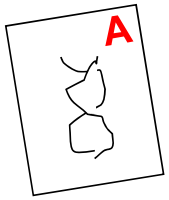
YIR032C	GATAAG
YIR032C	GGTAAG
YIR032C	GATAAG
YJL110C	GATAAT
YKR034W	GATAGA
YKR034W	GATAAC
YKR039W	GATAAG
YKR039W	GATAAC

8 GATA sites from SCPD database

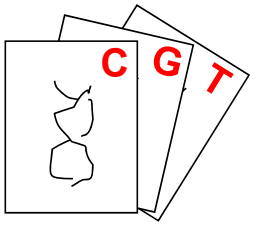
	A	*C*	*G*	*T*
1	0	0	8	0
2	7	0	1	0
3	0	0	0	8
⋮	8	0	0	0
	7	0	1	0
w	1	2	4	1

data matrix, X (*w* x |A|)

	A	C	G	T
1	0	0	8	0
2	7	0	1	0
3	0	0	0	8
⋮	8	0	0	0
	7	0	1	0
w	1	2	4	1



7



1

How to make the model of sequence specificity?

Probability, *p*

$p(\mathrm{A}) = 7 / 8$

$\equiv f_{\mathrm{A}}$

Let the data be $X = (7,0,1,0)$

$$L = p(X|\text{model}) = \prod_i f_i^{X_i}$$

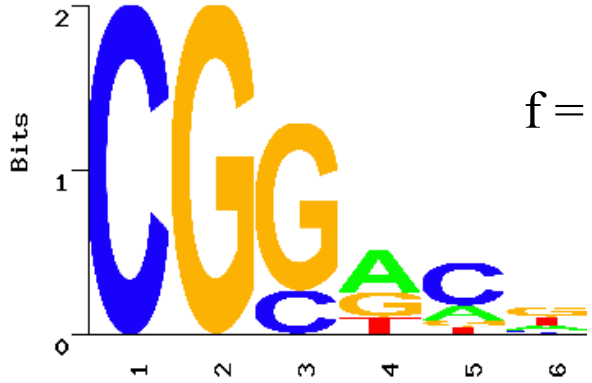
$$\log[L] = \sum_i X_i \log f_i$$

$$\frac{\partial}{\partial \mathrm{f}} \log[L] = \sum_i \frac{X_i}{f_i} = 0 \qquad \sum_i f_i = 1$$

$$\longrightarrow f_i = \frac{X_i}{\sum_i X_i}$$

7/8 is the ‘maximum likelihood’ estimate…

Probabilistic motif model



$f = (f_A, f_C, f_G, f_T)$

Probabilistic motif models are represented as "sequence logos" (Schneider et al. 1986)

Multinomial $P(X|f_1)$

$f_1 = (0, 1, 0, 0)$

Multinomial $P(X|f_2)$

$f_2 = (0, 0, 1, 0)$

Multinomial $P(X|f_3)$

$f_2 = (0, 0.7, 0.3, 0)$

Probabilistic models of regulatory motifs (sequence families, consensus patterns)

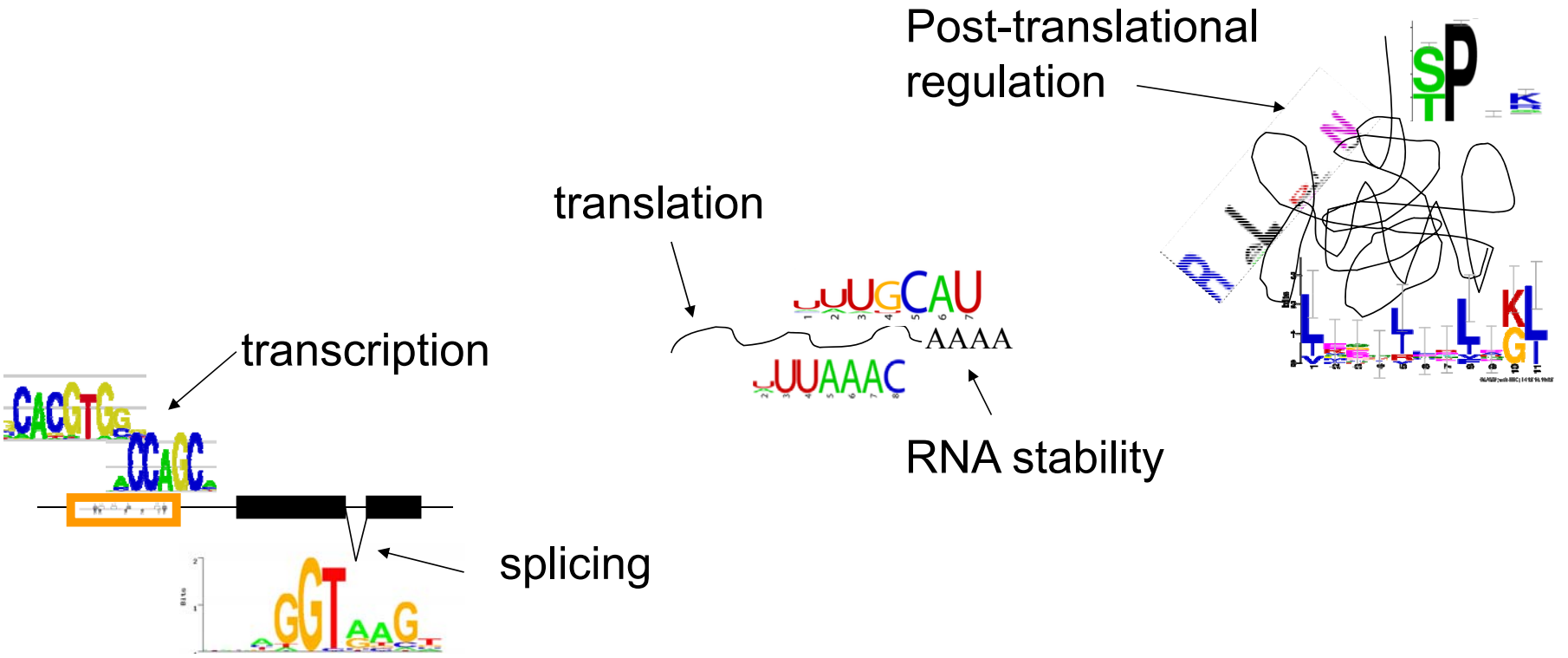
- Represent the sequence pattern quantitatively using a statistical model
- Identify “matches to the motif” or instances using a likelihood ratio score (Naïve Bayes classification)

$$S(X) = \log \frac{p(X \mid \text{motif})}{p(X \mid bg)} = \log \frac{\prod_{i=1}^{w} \prod_{b \in ACGT} f_{ib}^{X_{ib}}}{\prod_{i=1}^{w} \prod_{b \in ACGT} g_{b}^{X_{ib}}} = \sum_{i=1}^{w} \sum_{b \in ACGT} X_{ib} \log\left(\frac{f_{ib}}{g_b}\right)$$

