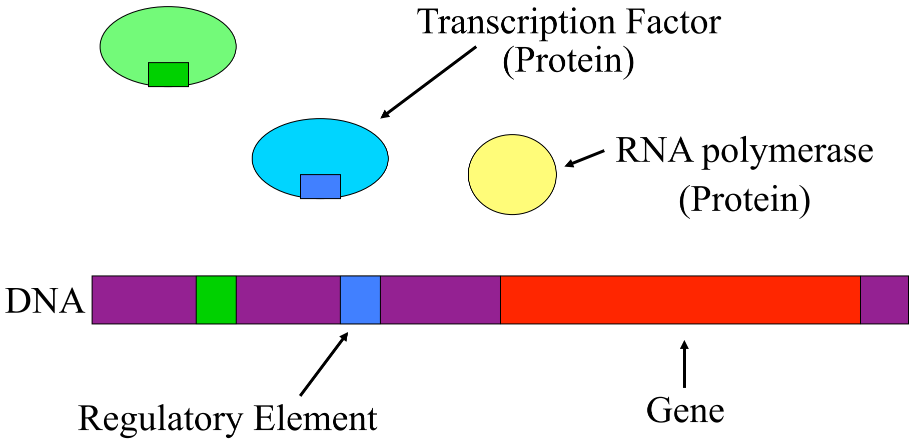
> ONE SHOULD NEVER
> MISTAKE PATTERN
> FOR MEANING
>
> IAIN BANKS

Regulation of Genes



Transcription Factor
(Protein)

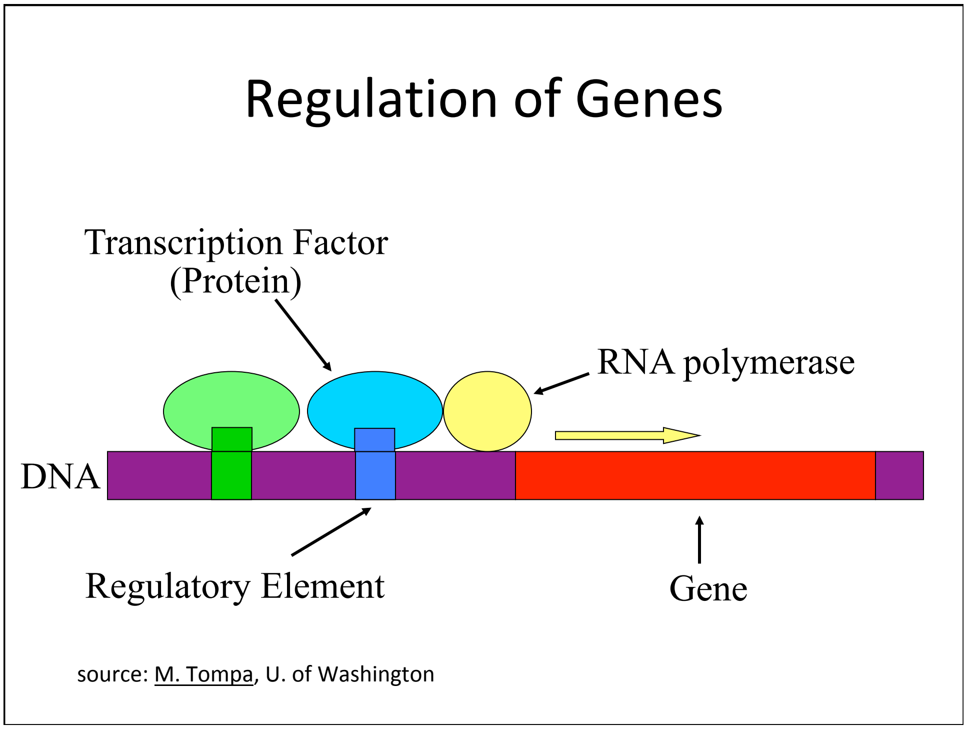
RNA polymerase
(Protein)

DNA

Regulatory Element

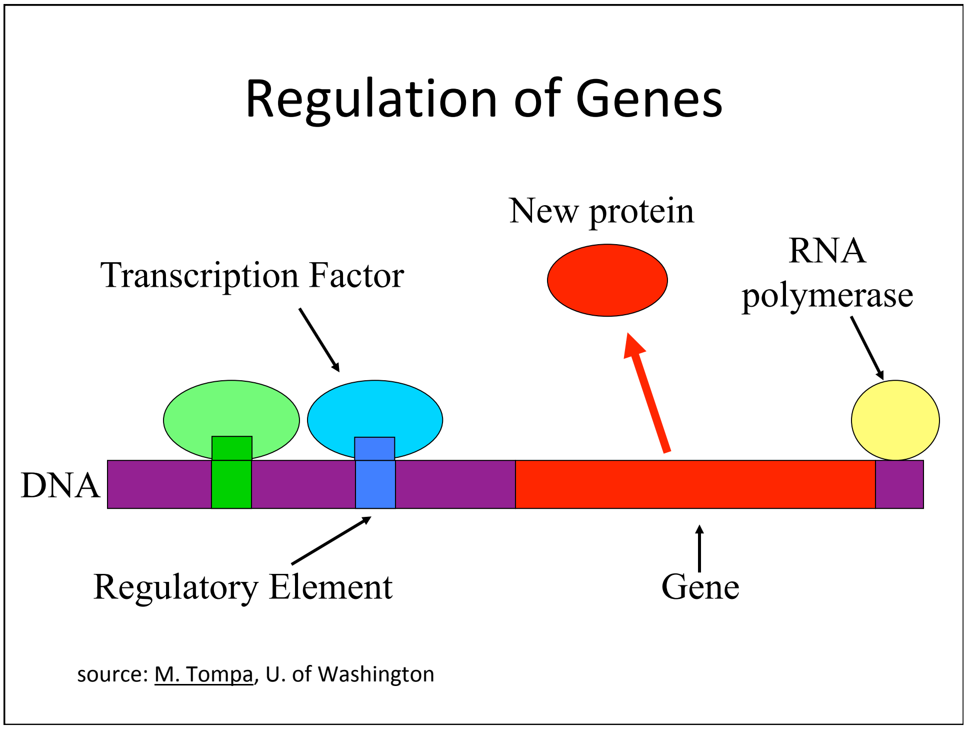
Gene

source: M. Tompa, U. of Washington



Regulation of Genes

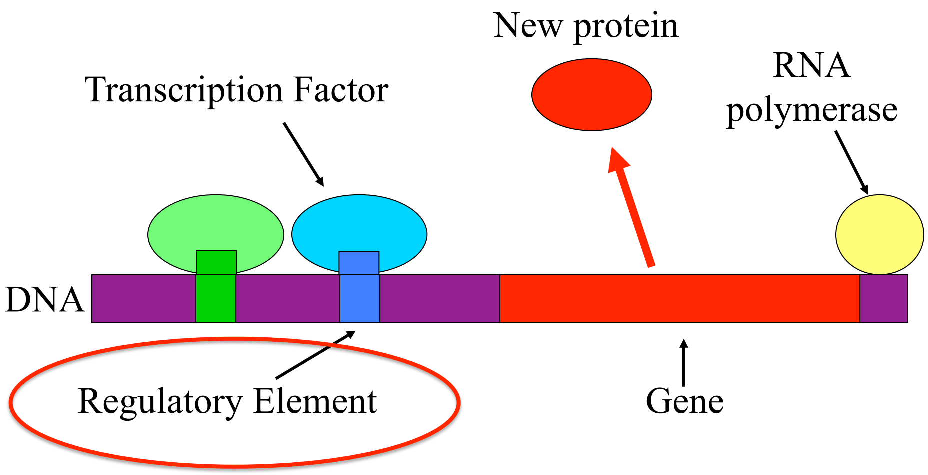
source: M. Tompa, U. of Washington



Regulation of Genes

source: M. Tompa, U. of Washington

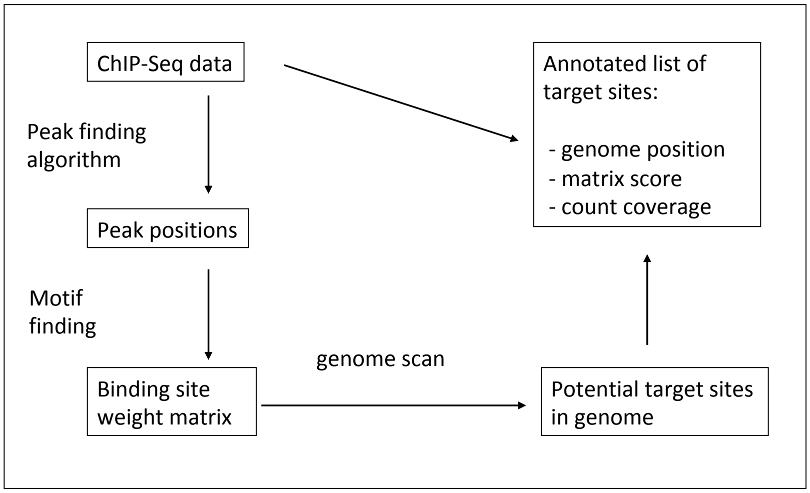
Regulation of Genes



source: M. Tompa, U. of Washington

Selective occupancy of genomic TF target sites

General data processing pipeline:



