CHAPTER 3

Predicting TF Binding

The main focus of this thesis lies on improving the detection of TF – target promoter interactions through an interpretable biophysical model and subsequently to find TFs that cause the regulation of groups of genes. The following chapter is therefore divided into two parts. The first part focuses on how TF binding can be inferred from experimentally verified TF binding sites. To this end, after describing the most frequent probabilistic models used to infer individual TF binding sites the biophysical model by Berg and von Hippel (1987) for the prediction of TF binding energies will be introduced. How the statistical and physical approaches are related will be briefly outlined. The second part of the chapter is centred on how to detect transcription factors that play a role in the regulation of groups of genes. In this context a number of frequently cited methods will be discussed.

3.1 Principles of TF binding site discovery

As described in the previous chapter various approaches exist for determining the binding preference of TFs. Most of the TF binding data available today stems from small scale experiments such as EMSA tests or from a rather limited number of high affinity sites derived from SELEX experiments. In this section I will discuss how such experimentally verified sites can be used to identify possible binding sites in the genome based on statistical and biophysical considerations.

3.1.1 Deriving a Consensus sequence

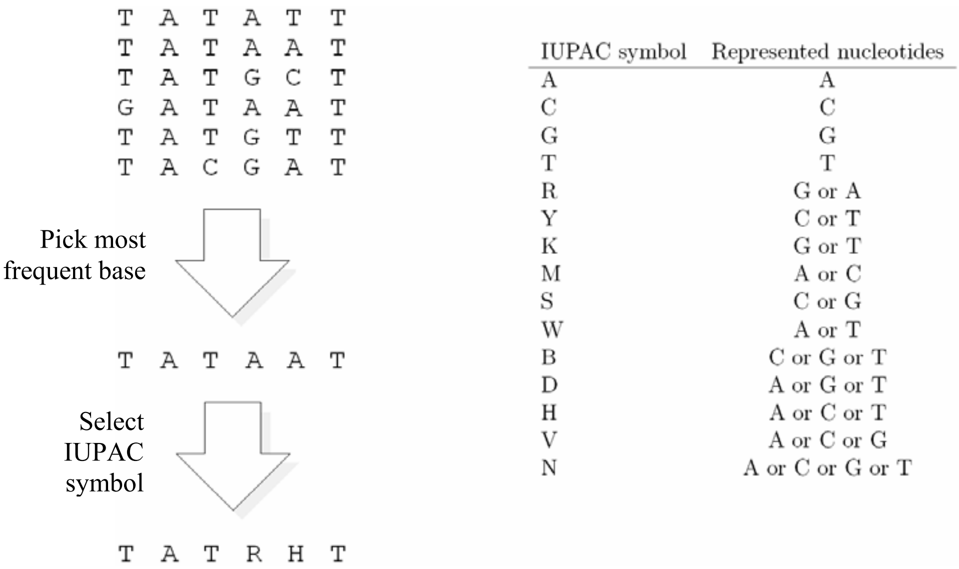
Basic site search

The simplest approach for computationally detecting new potential TF binding sites is to annotate all sites in a genome that share the sequence with one of the previously experimentally identified binding sites of the TF. Naturally, the smaller the number of experimentally verified binding sites the smaller will be the fraction of binding sites that can be detected in the genome. The success of this method is especially limited if the binding site of a TF contains bases not actually contacted by the TF. As was noted in Figure 2.5 such bases are often not constricted and therefore change randomly between different binding sites. Similarly, due to experimental limitations uninformative flanking sequences, not actually contacted by the TF, may be considered as belonging to a given binding site. In turn if such sites stem from a SELEX experiment they may not perfectly match to the genome at all.

