7.2 Transcription factor binding sites (TFBSs)

TFBS: DNA region that forms a **specific physical contact** with a particular TF.

TFBS are usually between 8 and 20 bp long
and contain a 5-8 bp long **core region** of well-conserved nucleotide bases.

Most TFs bind in the **major groove** of double-stranded DNA,
the others bind in the minor groove.

The **periodicity** of double-standed DNA is around 10 bp.
Thus, the core regions of TFBS are a bit longer than half a turn of dsDNA.

TFs may recognize DNA sequences that are similar,
but not identical, differing by a few nucleotides.

Sequence logos represent binding motifs

A **logo** represents each column of the alignment by a stack of letters.

The height of each letter is proportional to the **observed frequency** of the corresponding amino acid or nucleotide.

The overall height of each stack is proportional to the **sequence conservation** at that position.

Sequence conservation is defined as difference between the maximum possible entropy and the entropy of the observed symbol distribution:

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

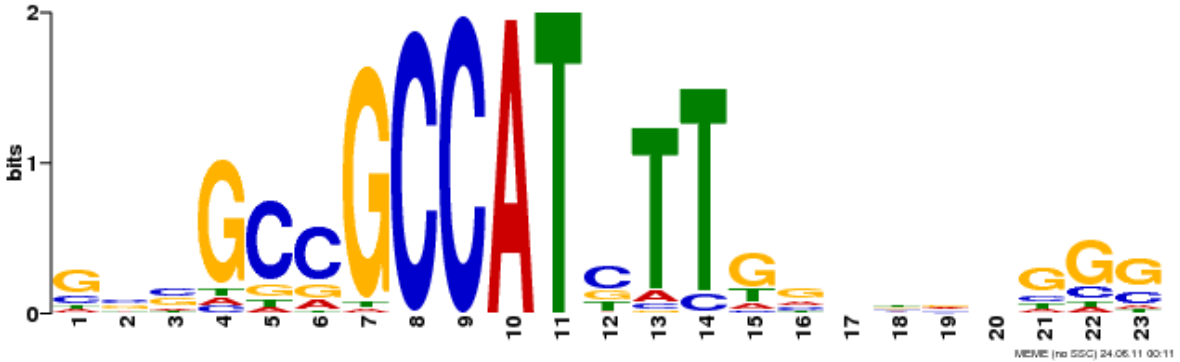
p_n : observed frequency of symbol n at a particular sequence position

N : number of distinct symbols for the given sequence type, either 4 for DNA/RNA or 20 for protein.

Crooks et al., Genome Research 14:1188–1190 (2004)

YY1 sequence logo

Sequence-logos are a convenient way to visualize the degree of degeneracy in the TFBS.



$$H_i = -\sum_{b=A}^{T} f_{b,i} \times log_2 f_{b,i}$$

$$R_i = log_2(4) - (H_i + e_n)$$

$$e_n = \frac{1}{ln2} \times \frac{s-1}{2n}$$

Sequence logo for the DNA binding motif that the TF YY1 (Yin Yang 1) binds to.

The motif was derived from the top 500 TF ChIP-seq peaks by the ENCODE consortium. For YY1, 468 out of 500 sequences contained this motif.

H_i : uncertainty (Shannon entropy) of position *i*

R_i : information content (y-axis) of position *i*

e_n : small-sample correction, s = 4 for nucleotides, n : number of sequences

Figure from Factorbook repository (Wang *et al.* 2013).