ChIP-Seq data

Peak finding algorithm

Peak positions

Motif finding

Binding site weight matrix

genome scan

Potential target sites in genome

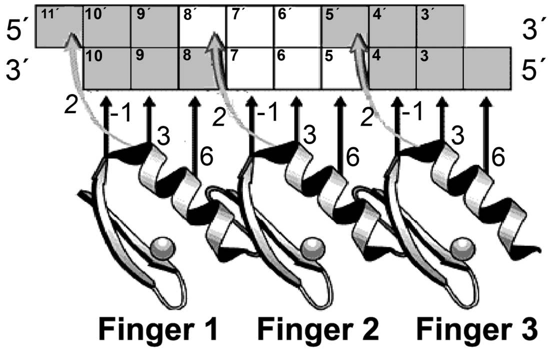
Annotated list of target sites:

- genome position
- matrix score
- count coverage

What is a motif?

- A subsequence that occurs in multiple sequences with a biological importance (e.g. at sites of protein binding).
- Motifs can be totally constant (substring) or have variable elements (pattern).
- Protein Motifs often result from structural features.
- DNA Motifs (regulatory elements)
  - Binding sites for proteins
  - Short sequences (5-25 bps)
  - A handful to millions of binding sites in the genome

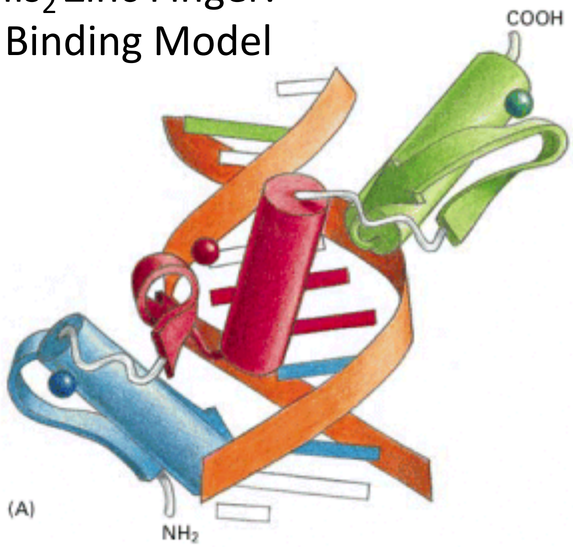
Cys$_2$His$_2$ Zinc Finger: Canonical DNA binding model



Residues at positions 6, 3, 2, and -1 (relative to the beginning of the a-helix) at each finger interact with adjacent nucleotides in the DNA molecule (interactions shown with arrows).

Kaplan. et al., PLoS Comput Biol, 2005

Cys$_2$His$_2$ Zinc Finger: DNA Binding Model



source: Molecular Biology of the Cell (4th ed.), A. Johnson, et al.

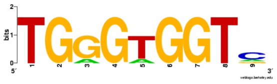
Sequence motif – a pattern of nucleotide or amino acid sequences

DNA motif:

```
GTATGTACTTACTATGGGTGGTCAACAAATCTATGTATGA
TAACATGTGACTCCTATAACCTCTTTGGGTGGTACATGAA
CTGGGAGGTCCTCGGTTCAGAGTCACAGAGCAGATAATCA
TTAGAGGCACAATTGCTTGGGTGGTGCACAAAAAAACAAG
AACAGCCTTGGATTAGCTGCTGGGGGGGTGAGTGGTCCAC
ATCAGAATGGGTGGTCCATATATCCCAAAGAAGAGGGTAG
```

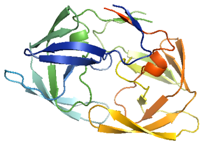
Transcription Factor Binding Sites (TFBS)

```
123456789
TGGGTGGTC
TGGGTGGTA
TGGGAGGTC
TGGGTGGTG
TGAGTGGTC
TGGGTGGTC
```



Protein motif:

```
DREDDPAfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPBfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
DREDDPCfly : MSASAIYRPFPKVKHFCIFPIAMAGSNLLIHLDTIDQNDLIYVERDMNFAQKVGLCFL-LYGDDHSDATYILQKLLAMTRSDFPQSDLLIKFAK
```



Motif Representations

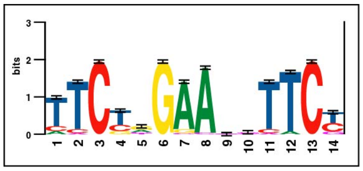
- Consensus sequence: a single string with the most likely sequence

  TTCTGGAACCTTCT

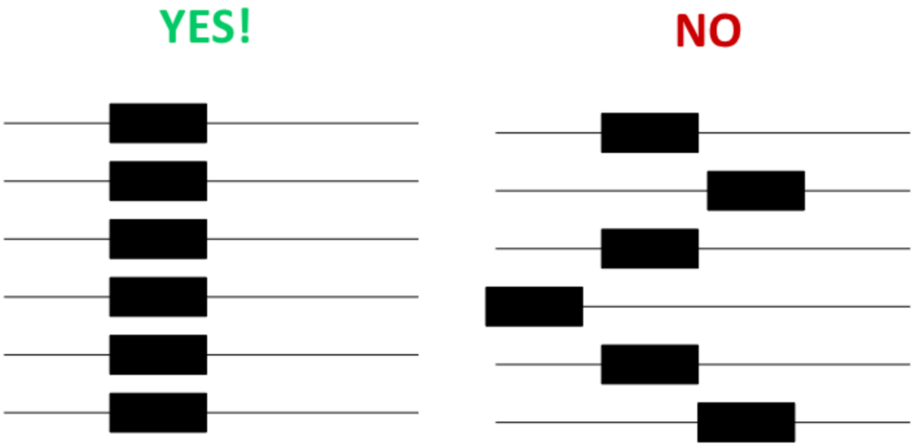
- Regular expression: a string with wildcards, constrained selection (+/- wildcards & ambiguity)

  YTCYXGAAXXTTCY

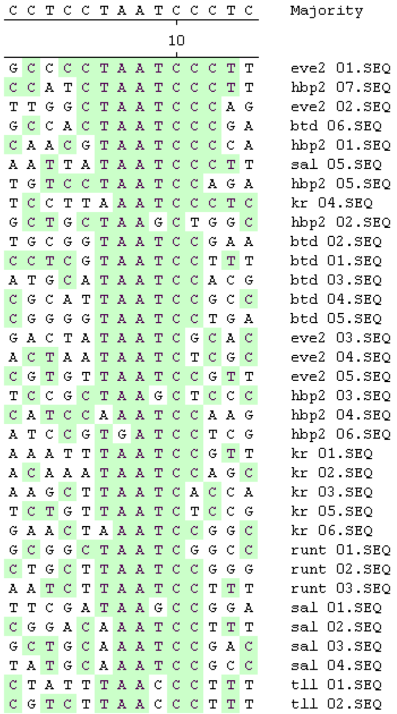
- Profile: a list of the letter frequencies at each position
- Sequence Logo:
  - graphical depiction of a profile
  - conservation of elements in a motif.



How do we describe a motif?