Consensus sequences

Early on, the observation was made that different binding sites of a given TF resemble each other. Individual binding sites thereby often diverge in only a few bases from a common consensus motif. Therefore, a first step towards making a meaningful generalization of the observed binding patterns has been to align all experimentally found binding sites of a given TF and to construct a consensus sequence from the alignment. This procedure is illustrated in Figure 3.1 for six instances of a Pribnow box motif (Pribnow, 1975; Schaller et al., 1975). The original description of a consensus was to take the most common base at each position in the alignment. The resulting consensus sequence provides a simple representation of the preferred binding motif, which can readily be used to scan for occurrences of this motif in a sequence. At a time when computers where not widely available this simplicity was an important feature. However, the rigidity of such a consensus description causes a high number of false negative predictions. For instance, for the example shown in Figure 3.1, if one would desire to detect a perfect match to the consensus one would identify only one of the six sites used to construct the consensus. At the same time one would detect a false positive match statistically only every 4^6 = 4096 bases. In order to recover more of the true positive sites, matching to the consensus can be made more flexible by allowing for one mismatch. In this case three true positives would be recovered while a false positive prediction would be made every 228 bases. In order to detect all six positive examples one needs to allow for two mismatches. However, this goes in hand with obtaining a false positive hit every 30 base pairs. The situation can be slightly improved if one uses a more descriptive alphabet such as provided by the IUPAC code. Using this alphabet the consensus sequence TATAAT can be rewritten as TATRHT. The use of ambiguity codes allows retaining more of the information provided by the verified sites. That is, a highly degenerate position in the alignment can be treated differently from a position conserved in all sites. Using TATRHT for the search one recovers four out of six sites while obtaining a random hit in the genome only about every 512 bases. Nevertheless, to recover all six known sites one would again have to allow for one mismatch leading back to a false positive rate of one in 30 base pairs. This example shows that while consensus motifs are useful visualisations of binding patterns and easy to construct their predictive ability is strongly dependent on the used alphabet and the rules applied for pattern matching when scanning the genome (Stormo, 2000; Day and Mcmorris, 1992).

Figure 3.1 – From a binding site alignment to a consensus sequence



IUPAC symbol	Represented nucleotides
A	A
C	C
G	G
T	T
R	G or A
Y	C or T
K	G or T
M	A or C
S	C or G
W	A or T
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	A or C or G or T