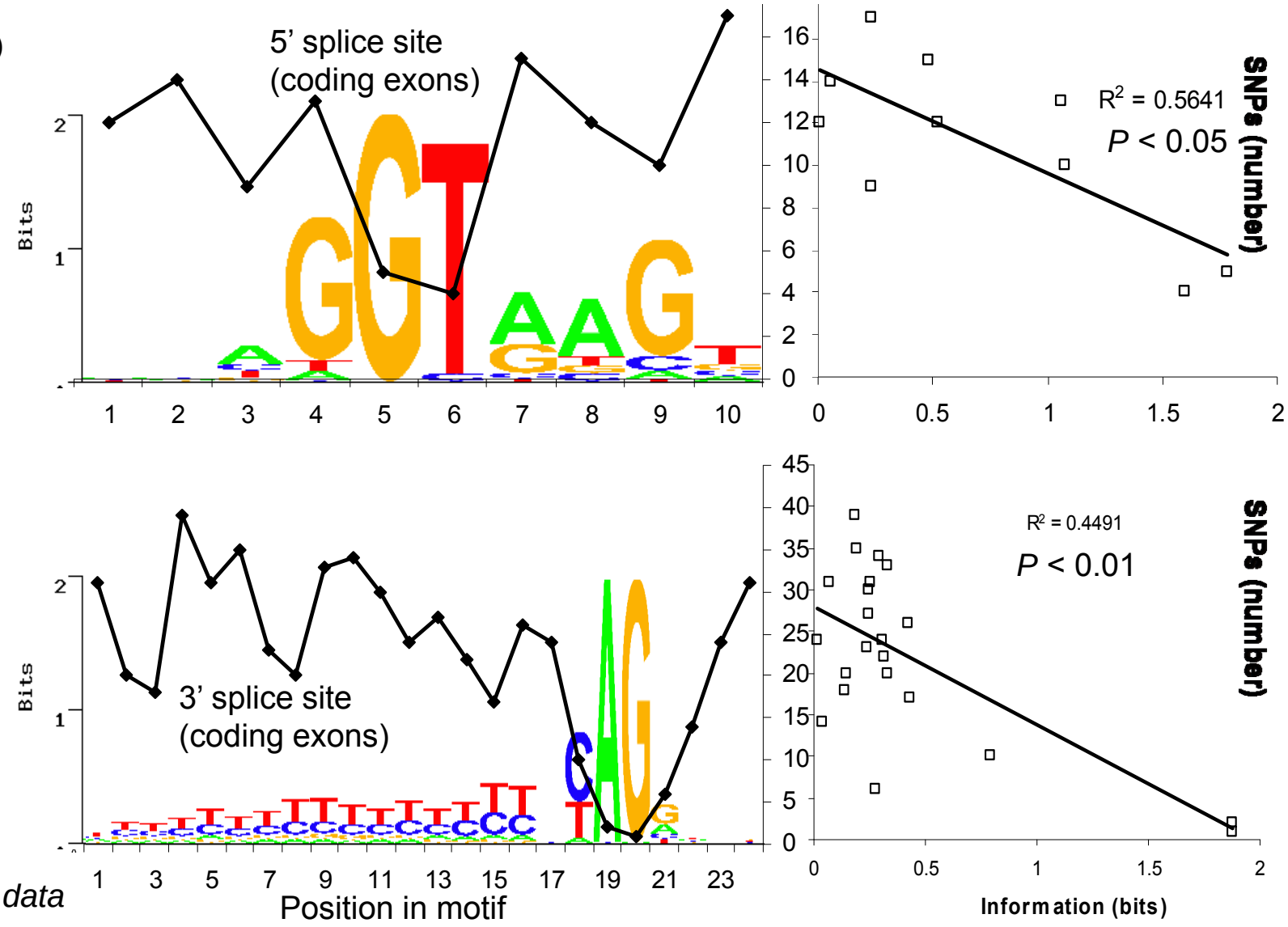
Likelihood ratio score

Sequence of length *w*

Probabilistic models of regulatory motifs (sequence families, consensus patterns)



Human Exon data (c. 2007)



Unpublished data

Human Exon data (c. 2007)

Evolution of motifs

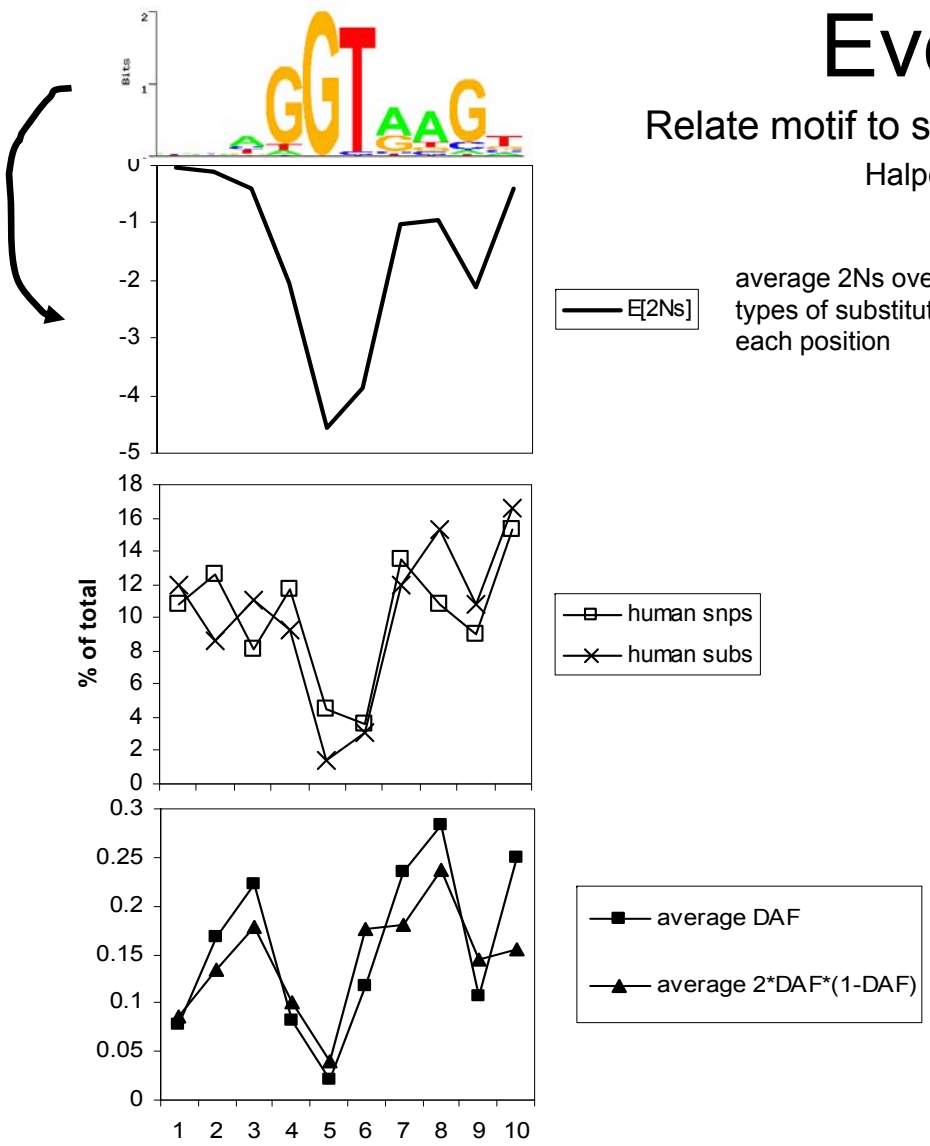
Relate motif to strength of selection:

Halpern & Bruno 1998

$$\frac{f_{ab}}{f_{ba}} \approx \frac{\frac{2s}{1 - e^{-2Ns}}}{\frac{-2s}{1 - e^{2Ns}}} = \frac{e^{2Ns} - 1}{1 - e^{-2Ns}} = e^{2Ns}. \quad (7)$$

Since relative equilibrium frequencies between a and b are determined by the ratios of their substitution rates, given reversibility, we may also express the ratio of fixation rates in terms of equilibrium frequencies (π_a, π_b) and mutation rates (p_{ab}, p_{ba}),

$$\frac{\pi_b p_{ba}}{\pi_a p_{ab}} = \frac{f_{ab}}{f_{ba}} = e^{2Ns}. \quad (8)$$