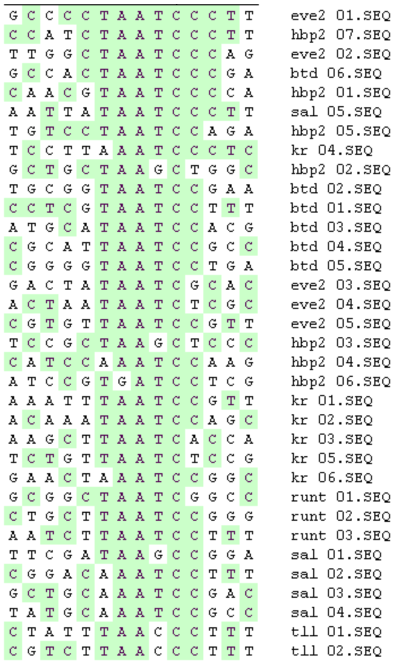
Given a multiple sequence alignment, it's trivial



```
C C T C C T A A T C C C T C    Majority
                  |
                 10
G C C C C T A A T C C C T T    eve2 01.SEQ
C C A T C T A A T C C C T T    hbp2 07.SEQ
T T G G C T A A T C C C A G    eve2 02.SEQ
G C C A C T A A T C C C G A    btd 06.SEQ
C A A C G T A A T C C C C A    hbp2 01.SEQ
A A T T A T A A T C C C T T    sal 05.SEQ
T G T C C T A A T C C A G A    hbp2 05.SEQ
T C C T T A A A T C C C T C    kr 04.SEQ
G C T G C T A A G C T G G C    hbp2 02.SEQ
T G C G G T A A T C C G A A    btd 02.SEQ
C C T C G T A A T C C T T T    btd 01.SEQ
A T G C A T A A T C C A C G    btd 03.SEQ
C G C A T T A A T C C G C C    btd 04.SEQ
C G G G G T A A T C C T G A    btd 05.SEQ
G A C T A T A A T C G C A C    eve2 03.SEQ
A C T A A T A A T C T C G C    eve2 04.SEQ
C G T G T T A A T C C G T T    eve2 05.SEQ
T C C G C T A A G C T C C C    hbp2 03.SEQ
C A T C C A A A T C C A A G    hbp2 04.SEQ
A T C C G T G A T C C T C G    hbp2 06.SEQ
A A A T T T A A T C C G T T    kr 01.SEQ
A C A A A T A A T C C A G C    kr 02.SEQ
A A G C T T A A T C A C C A    kr 03.SEQ
T C T G T T A A T C T C C G    kr 05.SEQ
G A A C T A A A T C C G G C    kr 06.SEQ
G C G G C T A A T C G G C C    runt 01.SEQ
C T G C T T A A T C C G G G    runt 02.SEQ
A A T C T T A A T C C T T T    runt 03.SEQ
T T C G A T A A G C C G G A    sal 01.SEQ
C G G A C A A A T C C T T T    sal 02.SEQ
G C T G C A A A T C C G A C    sal 03.SEQ
T A T G C A A A T C C G C C    sal 04.SEQ
C T A T T T A A C C C T T T    tll 01.SEQ
C G T C T T A A C C C T T T    tll 02.SEQ
```

T A A T C C C ← Motif

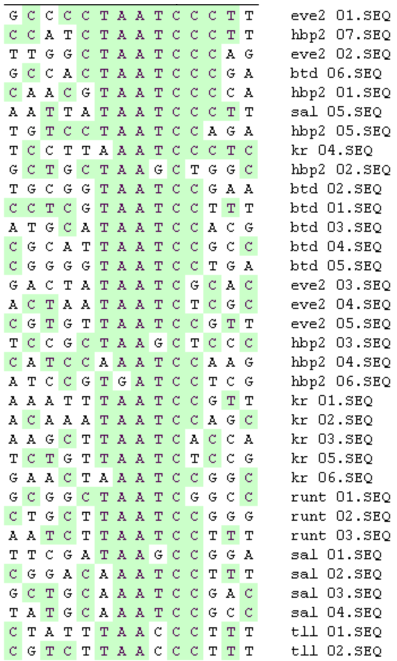
(“Consensus String”)



```
G C C C C T A A T C C C T T    eve2 01.SEQ
C C A T C T A A T C C C T T    hbp2 07.SEQ
T T G G C T A A T C C C A G    eve2 02.SEQ
G C C A C T A A T C C C G A    btd 06.SEQ
C A A C G T A A T C C C C A    hbp2 01.SEQ
A A T T A T A A T C C C T T    sal 05.SEQ
T G T C C T A A T C C A G A    hbp2 05.SEQ
T C C T T A A A T C C C T C    kr 04.SEQ
G C T G C T A A G C T G G C    hbp2 02.SEQ
T G C G G T A A T C C G A A    btd 02.SEQ
C C T C G T A A T C C T T T    btd 01.SEQ
A T G C A T A A T C C A C G    btd 03.SEQ
C G C A T T A A T C C G C C    btd 04.SEQ
C G G G G T A A T C C T G A    btd 05.SEQ
G A C T A T A A T C G C A C    eve2 03.SEQ
A C T A A T A A T C T C G C    eve2 04.SEQ
C G T G T T A A T C C G T T    eve2 05.SEQ
T C C G C T A A G C T C C C    hbp2 03.SEQ
C A T C C A A A T C C A A G    hbp2 04.SEQ
A T C C G T G A T C C T C G    hbp2 06.SEQ
A A A T T T A A T C C G T T    kr 01.SEQ
A C A A A T A A T C C A G C    kr 02.SEQ
A A G C T T A A T C A C C A    kr 03.SEQ
T C T G T T A A T C T C C G    kr 05.SEQ
G A A C T A A A T C C G G C    kr 06.SEQ
G C G G C T A A T C G G C C    runt 01.SEQ
C T G C T T A A T C C G G G    runt 02.SEQ
A A T C T T A A T C C T T T    runt 03.SEQ
T T C G A T A A G C C G G A    sal 01.SEQ
C G G A C A A A T C C T T T    sal 02.SEQ
G C T G C A A A T C C G A C    sal 03.SEQ
T A T G C A A A T C C G C C    sal 04.SEQ
C T A T T T A A C C C T T T    tll 01.SEQ
C G T C T T A A C C C T T T    tll 02.SEQ
```

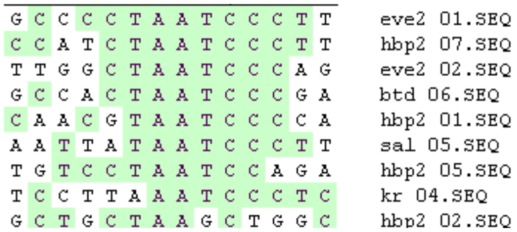
WAATCC**N** ← Motif (Regular Expression)

W = T or A
N = A,C,G,T



```
G C C C C T A A T C C C T T    eve2 01.SEQ
C C A T C T A A T C C C T T    hbp2 07.SEQ
T T G G C T A A T C C C A G    eve2 02.SEQ
G C C A C T A A T C C C G A    btd 06.SEQ
C A A C G T A A T C C C C A    hbp2 01.SEQ
A A T T A T A A T C C C T T    sal 05.SEQ
T G T C C T A A T C C A G A    hbp2 05.SEQ
T C C T T A A A T C C C T C    kr 04.SEQ
G C T G C T A A G C T G G C    hbp2 02.SEQ
T G C G G T A A T C C G A A    btd 02.SEQ
C C T C G T A A T C C T T T    btd 01.SEQ
A T G C A T A A T C C A C G    btd 03.SEQ
C G C A T T A A T C C G C C    btd 04.SEQ
C G G G G T A A T C C T G A    btd 05.SEQ
G A C T A T A A T C G C A C    eve2 03.SEQ
A C T A A T A A T C T C G C    eve2 04.SEQ
C G T G T T A A T C C G T T    eve2 05.SEQ
T C C G C T A A G C T C C C    hbp2 03.SEQ
C A T C C A A A T C C A A G    hbp2 04.SEQ
A T C C G T G A T C C T C G    hbp2 06.SEQ
A A A T T T A A T C C G T T    kr 01.SEQ
A C A A A T A A T C C A G C    kr 02.SEQ
A A G C T T A A T C A C C A    kr 03.SEQ
T C T G T T A A T C T C C G    kr 05.SEQ
G A A C T A A A T C C G G C    kr 06.SEQ
G C G G C T A A T C G G C C    runt 01.SEQ
C T G C T T A A T C C G G G    runt 02.SEQ
A A T C T T A A T C C T T T    runt 03.SEQ
T T C G A T A A G C C G G A    sal 01.SEQ
C G G A C A A A T C C T T T    sal 02.SEQ
G C T G C A A A T C C G A C    sal 03.SEQ
T A T G C A A A T C C G C C    sal 04.SEQ
C T A T T T A A C C C T T T    tll 01.SEQ
C G T C T T A A C C C T T T    tll 02.SEQ
```