YY1 binding motifs

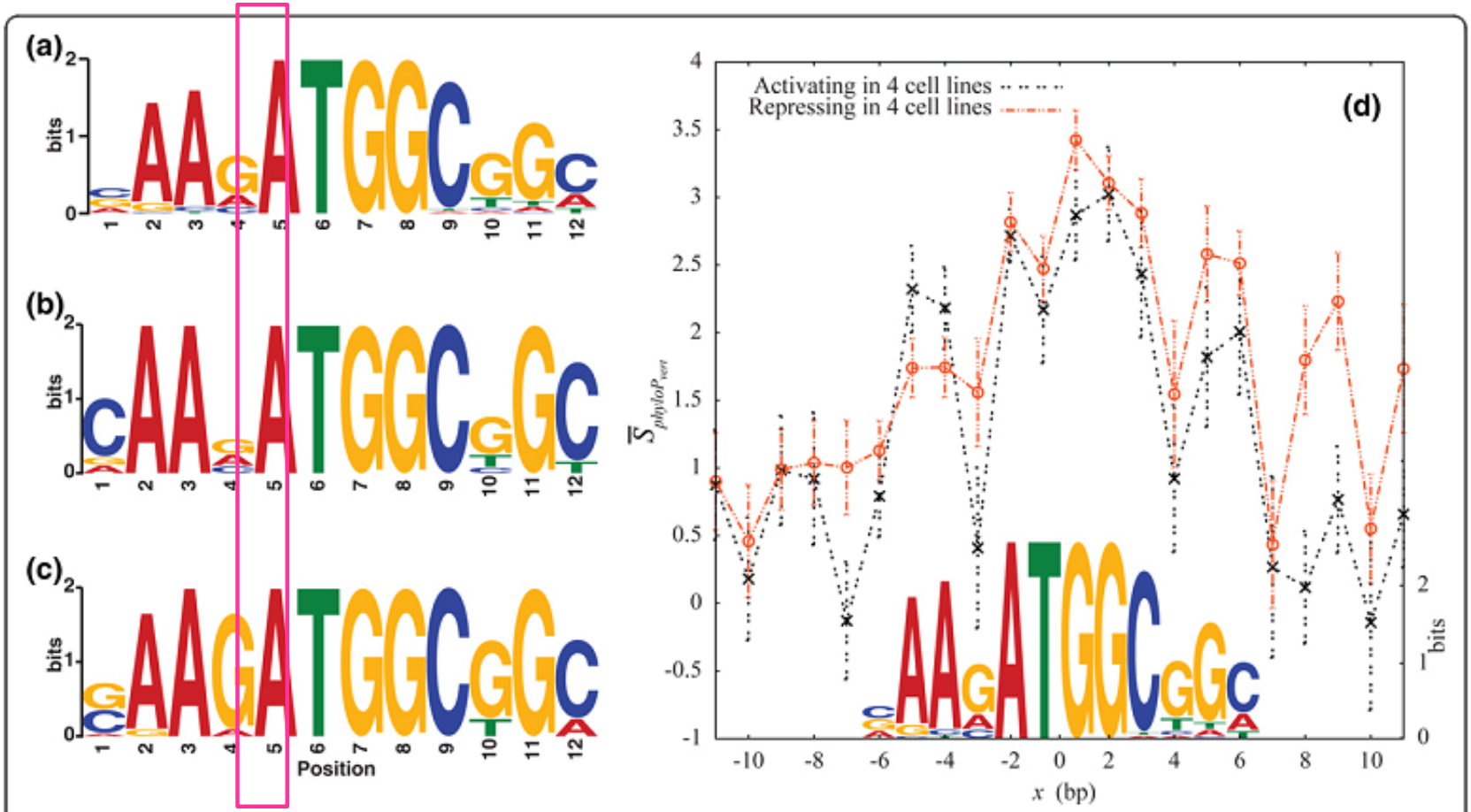
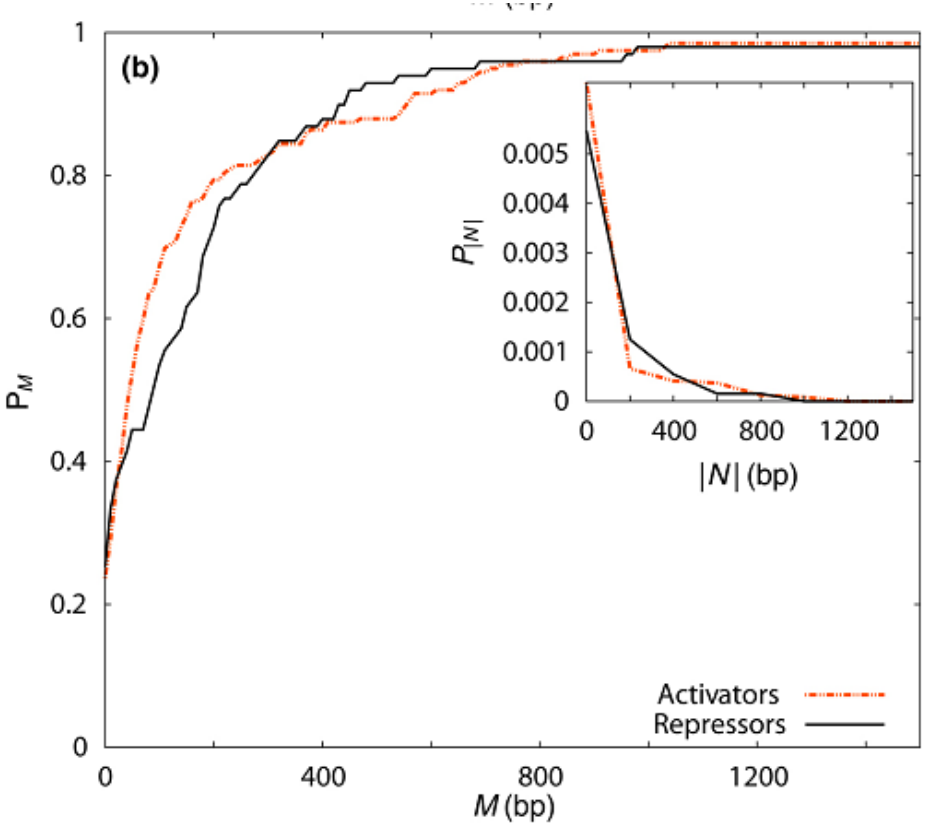