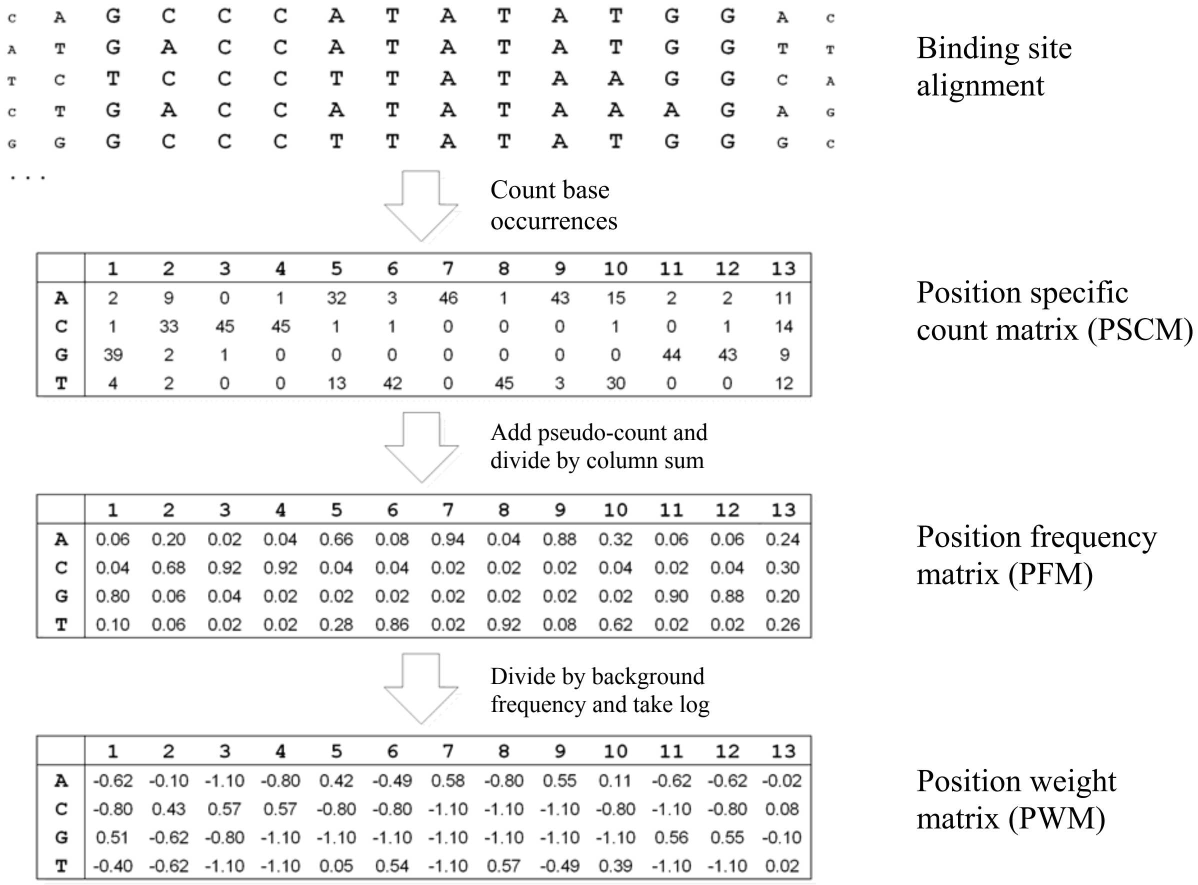
A simple consensus is derived by aligning experimentally found binding sites and selecting the most frequent base at each position of the alignment. An ambiguity alphabet such as the IUPAC code shown on the right, can be applied to retain more information about the motif flexibility indicated in the alignment.

3.1.2 Deriving a TF binding motif

The main problem of the consensus model described above is that it cannot fully account for the relative importance of each position in the binding sites. To alleviate this problem the concept of a position frequency matrix (PFM) was introduced by Harr et al. (1983) and Staden (1984). PFMs allow putting high weights on important positions in the binding site and low weights on positions that show little or no preference for a given base. PFMs as used today are two-dimensional arrays where the number of columns corresponds to the length of the binding site and the rows correspond to the bases A, C, G and T. As shown in Figure 3.2 the entries in the matrix are derived from the alignment of known binding sites. The number of occurrences of each base at each position in the alignment is thereby first entered into the corresponding position of a so called position specific count matrix (PSCM). Today, there exists a large number of PSCMs available for many TFs, which are stored in various databases such as TRANSFAC (Matys et al., 2003) and JASPAR (Sandelin et al., 2004). The counts in a PSCM can be converted into base frequencies by dividing each cell by the sum over the four entries in the corresponding column (after adding a pseudo count to each element) thereby giving rise to a position frequency matrix (PFM).

Figure 3.2 – From binding site alignments to PWMs

a)



```
c  A  G  C  C  C  A  T  A  T  A  T  G  G  A  c
A  T  G  A  C  C  A  T  A  T  A  T  G  G  T  T
T  C  T  C  C  C  T  T  A  T  A  A  G  G  C  A
c  T  G  A  C  C  A  T  A  T  A  A  A  G  A  G
G  G  G  C  C  C  T  T  A  T  A  T  G  G  G  c
. . .
```

Binding site alignment

Count base occurrences

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	2	9	0	1	32	3	46	1	43	15	2	2	11
C	1	33	45	45	1	1	0	0	0	1	0	1	14
G	39	2	1	0	0	0	0	0	0	0	44	43	9
T	4	2	0	0	13	42	0	45	3	30	0	0	12

Position specific count matrix (PSCM)