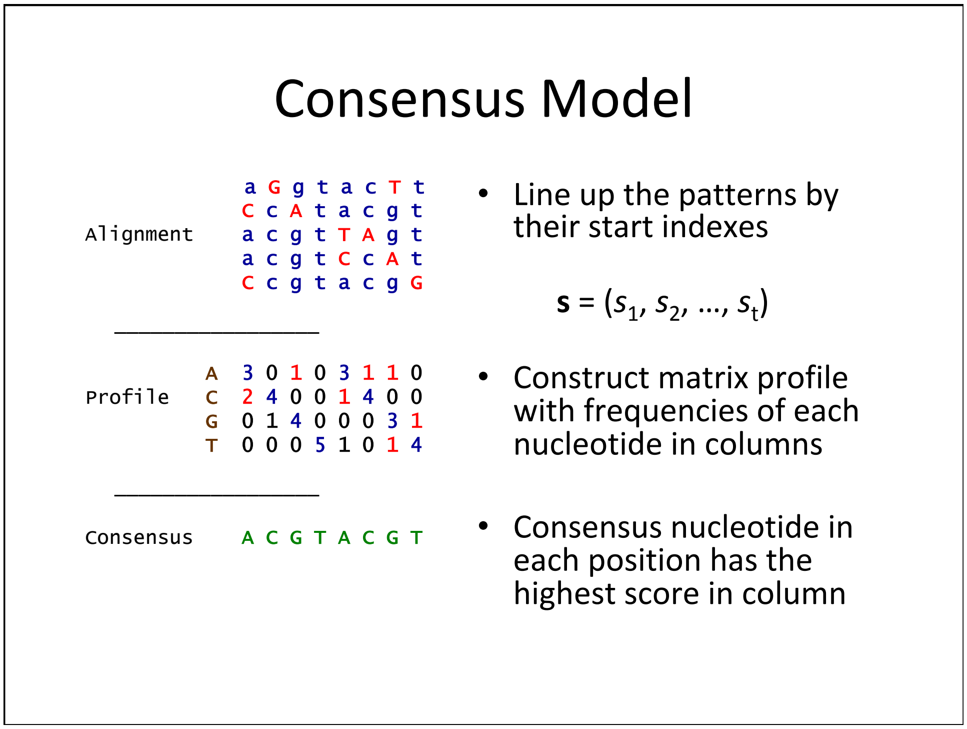
Alternative way to represent motif



```
G C C C C T A A T C C C T T    eve2 01.SEQ
C C A T C T A A T C C C T T    hbp2 07.SEQ
T T G G C T A A T C C C A G    eve2 02.SEQ
G C C A C T A A T C C C G A    btd 06.SEQ
C A A C G T A A T C C C C A    hbp2 01.SEQ
A A T T A T A A T C C C T T    sal 05.SEQ
T G T C C T A A T C C A G A    hbp2 05.SEQ
T C C T T A A A T C C C T C    kr 04.SEQ
G C T G C T A A G C T G G C    hbp2 02.SEQ
```

1	1	9	9	0	0	0	1	A
6	0	0	0	0	9	8	7	**C**
1	0	0	0	1	0	0	1	**G**
1	8	0	0	8	0	1	0	**T**

Position weight matrix (PWM)
Or simply, "weight matrix"
Sometimes also referred to as Position Specific Scoring Matrix (PSSM)
(in this instance it's a frequency matrix!)



Consensus Model

Alignment

```
a G g t a c T t
C c A t a c g t
a c g t T A g t
a c g t C c A t
C c g t a c g G
```

Profile

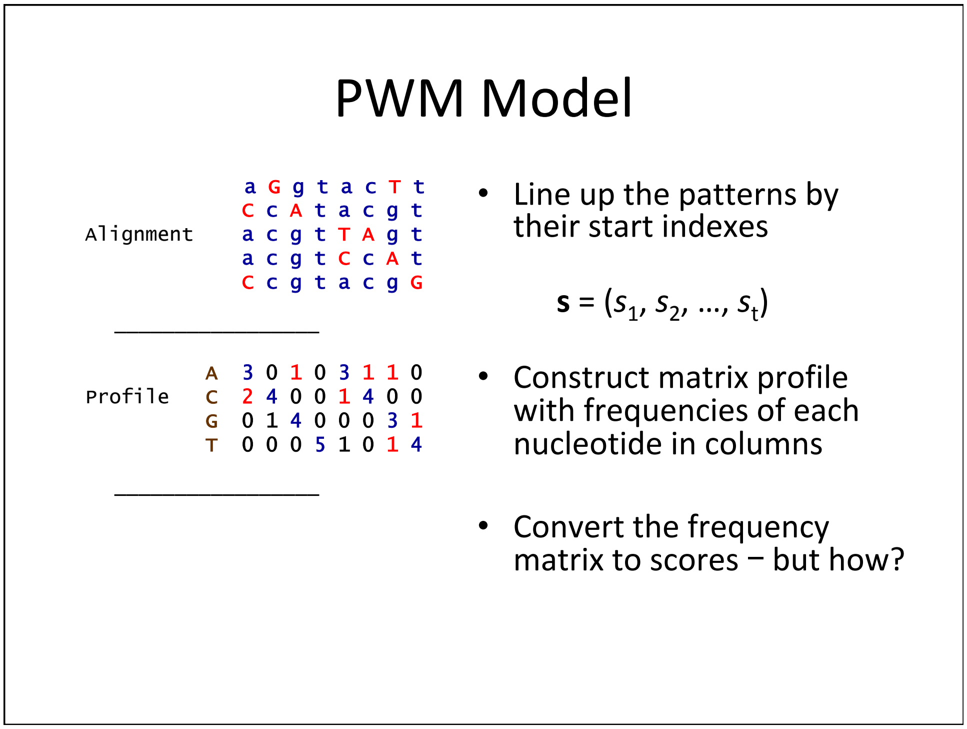
A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

Consensus A C G T A C G T

- Line up the patterns by their start indexes

  $\mathbf{s} = (s_1, s_2, \ldots, s_t)$

- Construct matrix profile with frequencies of each nucleotide in columns
- Consensus nucleotide in each position has the highest score in column



PWM Model

Alignment

```
a G g t a c T t
C c A t a c g t
a c g t T A g t
a c g t C c A t
C c g t a c g G
```

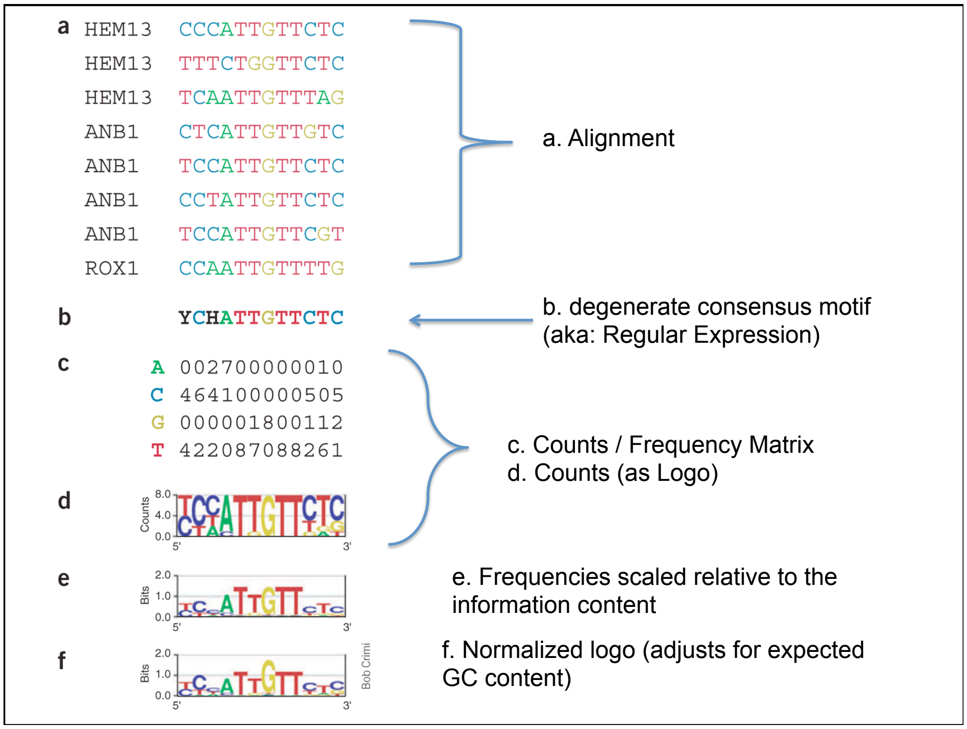
Profile

A	3	0	1	0	3	1	1	0
C	2	4	0	0	1	4	0	0
G	0	1	4	0	0	0	3	1
T	0	0	0	5	1	0	1	4

- Line up the patterns by their start indexes

  $\mathbf{s} = (s_1, s_2, \ldots, s_t)$

- Construct matrix profile with frequencies of each nucleotide in columns
- Convert the frequency matrix to scores – but how?