Figure 2 Characterization of functional YY1 binding sites. Sequence logo [102] for YY1 binding sites from **(a)** PWM and sites that are functionally **(b)** ubiquitously activating (9 BS) or **(c)** ubiquitously repressive (16 BS) in four human cell lines. In **(d)**, we plot the mean vertebrate phyloP conservation score [90] around functional YY1 binding sites. The mean score, $\bar{S}_{phyloP_{vert}}$, was computed at each base for sites where the binding event ubiquitously activated (black line) or repressed (red line) transcription in all four cell lines. The position weight matrix that was used to predict YY1 binding sites is shown (scale on the right axis).

No noticeable difference in binding motifs of activated or repressed target genes.

Whitfield et al. Genome Biology 2012, 13:R50

Where are TFBS relative to the TSS?



Inset: probability to find binding site at position *N* from transcriptional start site (TSS)

Main plot: cumulative distribution.

Activating TF binding sites are closer to the TSS than repressing TF binding sites ($p = 4.7 \times 10^{-2}$).

Whitfield et al. Genome Biology 2012, 13:R50

7.4. Position-specific scoring matrix

PSSMs are used to represent motifs (patterns) in biological sequences.

	Position 1	Position 2	Position 3	Position 4
Sequence 1	A	C	A	T
Sequence 2	A	C	C	T
Sequence 3	A	G	G	G
Sequence 4	C	C	T	G
Sequence 5	A	T	A	G
Sequence 6	C	A	G	T

Toy example of six DNA sequences that are 4 bp long.

	Position 1	Position 2	Position 3	Position 4
Frequency A	4	1	2	0
Frequency C	2	3	1	0
Frequency G	0	1	2	3
Frequency T	0	1	1	3

Frequency n_i^j of nucleotide bases *(i)* at the 4 positions *(j)*.

Out of 6 × 4 = 24 nucleotides in the four sequences, 7 are adenine, 6 are cytosine, 6 are guanine, and 5 are thymine. Thus, the frequencies p_i of the four nucleotides are 0.29 (A), 0.25 (C and G), and 0.21 (T).

7.4. Position-specific scoring matrix

From the frequency matrix, one computes the score matrix using

$$s_i^j = ln \frac{\left(n_i^j + p_i\right)/(N+1)}{p_i},$$

where, N is the number of considered sequences (here, $N = 6$).

	Position 1	Position 2	Position 3	Position 4
score A	0.75	-0.45	0.12	-1.94
score C	0.25	0.62	-0.34	-1.94
score G	-1.94	-0.34	0.25	0.62
score T	-1.94	-0.19	-0.19	0.78

Adding the frequencies p_i in the denominator and dividing by $N + 1$ avoids problematic cases with $n_i^j = 0$ where the logarithm would not be defined otherwise.

Positions with score $s_i^j = 0$ occur at the frequency that is expected randomly, positive entries denote enriched nucleotides at this position, negative entries denote the opposite case.

7.5 Binding free energy models

The binding of a TF to single- or double-stranded DNA is an elementary biomolecular association reaction.

The binding free energy model of Djordjevic (2003) describes the reversible binding of a TF to a short piece of DNA with sequence S,

$$TF + DNA \underset{k_{diss}}{\overset{k_{bind}}{\rightleftarrows}} TF - DNA$$

with the sequence-dependent rate constants k_{bind} and k_{diss} for TF binding and dissociation, respectively.

In equilibrium, $[TF] \cdot [S] \cdot k_{bind}(S) = [TF{:}S] \cdot k_{diss}(S)$

The ratio of the bound and free forms thus equals the ratio of the two rate constants and is equal to $\frac{[TF:S]}{[TF][S]} = \frac{k_{bind}(S)}{k_{diss}(S)} = \frac{1}{K_D} = c \cdot e^{-\frac{\Delta G(S)}{kT}}$, where c is a constant and $\Delta G(S)$ is the (usually negative) binding free energy of the TF to its recognition sequence S on the DNA.