Add pseudo-count and divide by column sum

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	0.06	0.20	0.02	0.04	0.66	0.08	0.94	0.04	0.88	0.32	0.06	0.06	0.24
C	0.04	0.68	0.92	0.92	0.04	0.04	0.02	0.02	0.02	0.04	0.02	0.04	0.30
G	0.80	0.06	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.90	0.88	0.20
T	0.10	0.06	0.02	0.02	0.28	0.86	0.02	0.92	0.08	0.62	0.02	0.02	0.26

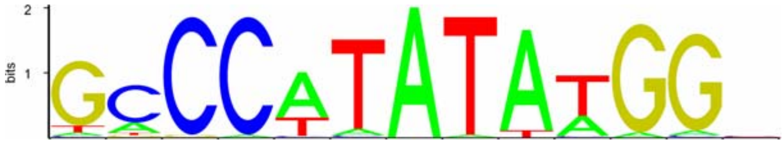
Position frequency matrix (PFM)

Divide by background frequency and take log

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	-0.62	-0.10	-1.10	-0.80	0.42	-0.49	0.58	-0.80	0.55	0.11	-0.62	-0.62	-0.02
C	-0.80	0.43	0.57	0.57	-0.80	-0.80	-1.10	-1.10	-1.10	-0.80	-1.10	-0.80	0.08
G	0.51	-0.62	-0.80	-1.10	-1.10	-1.10	-1.10	-1.10	-1.10	-1.10	0.56	0.55	-0.10
T	-0.40	-0.62	-1.10	-1.10	0.05	0.54	-1.10	0.57	-0.49	0.39	-1.10	-1.10	0.02

Position weight matrix (PWM)

b)



Sequence LOGO

a) Experimentally verified binding sites for the transcription factor SRF are aligned and the frequency of each base at each position is determined. The resulting count matrix can then be used to generate a frequency matrix and finally a weight matrix. Databases such as TRANSFAC and JASPAR have stored PSCMs for hundreds of TFs. b) The base preference of the factor at each position can be illustrated by a sequence LOGO where the height of a letter indicates the importance of the base.

The mathematical form of a position frequency matrix is given by:

$$PFM = \begin{bmatrix} v_{1,A} & \cdots & v_{M,A} \\ v_{1,C} & \cdots & v_{M,C} \\ v_{1,G} & \cdots & v_{M,G} \\ v_{1,T} & \cdots & v_{M,T} \end{bmatrix}$$

where *M* denotes the length of the sequence and $v_{m,\alpha}$ denotes the frequency with which base α is observed at position *m* in the known binding sites. Given a set of known binding sites the above definition of the PFM corresponds to the maximum likelihood estimate for the *M* × 4 parameters that are used by the PFM to model the sites. Instead of making a simple yes or no decision when classifying sites in the genome as was the case with consensus matching a PFM can be used to compute a continuous score that reflects the similarity between a given site and the binding site model. Since positions in the PFM are assumed to be independent of each other, a natural choice for the similarity score is the probability of a site *S* being generated by the model. This probability $p(S)$ can be computed by:

$$p(S) = \prod_{m=1}^{M} v_{m,\alpha} . \tag{3.1}$$

The consensus sequence is hereby always generated with the highest probability. For example, the consensus sequence GCCCATATATGGC of the transcription factor SRF is generated according to the PFM shown in Figure 3.2 with probability:

$$p = 0.8 \cdot 0.68 \cdot 0.92^2 \cdot 0.66 \cdot 0.86 \cdot 0.94 \cdot 0.92 \cdot 0.88 \cdot 0.62 \cdot 0.90 \cdot 0.88 \cdot 0.30 = 0.03 .$$