a

HEM13	CCCATTGTTCTC
HEM13	TTTCTGGTTCTC
HEM13	TCAATTGTTTAG
ANB1	CTCATTGTTGTC
ANB1	TCCATTGTTCTC
ANB1	CCTATTGTTCTC
ANB1	TCCATTGTTCGT
ROX1	CCAATTGTTTTG

a. Alignment

b YCHATTGTTCTC

b. degenerate consensus motif (aka: Regular Expression)

c

```
A 002700000010
C 464100000505
G 000001800112
T 422087088261
```

d

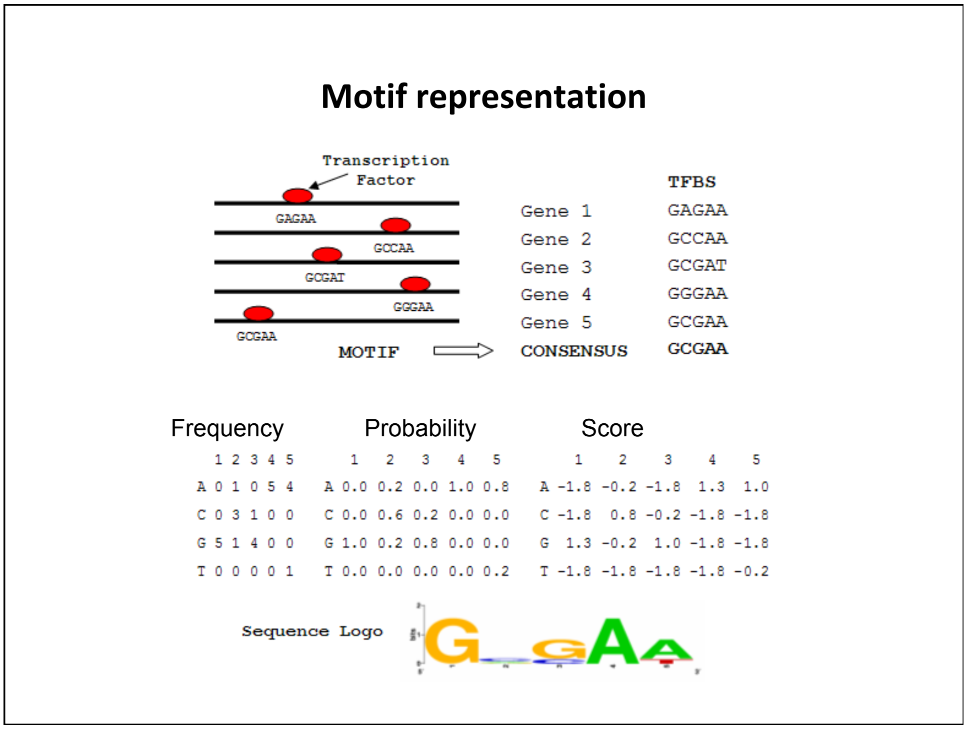
c. Counts / Frequency Matrix
d. Counts (as Logo)

e

e. Frequencies scaled relative to the information content

f

f. Normalized logo (adjusts for expected GC content)



Motif representation

	TFBS
Gene 1	GAGAA
Gene 2	GCCAA
Gene 3	GCGAT
Gene 4	GGGAA
Gene 5	GCGAA
CONSENSUS	GCGAA

Frequency

	1	2	3	4	5
A	0	1	0	5	4
C	0	3	1	0	0
G	5	1	4	0	0
T	0	0	0	0	1

Probability

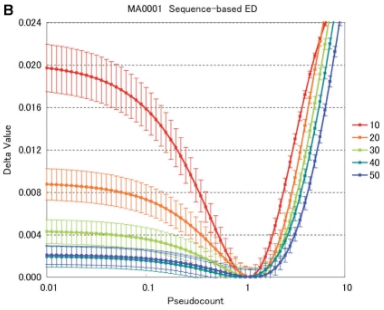
	1	2	3	4	5
A	0.0	0.2	0.0	1.0	0.8
C	0.0	0.6	0.2	0.0	0.0
G	1.0	0.2	0.8	0.0	0.0
T	0.0	0.0	0.0	0.0	0.2

Score

	1	2	3	4	5
A	-1.8	-0.2	-1.8	1.3	1.0
C	-1.8	0.8	-0.2	-1.8	-1.8
G	1.3	-0.2	1.0	-1.8	-1.8
T	-1.8	-1.8	-1.8	-1.8	-0.2

Sequence Logo

To avoid bias due to this small sample size, a certain numeric value, called a pseudocount, is usually allocated for each position, and its fraction according to the background base composition is added to each element.



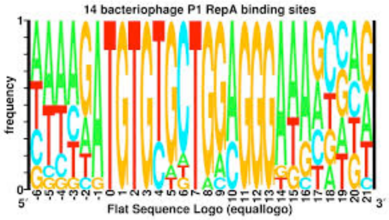
Avoids division by zero.

As more data is available, pseudocounts are overwhelmed and have negligible effect.

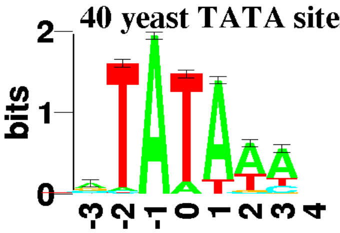
Effectively this is another form of a ‘prior’.

Motif Logos

How to create a motif model?


14 bacteriophage P1 RepA binding sites
frequency
1
0
5'
3'
Flat Sequence Logo (equallogo)

X-axis: position in sequence (assumes position independence)

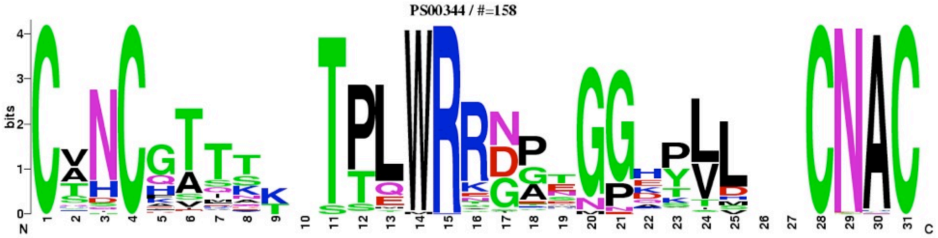


Y-axis: typically “bits” (occasionally frequency or probability)

assume statistical independence between positions in the pattern

Height of letter ≈ fraction of time that letter is observed at that position.

(Height of all the letters in a column ≈ to how conserved the column is)



Motif Position

What is a “bit” in this context?

Each position scaled with the information content of the base frequencies at that position:

$$I_i = 2 + \sum_b f_{b,i} \log_2 f_{b,i}$$

Positions are calculated using log likelihoods, the score of a sequence can be calculated by adding (rather than multiplying as is necessary with probabilities).

Perfect conservation = 2 bits
Every base equal = 0 bits

Assumes all four bases occur equally likely.

What is a “bit” in this context?

Better measure adjust for background bias:

$$I_{seq}(i) = -\sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

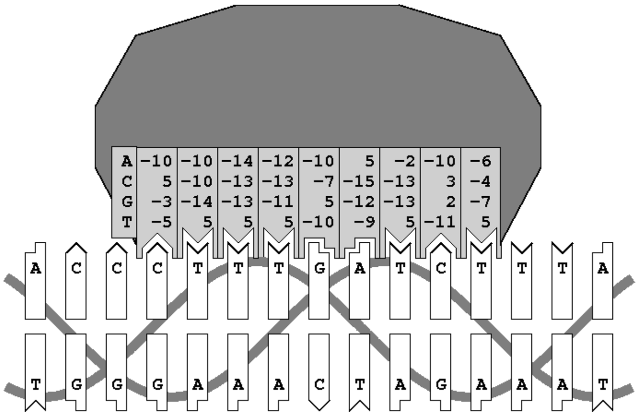
Aka. “relative entropy”

Notice its equivalent to a log-likelihood ratio!

Play with Logos:

Steven Brenner's WebLogo (http://weblogo.berkeley.edu/)
enoLOGOS3 (http://biodev.hgen.pitt.edu/enologos)

Physical interpretation of an weight matrix



A	-10	-10	-14	-12	-10	5	-2	-10	-6
C	5	-10	-13	-13	-7	-15	-13	3	-4
G	-3	-14	-13	-11	5	-12	-13	2	-7
T	-5	5	5	5	-10	-9	5	-11	5

Weight matrix elements represent relative binding energies between DNA base-pairs and protein surface areas (base-pair acceptor sites).