7.5 Binding free energy models

Let us consider the binding reaction of two molecules *L* and *M*:

$$L + M \underset{\leftarrow}{\overset{\rightarrow}{}} LM.$$

The dissociation equilibrium constant K_D is defined as:

$$K_D = \frac{[L][M]}{[LM]} = \frac{k_{diss}}{k_{bind}}$$

, where [*L*], [*M*], and [*LM*] are the molecular concentrations of *L* and *M* and of the complex *LM*.

In equilibrium, we may take *T* as the total concentration of molecule *L*

$$T = [L] + [LM].$$

y is the fraction of molecules *L* that have reacted (bound),

$$y = \frac{[LM]}{[LM] + [L]}$$

.

7.5 Binding free energy models

$$y = \frac{[LM]}{[LM] + [L]}$$

Substituting [*LM*] by [*L*] [*M*] / K_D gives

$$y = \frac{([L][M])/K_D}{([L][M])/K_D + [L]} = \frac{([M])/K_D}{([M])/K_D + 1}$$

.

When a solution contains both the DNA sequence and the TF with total concentration n_{tf}, the equilibrium probability that the DNA is bound to a TF molecule is (replace in upper eq. [M] by n_{tf}):

$$p(TF\ is\ bound\ to\ S) = \frac{\frac{1}{K_D} \cdot n_{tf}}{\frac{1}{K_D} \cdot n_{tf} + 1} = \frac{c \cdot e^{-\Delta G(S)/kT} \cdot n_{tf}}{c \cdot e^{-\Delta G(S)/kT} \cdot n_{tf} + 1}$$

We multiply this with $e^{+\Delta G(S)/kT}$ and divide by c·n_{tf}.

7.5 Binding free energy models

This gives: $P(TF\ is\ bound\ to\ S) = \frac{1}{1+\frac{e^{\Delta G(S_i)/kT}}{c \cdot n_{tf}}}$,

where $\Delta G(S_i)$: free energy of the TF binding to S_i .

We set $c \cdot n_{tf} = e^{\frac{\mu}{kT}}$ or $\mu = kT \cdot ln(c \cdot n_{tf})$

μ : chemical potential set by the TF concentration. This gives

$$P(TF\ is\ bound\ to\ S) = \frac{1}{1 + e^{(\Delta G(S_i) - \mu)/kT}}$$

,

This is the so-called Fermi-Dirac form of binding probability.

A sequence having a binding free energy well below the chemical potential ($\Delta G(S_i) - \mu \ll 0$) is almost always bound to the TF.
($P(TF\ is\ bound\ to\ S) \to 1$ because the exponential term is very small.)

In cases when the binding free energy is well above the chemical potential, the sequence is rarely bound.

7.5 Binding free energy models

The binding energy model (BEM) uses a vector of (free) energy contributions, $\vec{E}$.

For any sequence S_i, the binding energy predicted by the BEM model is

$$E(S_i) = \vec{E} \cdot \vec{S}_i$$

where $\vec{S}_i$ is the vector encoding of sequence S_i that can include whatever features of the sequence are relevant to its binding energy.

If the only relevant features are which bases occur at each position within the binding site, then $\vec{E}$ will be a PSSM with the characteristic that each element is a (free) energy contribution.

7.5 Binding free energy models

When the (free) energy contributions of each position are independent, $\vec{E} \cdot \vec{S}_i$ can be written as:

$$E(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \epsilon(b,m) S_i(b,m)$$

where L : length of the binding site, $\varepsilon(b, m)$: (free) energy contributions of base b at position m, and $S_i(b, m)$: indicator variable with $S_i(b, m) = 1$ if base b occurs at position m of sequence S_i and $S_i(b, m) = 0$ otherwise.

If the positions are not independent, one can include pairwise interactions between adjacent positions m and n by adding **interaction terms** to the energy function such that $\vec{E} \cdot \vec{S}_i$ is

$$E(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \epsilon(b,m) S_i(b,m) + \sum_{m=1}^{L-1} \sum_{n=m+1}^{L} \sum_{b=A}^{T} \sum_{c=A}^{T} \epsilon(b,m,c,n) S_i(b,m,c,n)$$

where $\varepsilon(b, m, c, n)$: energy contribution of having base b at position m and base c at position n.