In contrast, the site AACCAAAAAAGGA, which contains several substitutions in respect to the consensus, is generated with probability:

$$p = 0.06 \cdot 0.2 \cdot 0.92^2 \cdot 0.66 \cdot 0.08 \cdot 0.94 \cdot 0.04 \cdot 0.88 \cdot 0.32 \cdot 0.90 \cdot 0.88 \cdot 0.24 = 10^{-6}$$

and is thus by a factor of ~3×10^{-5} less likely to be a site sampled from the matrix model than the consensus. As illustrated in Figure 2.5 and as quantified in Figure 3.2 for SRF not all bases in a binding site interact specifically with a given TF, consequently, not all positions in the matrix are equally important for the generated scores. For instance, in position 13 of the SRF matrix all bases occur with similar frequency suggesting that this position does not play an important role for the specific interaction of the site with the TF. Accordingly, a deviation from the consensus at this position does not down-weight a site strongly. In contrast, at position three of the SRF matrix any base except the consensus base causes a severe reduction in the probability of a site being drawn from the matrix model. While strong differences in the weights are generally desired, an unwanted situation occurs for entries with $\nu_{m,\alpha} = 0$. Such positions introduce a total probability of 0 for a given sequence to be a binding site irrespective of how well all other bases match the consensus. Since TF have a general tendency to associate with DNA via unspecific interactions assuming a binding probability of 0 for any site is unrealistic. The existence of $\nu_{m,\alpha} = 0$ is in general only due to the small sampling effect introduced by building the matrix from a few strongly bound sites and is particularly harmful at degenerate positions where in reality there exists only a moderate preference for a particular base.

Pseudo counts

Various ways have been suggested for adding pseudo counts (PC) to the elements in a count matrix in order to avoid the occurrence of $\nu_{m,\alpha} = 0$. The most simple and common approach is to add a value of PC = 1 to all elements in the matrix (Bucher 1990, Berg & von Hippel 1987). Aside from avoiding the extreme case of $\nu_{m,\alpha} = 0$, making the PFM more general by adding a PC to all elements is likely of advantage if only a small number of observed binding sites is known. Naturally, the larger the number of observed binding sites the smaller will be the effect of adding the PC.

Besides such simple PC descriptions, other more involved methods have been proposed. For instance, Rahmann et al., (2003) proposed to add a PC that depends on the information content of a given column in the matrix. Such a scheme has the advantage that it can keep the informative core of a matrix virtually untouched while fluctuations in irrelevant positions can be evened out. For example, the expected values in a count matrix column C_m generated by picking nine sites from a random background model with base distributions of 0.25 would be C_m = {2, 2, 2, 3}. The corresponding column in the frequency matrix would thus be $F_m =$ {0.22, 0.22, 0.22, 0.33}. However, since none of the bases should in fact carry more weight than the others this method suggests to change these frequencies to {0.25, 0.25, 0.25, 0.25}.

3.1.3 Statistical methods for finding binding sites in the genome

Once a position frequency matrix has been derived it can be used to identify likely TF binding sites in the genome. As described above the probability of a site being generated by the PFM can be computed by taking the product over all corresponding frequencies in the matrix. However, using these probabilities to identify likely binding sites is really valid only if all four bases occur with equal frequencies in the genome.

Requirement for a background model (BM)

If the bases in the genome do not occur with equal probability that is, $b = \{b_{\mathrm{A}}, b_{\mathrm{C}}, b_{\mathrm{G}}, b_{\mathrm{T}}\} \neq \{0.25,\ 0.25,\ 0.25,\ 0.25\}$, using raw probability is inadequate. To illustrate the problem consider the following PFM with consensus sequence TAAAT stemming from 21 hypothetical binding sites (after adding a PC of 1 to all elements):

$$PFM = \begin{bmatrix} 0.44 & 0.48 & 0.36 & 0.48 & 0.44 \\ 0.04 & 0.04 & 0.04 & 0.04 & 0.04 \\ 0.04 & 0.04 & 0.32 & 0.04 & 0.04 \\ 0.48 & 0.44 & 0.28 & 0.44 & 0.48 \end{bmatrix}.$$

Two sites found in a given promoter, S_1 = TAAAT and S_2 = TAGAT would then be generated by the model with probabilities:

P(TAAAT | PFM) = 0.48 * 0.48 * 0.36 * 0.48 * 0.48 = 0.019

P(TAGAT | PFM) = 0.48 * 0.48 * 0.32 * 0.48 * 0.48 = 0.017.