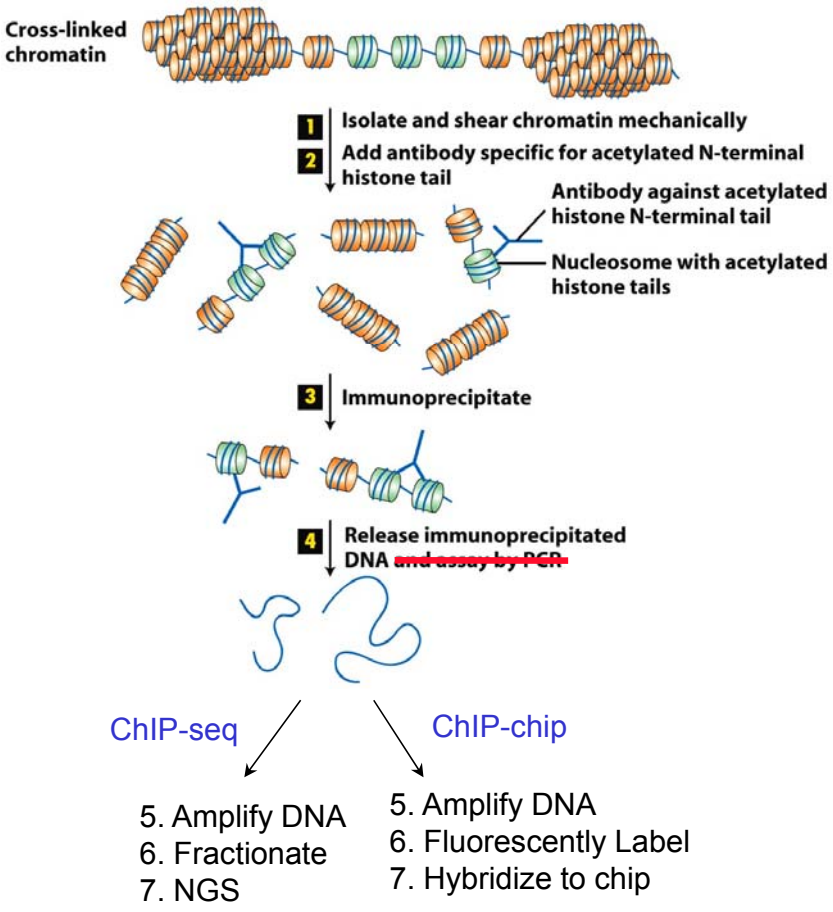


average 2Ns over all types of substitutions at each position

Patterns of genetic variation quantitatively reflect preservation of motifs by natural selection

Moses et al. 2003

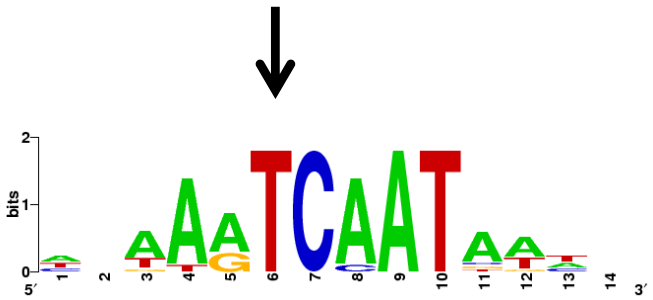
ChIP-chip and ChIP-seq



Can be used to measure binding of transcription factors to the entire genome

Allelic imbalance in ChIP-seq

```
HNF6 (from Lannoy et al. JBC 1998)
AAAAATCAATATCG    r.hnf-3b -126/-139
CTAAGTCAATAATC    m.ttr             -105/-92
TGAAATCAATTTCA    r.pfk-2 prom.     -211/-198
AAAAATCCATAACT    r.pfk-2 GRUc      1st intron
AGAAGTCAATGATC    m.hnf-4           -389/-376
TTTAGTCAATCAAA    r.pepck           -258/-245
TTTTATCAATAATA    r.a2-ug           -183/-196
CAGAATCAATATTT    r.cyp2c13         -41/-54
AAAAATCAATATTT    r.cyp2c12         -36/-49
TCTAATCAATAAAT    m.mup             -132/-119
ATAAATCAATAGAC    r.tog             -208/-221
CAAAGTCAATAAAG    r.a-fp            -6103/6090
```



I used these 10 columns

Look at peaks in liver of a Heterozygous Mouse

data from Mike Wilson c. 2011

chr19:34758655-34758994 Between Slc16a12 (monocarboxylic acid transporter) and Pank1 (liver CoA biosynthesis)

HNF6 peak score=162.35

rs31029568 ref=A var=G 1/1:99:52:0:255,157,0:1

```
Alignment block 1 of 2 in window, 34758754 - 34758795, 42 bps
B D            Mouse  gctca-tttgatgaggacagaa-aaagtcaatagaaccaagtga
B D              Rat  ggtcg-tttgatgagaacagaa-aaagtcggtagaacaaagtga
B D     Kangaroo rat  aaccattttgatgagaagagag-aaaatcaatggacaaagtga
B D   Naked mole-rat  catcattttgatgagatcagaa-aa--tcaatggacaaagtga
B D        Guinea pig  agtcatctttgtgagggcagaa-aa--tcaatggacaaagcaa
B D          Squirrel  agccatttcaatgagaactggg-taaatcaatggacaaagaga
B D            Rabbit  agtcattttgatgagaacagag-aaaacca--tagataaagtga
B D              Pika  agtcattttgatgagaatagag-aaaccccaataagataaagtga
B D             Human  a------ttgatgagaatggag-aaaatcaatggacaaagtga
B D             Chimp  agtcattttgatgagaatggag-aaaatcaatggacaaagtga
B D           Gorilla  agtcattttgatgagaatggag-aaaatcaatggacaaagtga
B D         Orangutan  agtcattttgatgagaatggag-aaaatcaatggacaaagtga
B D            Gibbon  agtcattttgatgagaatggag-aaaatcaatggacaaggtga
B D            Rhesus  agtcattttgatgagaatggag-aaaatcaatggacaaagtga
B D            Baboon  agtcattttgatgagaatggag-aaaatcaatggacaaagtga
B D          Marmoset  agtcattttgatgagaatggag-aaaatcaatcggacaaagtga
B D  Squirrel monkey  agtcattttgatgagaatggag-aaaatcaattggacaaagtga
B D       Mouse lemur  catcattctgataagaatggaa-aaaatcaatggacaaagtga
B D          Bushbaby  tgtcattttgataaagatggag-aaaatcaatggccaaagtga
B D               Pig  agtcgttttgacgagaacaggg-aaaatcaatggacaaagtga
B D            Alpaca  agtcattttgatgagaacagag-aaaatcaatggacaaaggga
B D           Dolphin  agtcattttgatgagaacaggg-aaaatcaatggacaaagtga
B D             Sheep  agtcttttgatgagaacaggg-aaaatcaatggacaaagtga
B D               Cow  agtcttttgatgagaacaggg-aaaatcaatggacaaagtga
B D               Cat  agtcattttgaggagaacagagaaaaatcaatggacaaagtgc
B D               Dog  agtcattttgatgagggcagag-aaaatcaatggatagagtgc
B D             Panda  agtcattttgatgagagcagag-aaaatcaatggacaaagtgc
B D             Horse  agtcattttgatgaaaacag---aaaatcaatggacaaagtga
B D          Microbat  a--------tacgagaacagag-gaaatcaatagaa-aaagtga
B D           Megabat  agtcattttgacgagaacagag-aaaatcaatggacaaaatga
B D             Shrew  agtcattttgaggagaacagagagaaatcggtgcgacaaaatga
B D          Elephant  -gcaattttgaagagaacagag-gaaatcaatggacaaagtga
B D            Tenrec  -gccgttttgaagagaacagag-gaaatcaatggacagtgtga
B D           Manatee  -gccattttgaaaagaacagag-gaaatcaatggacaaaatga
B D         Armadillo  -gtcattttgatgagaacagag-aaaatcaatggacaaagtga
B D             Sloth  -ttcaatttgatgagaacagag-aaaatcaatggacaaagtga
```

HNF6: 

ref: AAGTCA**A**TAG S=7.56

var: AAGTCA**G**TAG S=5.40

rs31029568 ΔS=2.16

In do40:

70 reads with ‘A’

10 reads with ‘G’ Q>20

Binomial tests:

$Z_{f=0.5} = 6.7$ Heterozygote

$Z_{f=0.99} = -10.3$ Homozygote, 1% error

Strong evidence for imbalance in the direction predicted by binding affinity

chr1:39813171-39813514 Not near any genes of known function

HNF6 peak score=301.85

rs47325601 ref=C var=T 1/1:99:32:0:255,96,0:1

```
Alignment block 1 of 2 in window, 39813380 - 39813712, 33 bps
B D            Mouse  cgtgagcaaata-----t----aatcaaccaaagca-ca--tgag
B D              Rat  tgtgagcaaata-----c----aatcaatcagagca-caagagag
B D   Naked mole-rat  cataagcaaaca-----cgatgaatcaatcaatgca-ta--ggtg
B D        Guinea pig  cataagcaaata-----cggtgaaccaatcaatgtg-ta--gcca
B D          Squirrel  cagaagcaaagg-----c----aatgactcaatgca-ca--gcag
B D            Rabbit  cacaagcaaata-----c----aatgaatcaatgca-tg--agtg
B D             Human  cacaagcaaata-----t----aatgaatcaatgca-ta--gccg
B D             Chimp  cacaagcaaata-----t----aatgaatcaatgca-ta--gccg
B D           Gorilla  cacaagcaaata-----t----aatgaatcaatgca-ta--gccg
B D            Gibbon  cacaagcaaata-----t----aatgaatcaatgca-ta--gctg
B D            Rhesus  cataagcaaata-----t----aatgaatcaatgca-ta--gcgg
B D            Baboon  cataagcaaata-----t----aatgaatcaatgca-ta--gcgg
B D          Marmoset  caagagcaaata-----c----aatgaatcaatgca-ta--gccg
B D  Squirrel monkey  cacgagcaaata-----c----aatgaatcaatgca-ta--gccg
B D           Tarsier  cacaagtaaata-----c----aattaatcaatgca-tg--gcag
B D       Mouse lemur  cacaagcaaatattatgc----aatgaatcaatgca-ta--gcaa
B D         Bushbaby  cacaagcaaata-----c----catgagtcaatgca-ca--gcag
B D        Tree shrew  cat-agcaaata-----t----aaggaatcaatgca-tc--gcag
B D               Pig  cac-agcaaaca-----c----aatgaatcaatgca-ta--gaag
B D            Alpaca  cacaagcagacg-----t----agctgagcaggacgcta--tgag
B D           Dolphin  cacaagcaaata-----c----aataaatcaaagca-ta--gaag
B D               Cat  cataagcaaata-----c----aatgaatcaacaca-ta--gaag
B D               Dog  cacaagcaaaca-----c----aatgaatcaatgca-tt--gccc
B D             Panda  cacaagcaaata-----c----aatgaatcaaagca-ta--gaag
B D             Horse  cacaagcaactg-----c----aatg---------a-tt--gaag
B D          Microbat  c--acgcaaata-----c----aataaatcaatgcc-ta--gaag
B D           Megabat  caaaagcaaaca-----c----aatgaatcaatgcc-ta--gaag
B D          Elephant  cacaagcaaata-----c----actgaatcaacgca-ta--gaag
B D           Manatee  cgcaagcaaata-----c----aatgaatcagtgca-ta--gaag
B D         Armadillo  cacaagcaaata-----g----aatgaatcaaggca-ca--gaag
```

HNF6: 

ref: TAATCAA**C**CA S=4.54

var: TAATCAA**T**CA S=6.71

rs47325601 $\Delta S = -2.16$

In do40:

21 reads with ‘C’

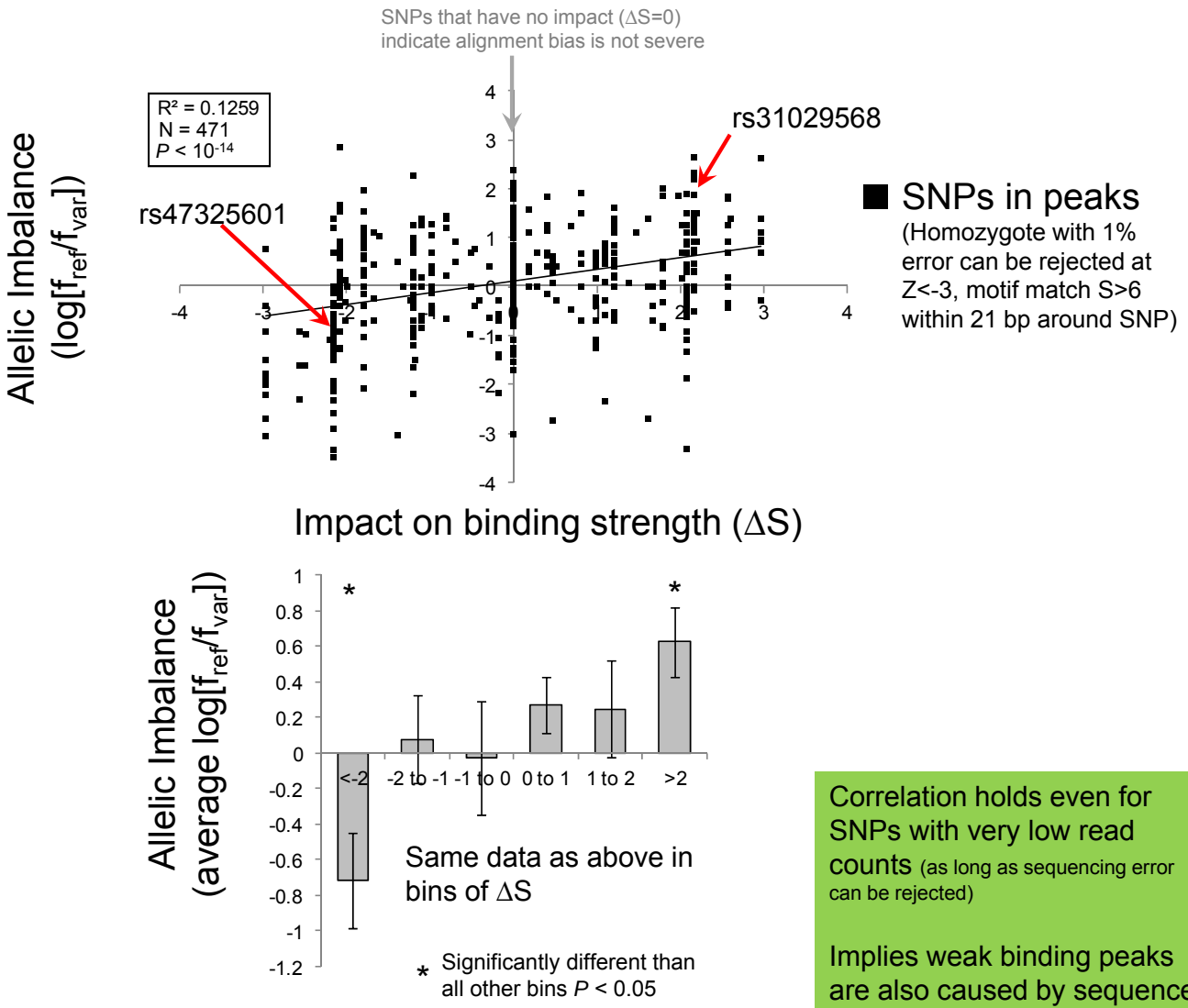
59 reads with ‘T’ Q>20

Binomial tests:

$Z_{f=0.5} = 4.2$ Heterozygote

$Z_{f=0.99} = -65.4$ Homozygote, 1% error

Strong evidence for imbalance in the direction predicted by binding affinity



Correlation holds even for SNPs with very low read counts (as long as sequencing error can be rejected)

Implies weak binding peaks are also caused by sequence specific TF-DNA interactions