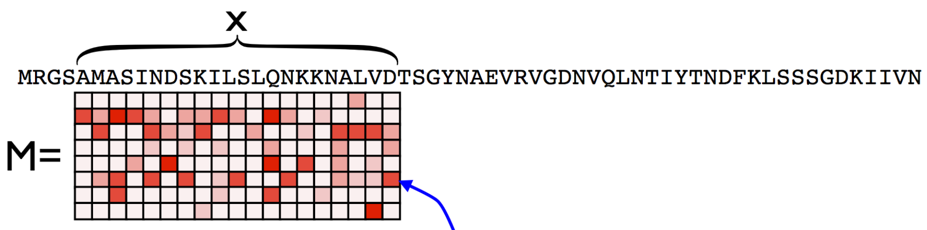
A weight matrix column describes the base preferences of a base-pair acceptor site.

Represents sequence of one strand, though physical interactions may be with other strand.

Motif scanning

- Given a motif (e.g., consensus string, regular expression, or weight matrix), find the binding sites in an input sequence (e.g. a genome)

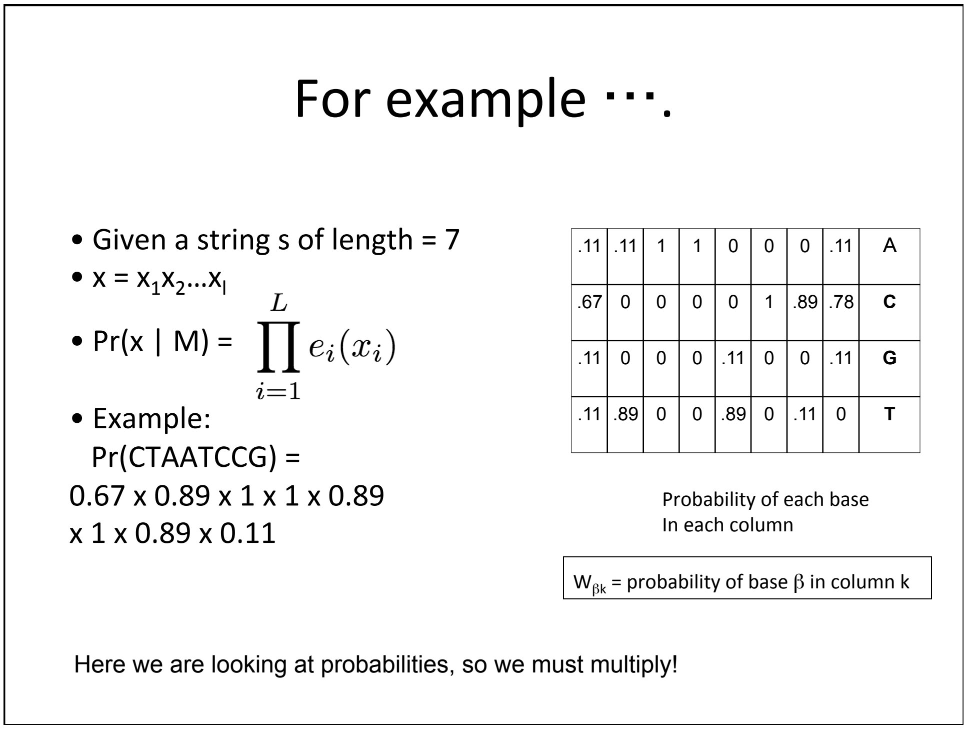
- For consensus string, problem is trivial
  - For each position in input sequence, check if substring starting at position *I* matches the motif.
- For weight matrix, not quite so trivial


x
MRGSAMASINDSKILSLQNKKNALVDTSGYNAEVRVGDNVQLNTIYTNDFKLSSSGDKIIVN
M=

Color ≈ Probability that the i^th position has the given amino acid = e_i(x).

$$\text{Score}(x) = \Pr(x \mid M) = \prod_{i=1}^{L} e_i(x_i)$$

Score of a string according to profile M = Product of the probabilities you would observe the given letters.



For example ….

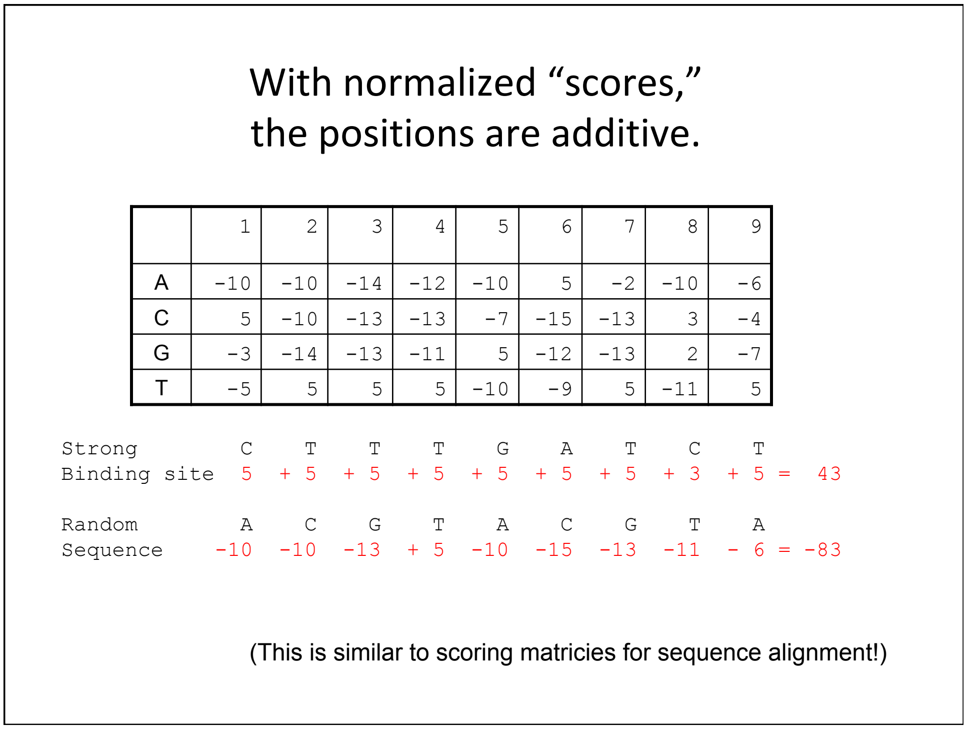
- Given a string s of length = 7
- x = $x_1x_2\ldots x_l$
- Pr(x | M) = $\prod_{i=1}^{L} e_i(x_i)$
- Example:
  Pr(CTAATCCG) =
  0.67 x 0.89 x 1 x 1 x 0.89
  x 1 x 0.89 x 0.11

.11	.11	1	1	0	0	0	.11	**A**
.67	0	0	0	0	1	.89	.78	**C**
.11	0	0	0	.11	0	0	.11	**G**
.11	.89	0	0	.89	0	.11	0	**T**

Probability of each base
In each column

$W_{\beta k}$ = probability of base β in column k

Here we are looking at probabilities, so we must multiply!



With normalized "scores," the positions are additive.

	1	2	3	4	5	6	7	8	9
A	-10	-10	-14	-12	-10	5	-2	-10	-6
C	5	-10	-13	-13	-7	-15	-13	3	-4
G	-3	-14	-13	-11	5	-12	-13	2	-7
T	-5	5	5	5	-10	-9	5	-11	5

```
Strong             C     T     T     T     G     A     T     C     T
Binding site       5   + 5   + 5   + 5   + 5   + 5   + 5   + 3   + 5 =   43

Random             A     C     G     T     A     C     G     T     A
Sequence         -10   -10   -13   + 5   -10   -15   -13   -11   - 6 = -83
```

(This is similar to scoring matricies for sequence alignment!)

Binding sites from a weight matrix motif

- Given sequence S (e.g., 1000 base-pairs long)
- For each substring x of S,
  - Compute Pr(x|M)
  - If Pr(x|M) > some threshold, call that a binding site
- Look at S, *as well as its reverse complement*

But what threshold?

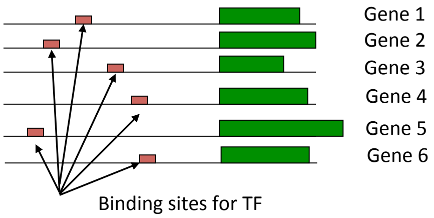
- In reality, every protein binds to every sequence with at least ***some*** affinity. But poorer matches to matrix are weaker (i.e. less frequent and transient) binding.
- Therefore, cutoff is related to how "strong" a site you are looking for … depends on protein (TF), its concentration and the question at hand.
- However, empirically 60% of max is often used as a cutoff. This is another arbitrary but commonly used threshold (like 0.05 as a p-value cutoff).