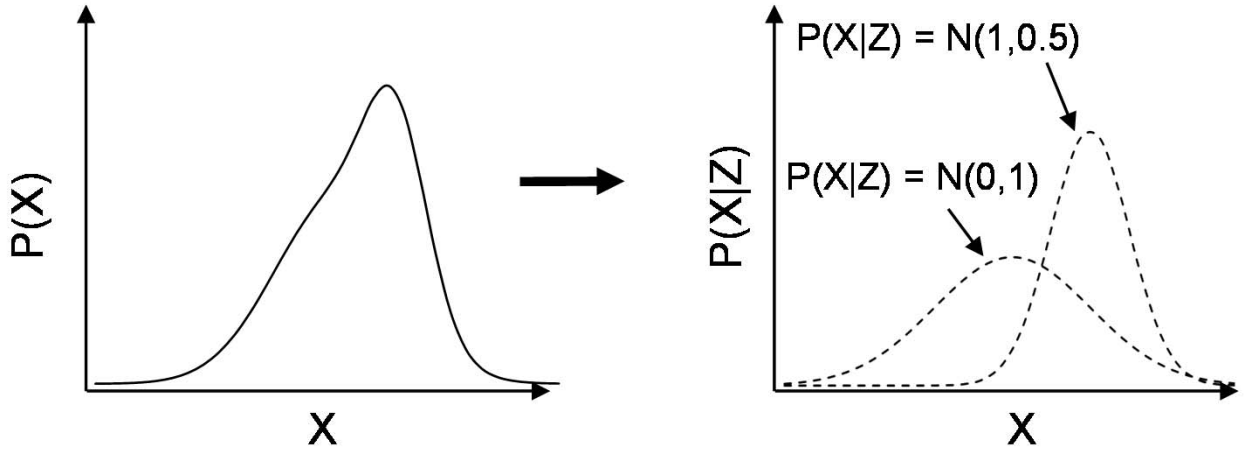
Unpublished data

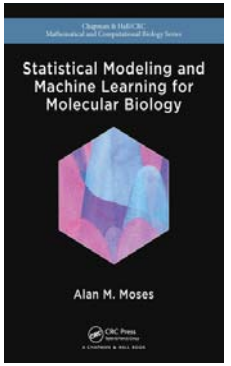
De novo motif-finding with the probabilistic model

- Search for statistically enriched sequence families in non-coding DNA

- DNA sequences contain some w-mers from the motif model, and some from the background model

Mixture model



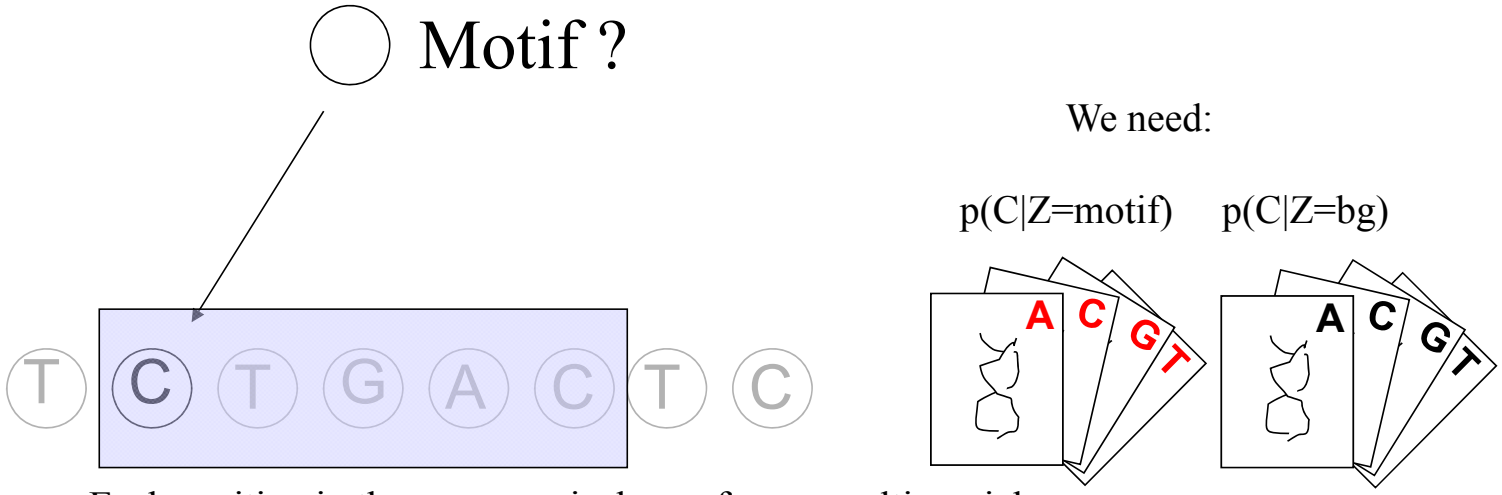


A two-component Mixture Model

We assume that…

Each position in the sequence is drawn from a multinomial representing the background or a position in the motif

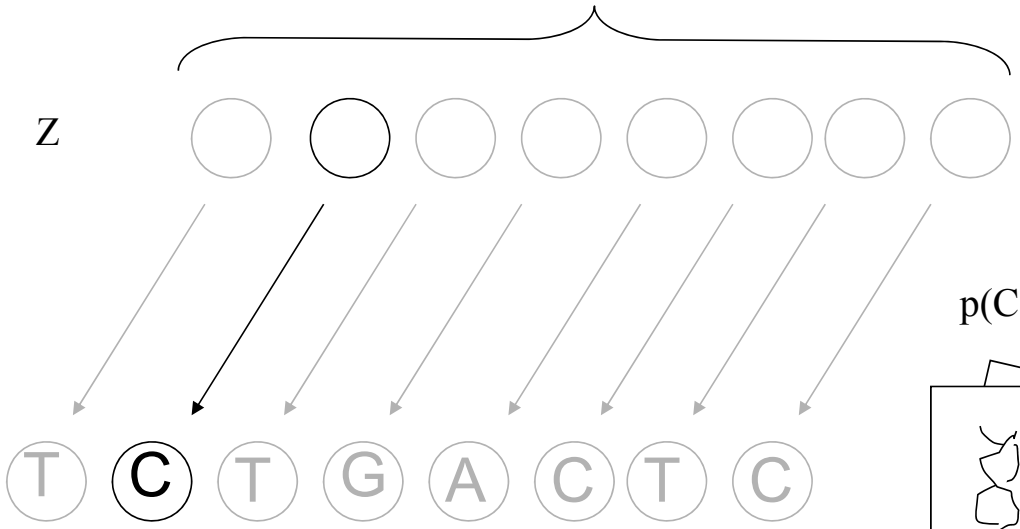
A two-component Mixture Model



Each position in the sequence is drawn from a multinomial representing the background or a position in the motif

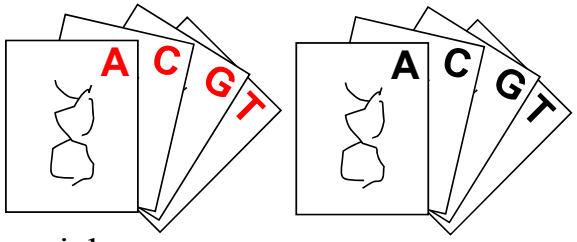
A two-component Mixture Model

Unobserved indicator variables determine whether each position is drawn from motif or background component



We need:

p(C|Z=motif) p(C|Z=bg)



Each position in the sequence is drawn from a multinomial representing the background or a position in the motif

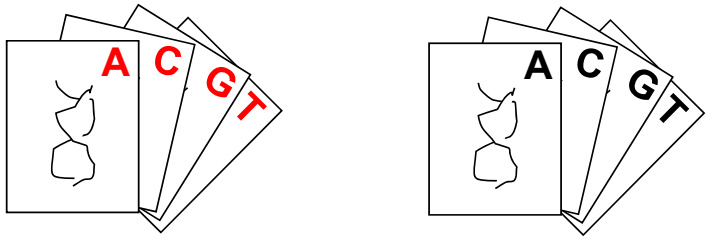
E.g., MEME

Bailey TL, Elkan C *ISMB*, pp. 28-36, AAAI Press, Menlo Park, California, (1994.)

How to estimate parameters?

Even though there are "hidden" variables, it's still possible to write the likelihood and find Maximum-Likelihood estimates for

$p(X|Z=\text{motif})$ $p(X|Z=\text{bg})$

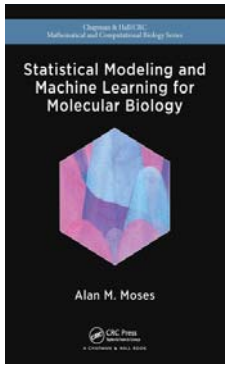


Using various optimization schemes :

Stormo GD, Hartzell GW 3rd. *Proc Natl Acad Sci U S A*. (1989) Feb;86(4):1183-7.

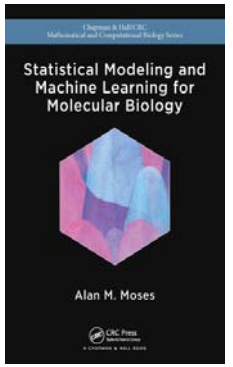
Lawrence CE, Reilly AA. *Proteins*. 1990;7(1):41-51.

Lawrence CE et al. *Science*. (1993) Oct 8;262(5131):208-14.



EM notes

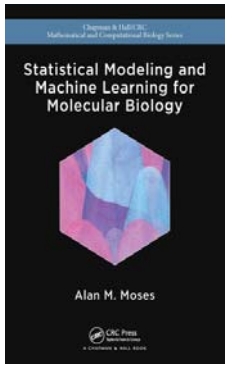
- Guaranteed to increase the likelihood at each iteration
- Can get stuck in local maxima
- Non-linear gradient ascent – can be very fast
- Usually not possible to derive analytic updates for all parameters
- General formulation based on graph theory allows application to complex models



Unsupervised classification

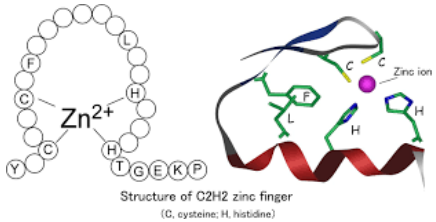
- Classification when you don't have a training set is called "unsupervised".
- Finding motifs "*de novo*" is a form of unsupervised classification (or clustering)
- Fundamentally hard problem because motifs appear by chance in long enough sequences

See e.g., Zia & Moses 2012



Motif models so far assume all fixed width, *w*

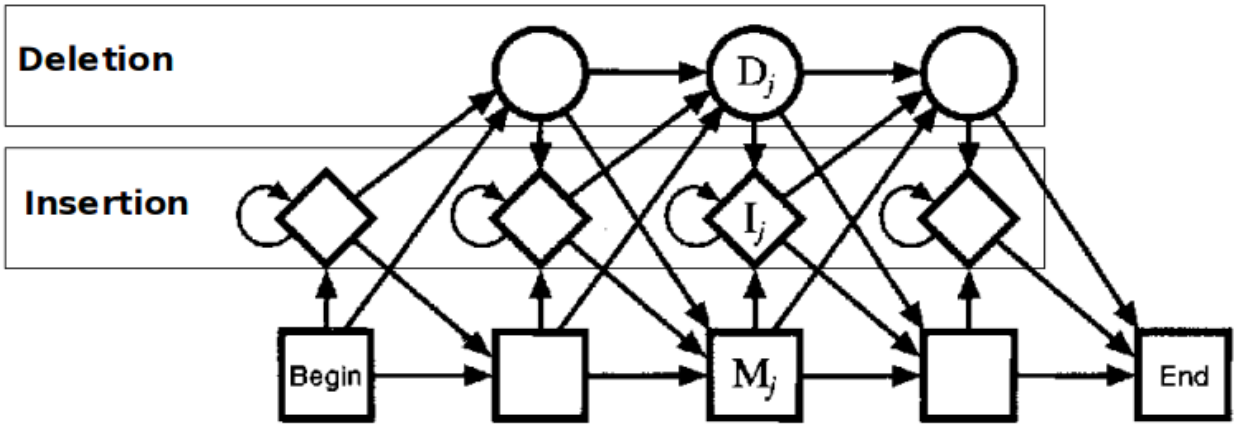
- Many biological patterns have variable lengths
- Especially important for protein families and domains (but also for some TFs, e.g., p53)
- Like the residues, insertions and deletions have position-specific preferences



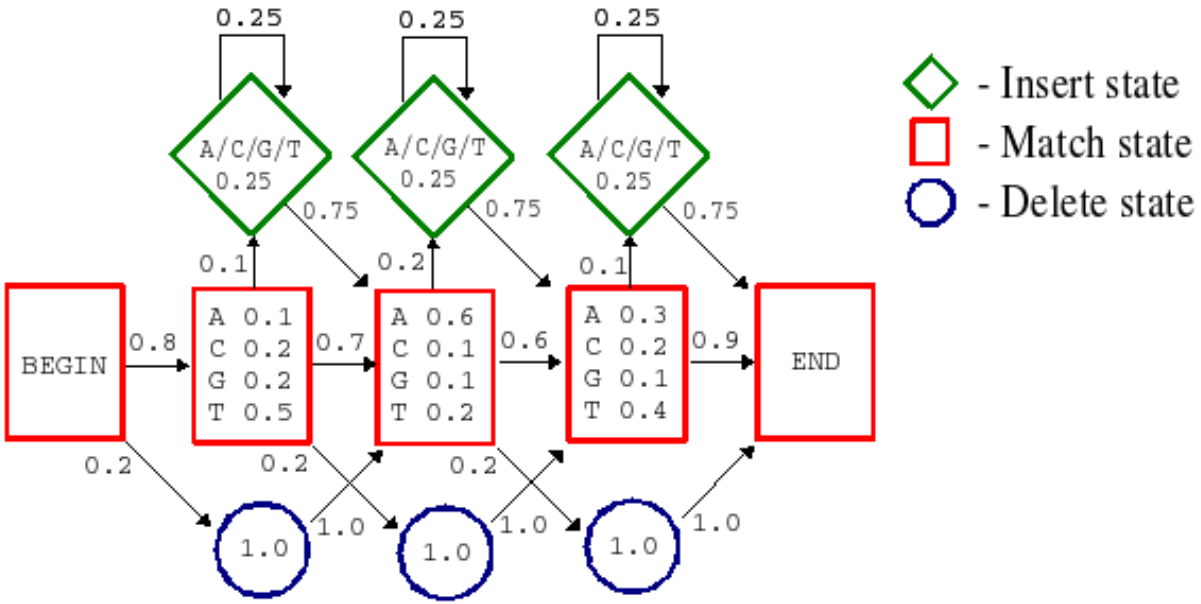

```
EVI1_HUMAN/131-154          YECE...NCA...KVFTDPSNLQRHIRS.QH
TRA1_CAEEL/337-362          YSCQI.PQCT...KSYTDPSSLRKHIKA.VH
H9ZJM4_DROME/370-392        CKCN...LCG...KAFSRPWLLQGHIRT..H
ZSC22_HUMAN/352-374         YKCG...ECG...KTFSRSTHLIQHQRV..H
ZN239_MOUSE/6-28            YKCD...KCG...KGFTRSSSLLVHHSV..H
A0A024RC04_HUMAN/488-510    IECD...ECG...KHFSHAGALFTHKMV..H
G2HH24_PANTR/517-539        YKCN...QCG...IIFSQNSPFIVHQIA..H
ZNF17_HUMAN/442-464         YECN...KCG...KFFRYCFTLNRHQRV..H
ZFP59_MOUSE/493-515         FECK...VCG...KSFKRESNLIQHGAV..H
ZFP60_MOUSE/428-450         YQCK...DCW...EFFRRRSNFIEHQSI..H
A0A023ZFK3_YEASX/50-73      FQCN..I.CL...KFFSRIDNLRQHQSS.VH
ZFP60_MOUSE/344-366         FECK...QCG...KIFSNGSYLLRHYDT..H
ZFP59_MOUSE/326-348         FECN...VCG...SAFRLQLYLSEHQKT..H
ZFP60_MOUSE/484-506         FECK...ECG...KAFHFSSQLNNHKTS..H
ZFP59_MOUSE/270-292         FQCK...DCG...KGFIVLAHLTRHQSS..H
TRA1_CAEEL/306-331          YKCEF.ADCE...KAFSNASDRAKHQNR.TH
```