A couple things to ponder …

- Note how the scoring scheme used for motifs resembles the scoring schemes we discussed for alignment, only now they are position dependent.
- Also notice that a position specific scoring matrix is essentially a very simple HMM.

We will leverage "patterns" to identify transcription factor motifs.

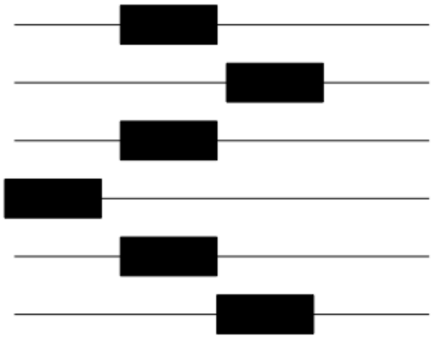
- Say a transcription factor (TF) controls six different genes
- Each of the six genes will have binding site(s) for the TF in their promoter region



The motif finding problem

- Now suppose we are given the regions of six genes G1, G2, … G6 identified by ChIP as bound by TF.

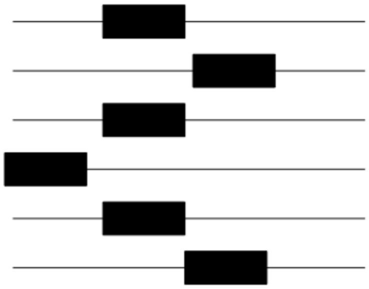
- Can we find the binding sites of TF, without knowing about them *a priori* ?
  - Binding sites are similar to each other, but not necessarily identical (e.g. we expect a PWM)



Identifying Motifs: Complications

- We do not know the motif sequence

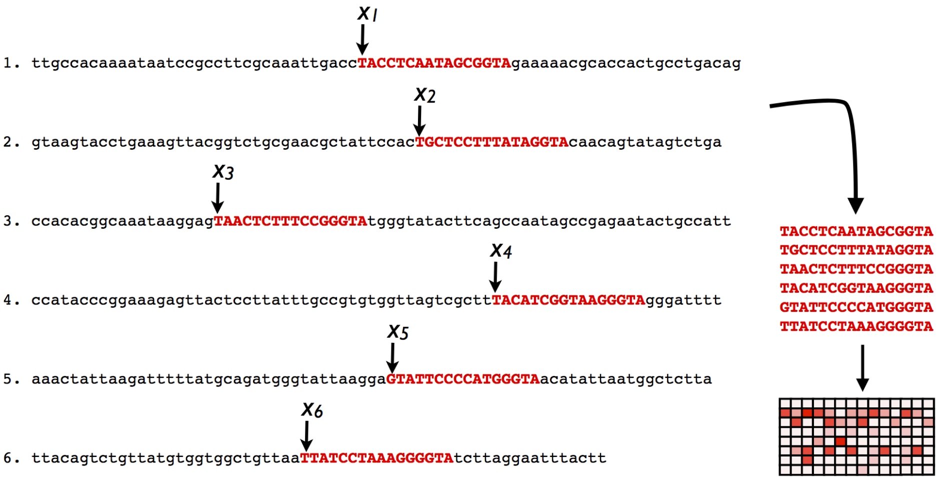
- We do not know where it is located within the sequence fragment



- Motifs can differ slightly from one gene to another

- How to discern it from “random” sequence?

If we knew the starting point of the motif in each sequence, we could construct a Sequence Profile (PSSM) for the motif:



1. ttgccacaaaataatccgccttcgcaaattgaccTACCTCAATAGCGGTAgaaaaacgcaccactgcctgacag

2. gtaagtacctgaaagttacggtctgcgaacgctattccacTGCTCCTTTATAGGTAcaacagtatagtctga

3. ccacacggcaaataaggagTAACTCTTTCCGGGTAtgggtatacttcagccaatagccgagaatactgccatt

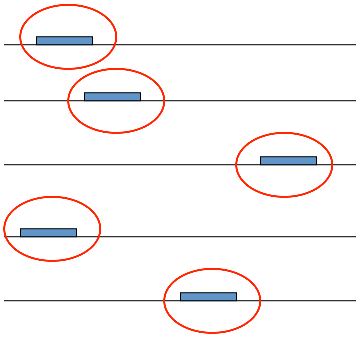
4. ccatacccggaaagagttactccttatttgccgtgtggttagtcgcttTACATCGGTAAGGGTAgggatttt

5. aaactattaagatttttatgcagatgggtattaaggaGTATTCCCCATGGGTAacatattaatggctctta

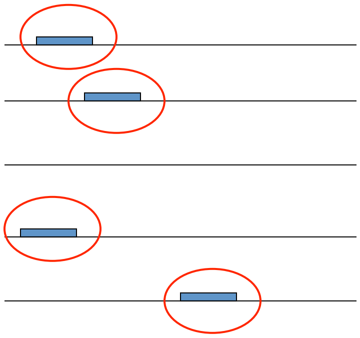
6. ttacagtctgttatgtggtggctgttaaTTATCCTAAAGGGGTAtcttaggaatttactt

TACCTCAATAGCGGTA
TGCTCCTTTATAGGTA
TAACTCTTTCCGGGTA
TACATCGGTAAGGGTA
GTATTCCCCATGGGTA
TTATCCTAAAGGGGTA

Gibbs sampling: basic idea

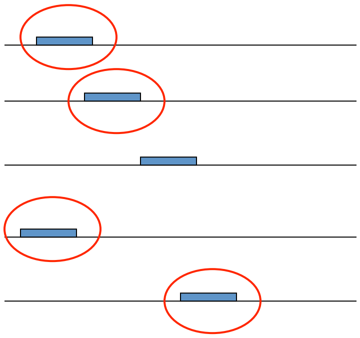


Gibbs sampling: basic idea



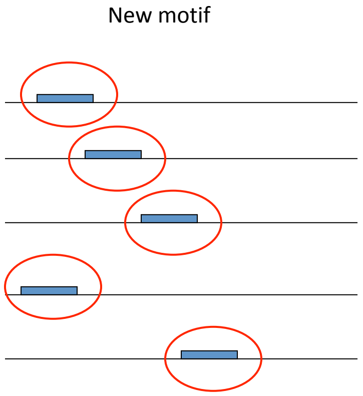
Delete one substring

Gibbs sampling: basic idea



Try a replacement:
Compute its score,
Accept the replacement
depending on the score.

Gibbs sampling: basic idea



Not guaranteed to find “best motif”.

Works well in practice.

Best if you “restart” process many times (resample).

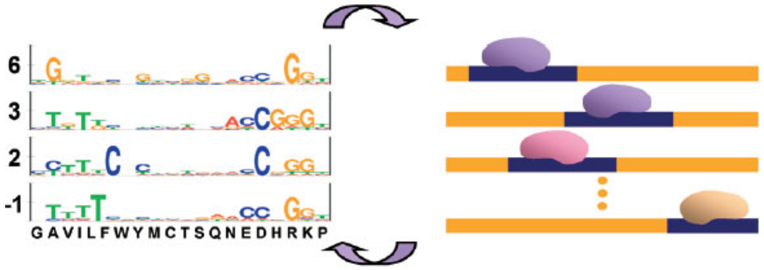
Expectation Maximization

- Popular algorithm for motif discovery
- Motif model: Position Weight Matrix
- Local search algorithm
  - Move from current choice of motif to a new similar motif, so as to improve the score
  - Keep doing this until no more improvement is obtained : Convergence to local optima

Estimating DNA Recognition Preferences

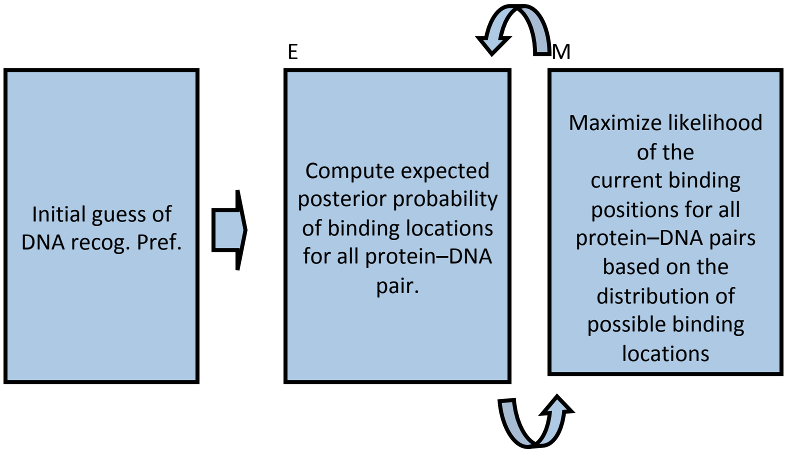
- Apply Expectation Maximization

Identify binding locations



Optimize recognition preferences

Expectation Maximization algorithm



E

M

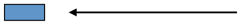
Initial guess of DNA recog. Pref.