Profile Hidden-Markov Models

- Probabilistic model with three types of states



- The states are ‘hidden’, but they explain the sequences that we observe


0.25
0.25
0.25
A/C/G/T
0.25
A/C/G/T
0.25
A/C/G/T
0.25
0.75
0.75
0.75
0.1
0.2
0.1
BEGIN
0.8
A 0.1
C 0.2
G 0.2
T 0.5
0.7
A 0.6
C 0.1
G 0.1
T 0.2
0.6
A 0.3
C 0.2
G 0.1
T 0.4
0.9
END
0.2
0.2
0.2
1.0
1.0
1.0
1.0
1.0
- Insert state
- Match state
- Delete state

TAT

T

TACT

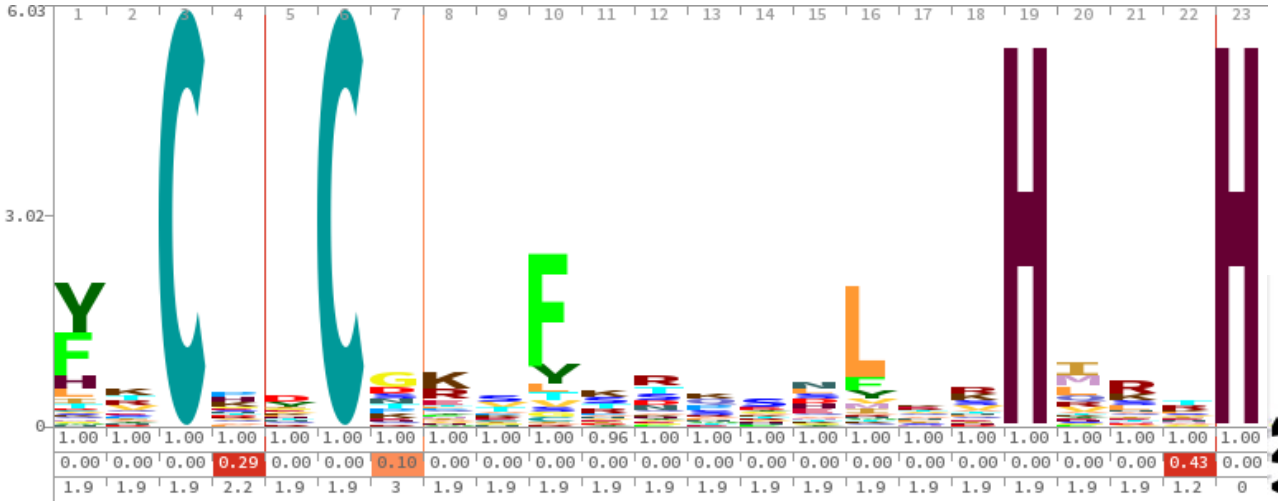
Profile Hidden-Markov Models

- How to get the parameters?

  – For HMMer (Pfam) this is usually done starting with a (manual) alignment and priors

  – GLAM2 offers unsupervised estimation of HMMs, also relies heavily on priors

```
EVI1_HUMAN/131-154          YECE...NCA...KVFTDPSNLQRHIRS.QH
TRA1_CAEEL/337-362          YSCQI.PQCT...KSYTDPSSLRKHIKA.VH
H9ZJM4_DROME/370-392        CKCN...LCG...KAFSRPWLLQGHIRT..H
ZSC22_HUMAN/352-374         YKCG...ECG...KTFSRSTHLTQHQRV..H
ZN239_MOUSE/6-28            YKCD...KCG...KGFTRSSSLLVHHSV..H
A0A024RC04_HUMAN/488-510    IECD...ECG...KHFSHAGALFTHKMV..H
G2HH24_PANTR/517-539        YKCN...QCG...IIFSQNSPFIVHQIA..H
ZNF17_HUMAN/442-464         YECN...KCG...KFFRYCFTLNRHQRV..H
ZFP59_MOUSE/493-515         FECK...VCG...KSFKRESNLIQHGAV..H
ZFP60_MOUSE/428-450         YQCK...DCW...EFFRRRSNFIEHQSI..H
A0A023ZFK3_YEASX/50-73      FQCN..I.CL...KFFSRIDNLRQHQSS.VH
ZFP60_MOUSE/344-366         FECK...QCG...KIFSNGSYLLRHYDT..H
ZFP59_MOUSE/326-348         FECN...VCG...SAFRLQLYLSEHQKT..H
ZFP60_MOUSE/484-506         FECK...ECG...KAFHFSSQLNNHKTS..H
ZFP59_MOUSE/270-292         FQCK...DCG...KGFIVLAHLTRHQSS..H
TRA1_CAEEL/306-331          YKCEF.ADCE...KAFSNASDRAKHQNR.TH
```



Wheeler et al. *BMC Bioinformatics* 2014

What are priors?

Why are priors so important for profile HMMs ?

CABIOS

Vol. 12 no. 4 1996
Pages 327–345

Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology

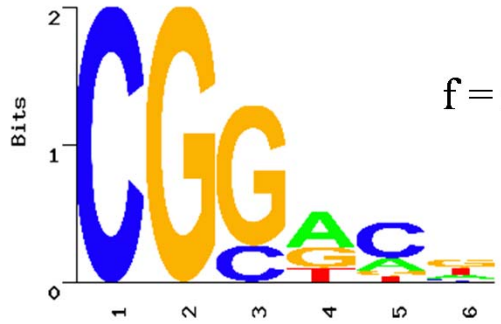
Kimmen Sjölander[3], Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh[1], I.Saira Mian[2] and David Haussler

With accurate prior information about which kinds of amino acid distributions are reasonable in columns of alignments, it is possible with only a few sequences to identify which prototypical distribution may have generated the amino acids observed in a particular column. Using this informed guess, we adjust the expected amino acid probabilities to include the possibility of amino acids that may not have been seen but are consistent with observed amino acid distributions. The statistical models produced are more effective at generalizing to previously unseen data, and are often superior at database search and discrimination experiments (Brown *et al.*, 1993; Tatusov *et al.*, 1994; Bailey and Elkan, 1995; Karplus, 1995a; Hughey and Krogh, 1996; Henikoff and Henikoff, 1996; Wang *et al.*, 1996).

- For DNA/RNA, it's not as important what priors you use, but you still need something that gives psuedocounts

$$S(X) = \log \frac{p(X \mid \text{motif})}{p(X \mid bg)} = \log \frac{\prod_{i=1}^{w} \prod_{b \in ACGT} f_{ib}^{X_{ib}}}{\prod_{i=1}^{w} \prod_{b \in ACGT} g_b^{X_{ib}}} = \sum_{i=1}^{w} \sum_{b \in ACGT} X_{ib} \log\left(\frac{f_{ib}}{g_b}\right)$$

- For proteins, without priors, models of highly variable protein motifs/domains with few known examples will always be overfit



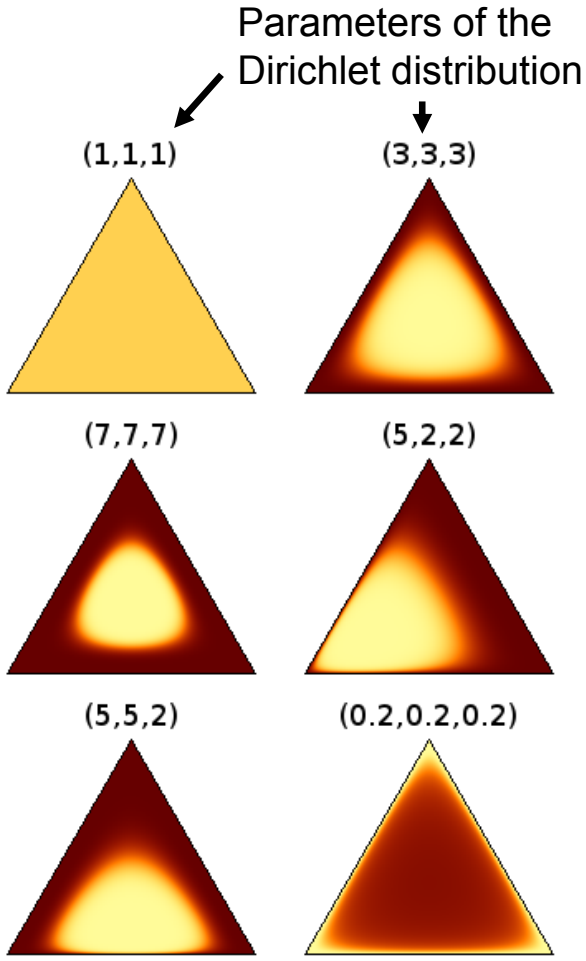
$f = (f_A, f_C, f_G, f_T)$

Multinomial
$P(X|f_1)$

$f_1 = (0, 1, 0, 0)$

How do we put a prior on DNA or protein data matrices?

We need a prior belief about the multinomial parameters?



Dirichlet distribution
(3 dimensional distribution is illustrated, for DNA we need 4 dimensions and for protein we need 20 dimensions)

	A	C	G	T
1	0	0	8	0
2	7	0	1	0
3	0	0	0	8
⋮	8	0	0	0
	7	0	1	0
w	1	2	4	1

$X = (7,0,1,0)$

$p(\text{A}) = 7 / 8$

$\equiv f_{\text{A}}$

$$f_i = \frac{X_i}{\sum_i X_i}$$

7/8 is the ‘maximum likelihood’ estimate…

Observed residues

Dirichlet distribution

Parameters of the Dirichlet distribution

$$f_i = \int_{\vec{p}} p_i \, \text{Prob}\,(\vec{p}|\vec{\alpha}, X)\mathrm{d}\vec{p} = \frac{X_i + \alpha_i}{\sum_i X_i + |\vec{\alpha}|}$$

Modified from Sjolander et al. 1996

With priors, we have a “Bayesian” estimator: mean posterior estimate

E.g., If $\vec{\alpha} = (3,2,2,3)$ Where did I get these parameters?

$p(\text{A}) = 10 / 18 = 5 / 9$

$\equiv f_{\text{A}}$

- For proteins, a single column of parameters in not enough
  - Biochemical similarity
  - Surface vs. core
  - Secondary structure vs. disordered
  - And more
- Amino acid probabilities in proteins have a very complicated distribution, and these are not uniform across the protein sequence

Use a mixture of Dirichlet distributions!

- Sjolander et al. 1996 trained a 9-component mixture of Dirichelet priors (Blocks9)

The mixture of Dirichlet prior

Table II. Preferred amino acids of Blocks9

Component	Ratio (r) of amino acid frequency relative to background frequency							
	$8 \leq r$	$4 \leq r \leq 8$	$2 \leq r \leq 4$	$1 \leq r \leq 2$	$1/2 \leq r < 1$	$1/4 \leq r < 1/2$	$1/8 \leq r < 1/4$	$r < 1/8$
1			SAT	CGP	NVM	QHRIKFLDW	EY	
2	Y	FW	H		LM	NQICVSR	TPAKDGE	
3			QE	KNRSHDTA	MPYG	VLIWCF		
4		KR	Q	H	NETMS	PWYALGVCI	DF	
5		LM	I	FV		WYCTQ	APHR	KSENDG
6		IV		LM	CTA	F	YSPWN	EQKRDGH
7		D	EN	QHS	KGPTA	RY	MVLFWIC	
8			M	IVLFTYCA	WSHQRNK	PEG	D	
9			PGW	CHRDE	NQKFYTLAM	SVI		

The function used to compute the ratio of the frequency of amino acid i in component j relative to the background frequency predicted by the mixture as a whole is $(\alpha_{j,i}/|\vec{\alpha}|)/(\sum_k q_k \alpha_{k,i}/|\vec{\alpha}_k|)$.
An analysis of the amino acids favored by each component reveals the following:
Component 1 favors small neutral residues.
Component 2 favors the aromatics.
Component 3 gives high probability to most of the polar residues (except for C, Y and W).
Component 4 gives high probability to positively charged amino acids and residues with NH_2 groups.
Component 5 gives high probability to residues that are aliphatic or large and non-polar.
Component 6 prefers I and V (aliphatic residues commonly found in β sheets), and allows substitutions with L and M.
Component 7 gives high probability to negatively charged residues, allowing substitutions with certain of the hydrophilic polar residues.
Component 8 gives high probability to uncharged hydrophobics, with the exception of glycine.
Component 9 gives high probability to distributions peaked around individual amino acids (especially P, G, W and C).

Profile Hidden-Markov Models

- How to get the parameters?

  – For HMMer (Pfam) this is usually done starting with a (manual) alignment and priors

  – GLAM2 offers unsupervised estimation of HMMs, also relies heavily on priors

- How to scan DNA/RNA/Protein sequences?

  – HMMer uses forward algorithms (after some speedup tricks)

  – GLAM2 uses an alignment based approach

Scoring DNA/proteins sequences with an HMM

- We want to calculate the probability of a sequence, X, given our profile HMM model, and compare this to a background model.
- In general, multiple paths through an HMM can give us the same sequence (unlike for a PSSM)
- We need to sum over all of the possible paths through the model, weighting each by their probability
- This is done using the “forward algorithm”

Forward algorithm for HMMs

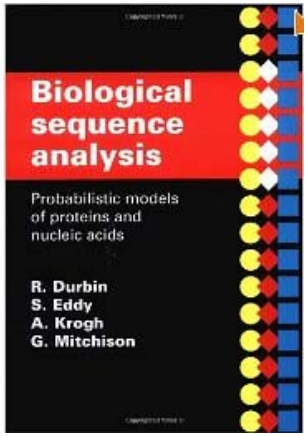
- Even though there are exponentially many paths, they are all made out of the same pieces.
- The trick is to add up the probability of all the paths that can get you to this point in the HMM, but don't keep track of all of the paths.
- Example of "dynamic programming"

Forward algorithm for HMMs

Step 1: Recursively fill a n x 3 matrix

$$F_{Zi} = p(X\ until\ i|Z_i) = p(X_i|Z_i) \sum_{Z_{i-1} \in (match,ins,del)} F_{Z_{i-1}i-1}\ p(Z_i|Z_{i-1})$$

'transition' probabilities between hidden states



$$p(X|profile\ HMM) = p(X\ until\ n) = \sum_{Z \in (match,ins,del)} F_{Zn}$$

Step 2: Add up the likelihoods at the last state

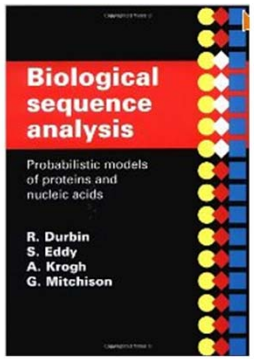
How does HMMer score sequences?

The null model is a one-state HMM configured to generate “random” sequences of the same mean length as the target sequence, with each residue drawn from a background frequency distribution (a standard i.i.d. model: residues are treated as independent and identically distributed).

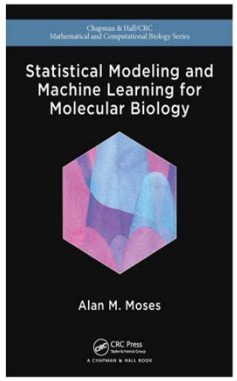
$$\text{Log}\ \frac{p(X \mid \text{profile HMM})}{p(X \mid \text{background HMM})}$$

Classic Bioinformatics topics!

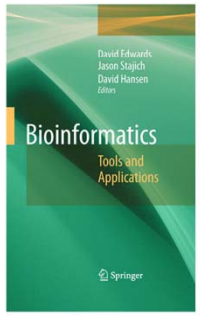
- consensus sequences
- regular expressions
- PWMs / PSSMs / matrix models
  - Motif scanning / matrix matching
  - *De novo* motif finding
- profile HMMs



Intermediate textbook



Beginner textbook



Chapter 7
Regulatory Motif Analysis

Alan Moses and Saurabh Sinha

Advanced book chapter

Questions for Glam2 discussion

- What is the prior model for indels in Glam2?
- How does the Glam2 model differ from a profile HMM?
  - Why does this make it easier to estimate parameters?
- What's the difference between an alignment-based score and the full Forward algorithm?
- What is the optimization strategy used by Glam2? Why not use EM?