Compute expected posterior probability of binding locations for all protein–DNA pair.

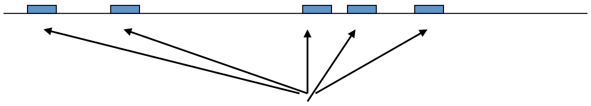
Maximize likelihood of the current binding positions for all protein–DNA pairs based on the distribution of possible binding locations

Basic idea of iteration

PWM

1.  Current motif

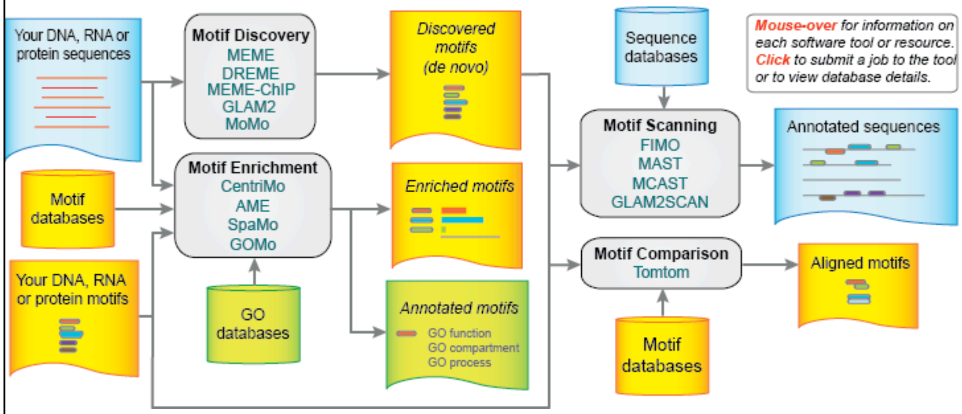
2. Scan sequence(s) for good matches to the current motif.



3. Build a new PWM out of these matches, and make it the new motif

MEME

Multiple Em for Motif Elicitation



Your DNA, RNA or protein sequences

Motif databases

Your DNA, RNA or protein motifs

Motif Discovery
MEME
DREME
MEME-ChIP
GLAM2
MoMo

Motif Enrichment
CentriMo
AME
SpaMo
GOMo

GO databases

Discovered motifs (de novo)

Enriched motifs

Annotated motifs
GO function
GO compartment
GO process

Sequence databases

Motif Scanning
FIMO
MAST
MCAST
GLAM2SCAN

Motif Comparison
Tomtom

Motif databases

Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.

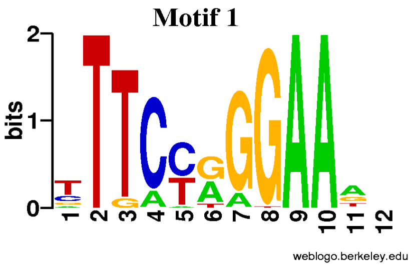
Annotated sequences

Aligned motifs

STAT1 Sequence Motif Defined by ChIP-Seq data

Input:
4446 ChIP peak regions,
200 bp

→



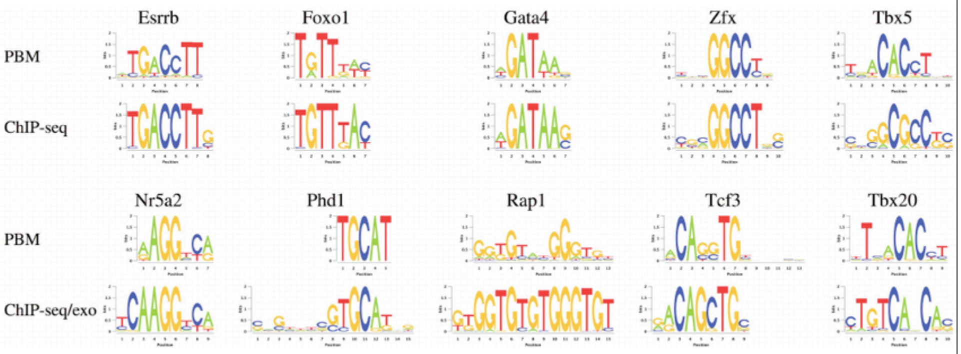
Matrix from experimental *in vivo* sites

	-7	-6	-5	-4	-3	-2	-1	0
A	23	38	15	0	2	29	8	19
C	33	17	13	0	0	67	56	31
G	35	17	12	4	2	4	15	31
T	10	27	60	96	96	0	21	19

Matrix from SELEX

	-7	-6	-5	-4	-3	-2	-1	0
A	6	62	26	2	2	5	2	2
C	57	13	27	2	3	89	95	48
G	23	14	10	2	2	2	2	49
T	14	11	36	95	93	4	2	2

Protein binding microarrays are another method of determining motifs.

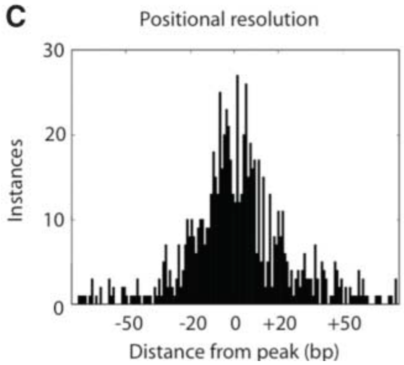


Motif Databases

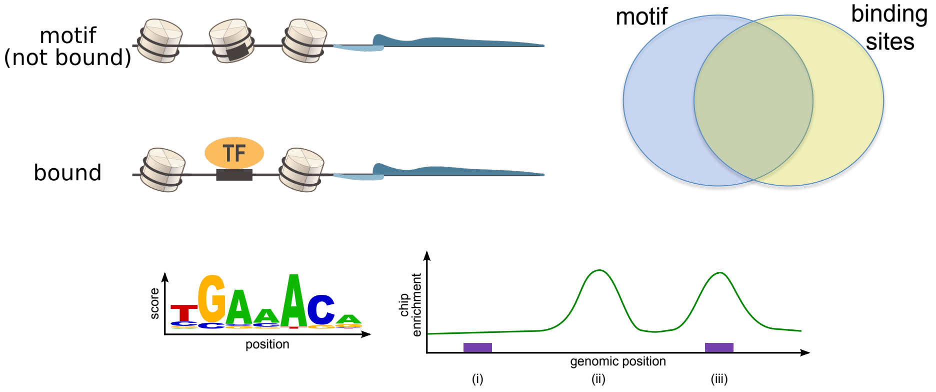
TRANSFAC®

Precision of ChIPSeq

- Evaluated against the center of high-scoring canonical motifs.
- 94% of these strong motifs fall *within 50bp* of the called experimental peak.



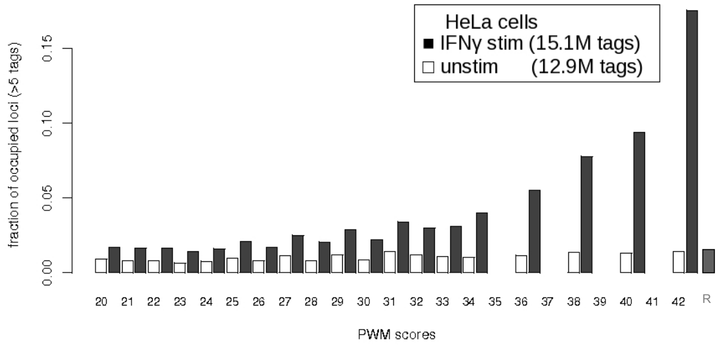
But .. only a fraction of motifs are bound



Dowell (2010)

Binding influenced by nucleosomes …
e.g. competition!

But higher quality motif sites (better matches to PSSM) are *more likely* to be bound.



STAT1 matrix

A:	0	-12	-12	-1	-6	0	0	-8	5	5	2
C:	0	-9	-10	5	4	0	-13	-12	-4	-13	-2
G:	-2	-13	-4	-12	-13	0	4	5	-10	-9	0
T:	2	5	5	-8	0	0	-6	-1	-12	-12	0