The consensus sequence, S_1, would consequently be recognized as binding site with greatest significance. However, now consider these sites are located in an extremely AT rich genome with GC content 0.02, $b = \{0.49,\ 0.01,\ 0.01,\ 0.49\}$. Sequences such as TAAAT would therefore be found randomly in the genome every ≈ 32 bps while a sequence like TAGAT would be found only every ≈ 1600 bps. That 1/3 of the known binding sites contain a G would thus indicate a strong bias of the TF to recognize a G in the central position. Given the vast excess of AT containing sequences in the genome the other 2/3 of the experimentally found sites might have been bound by the factor only in an unspecific fashion. Sites in the genome resembling TAGAT are hence more indicative of a true binding site than a sequence consisting only of A's and T's. To take care of the bias introduced by skewed background distributions one can first compute the probability that the sites were generated by the background. For S_1 and S_2 the probability of stemming from a background model (which assumes the above base distribution *b* and independence of bases) would be:

P(TAAAT | BM) = 0.49 * 0.49 * 0.49 * 0.49 * 0.49 = 0.028

P(TAGAT | BM) = 0.49 * 0.49 * 0.01 * 0.49 * 0.49 = 0.00058.

Given these probabilities one can now ask the question of whether a given site is more likely to be generated by the PFM or by the background model. To this end one computes the log likelihood ratio given by:

$$\Lambda = \log\left(\frac{P(S \mid PFM)}{P(S \mid BM)}\right) = \log\left(\frac{L(PFM \mid S)}{L(BM \mid S)}\right). \qquad (3.2)$$

where $L(PFM \mid S)$ and $L(BM \mid S)$ are the likelihoods of the PFM model parameters and background model parameters, respectively, given the observed site S. The larger Λ the more likely it is that the site represents a true binding site of the TF. For S_1 and S_2 the log likelihood ratios would be:

$$\Lambda_{S_1} = \log\left(\frac{P(TAAAT \mid PFM)}{P(TAAAT \mid BM)}\right) = \log\left(\frac{0.019}{0.028}\right) = -0.2$$

$$\Lambda_{S_2} = \log\left(\frac{P(TAGAT \mid PFM)}{P(TAGAT \mid BM)}\right) = \log\left(\frac{0.017}{0.00058}\right) = +1.5\,.$$

The consensus site is generated with slightly higher probability by the BM (Λ = -0.2) and would thus not be indicative of a true binding site. In contrast, S_2 is ≈ 30 times more likely (Λ = 1.5) to represent a true binding site than a site from the BM. While this was an extreme example it demonstrates the principal requirement for the use of a background models if bases do not occur with uniform frequency in the genome. Using genome frequencies and independence between the bases constitutes the simplest BM. Other more sophisticated models have been proposed including higher order Markov chains (Kim et al., 2006). Such models take into account that promoter sequences tend to contain CpG islands and AT rich stretches (consecutive bases are thus not independent). On the other hand such models cannot be applied blindly as they might filter out the actual binding signal for certain factors such as for the transcription factor SP1 which binds GC rich sequences or TBP which binds TATAA.

When using the simple BM, fast scanning of sequences for binding sites can be facilitated by pre-computing the log likelihood ratios. To this end every element of the PFM is divided by the background frequency b_α of the corresponding base α. Taking the log of each resulting element yields a so called position specific scoring matrix (PSSM) also referred to as position weight matrix (PWM) with weights $w_{m,\alpha}$ replacing the $v_{m,\alpha}$ of the PFM (see Figure 3.2). The total score (Λ) of a site S is then given by:

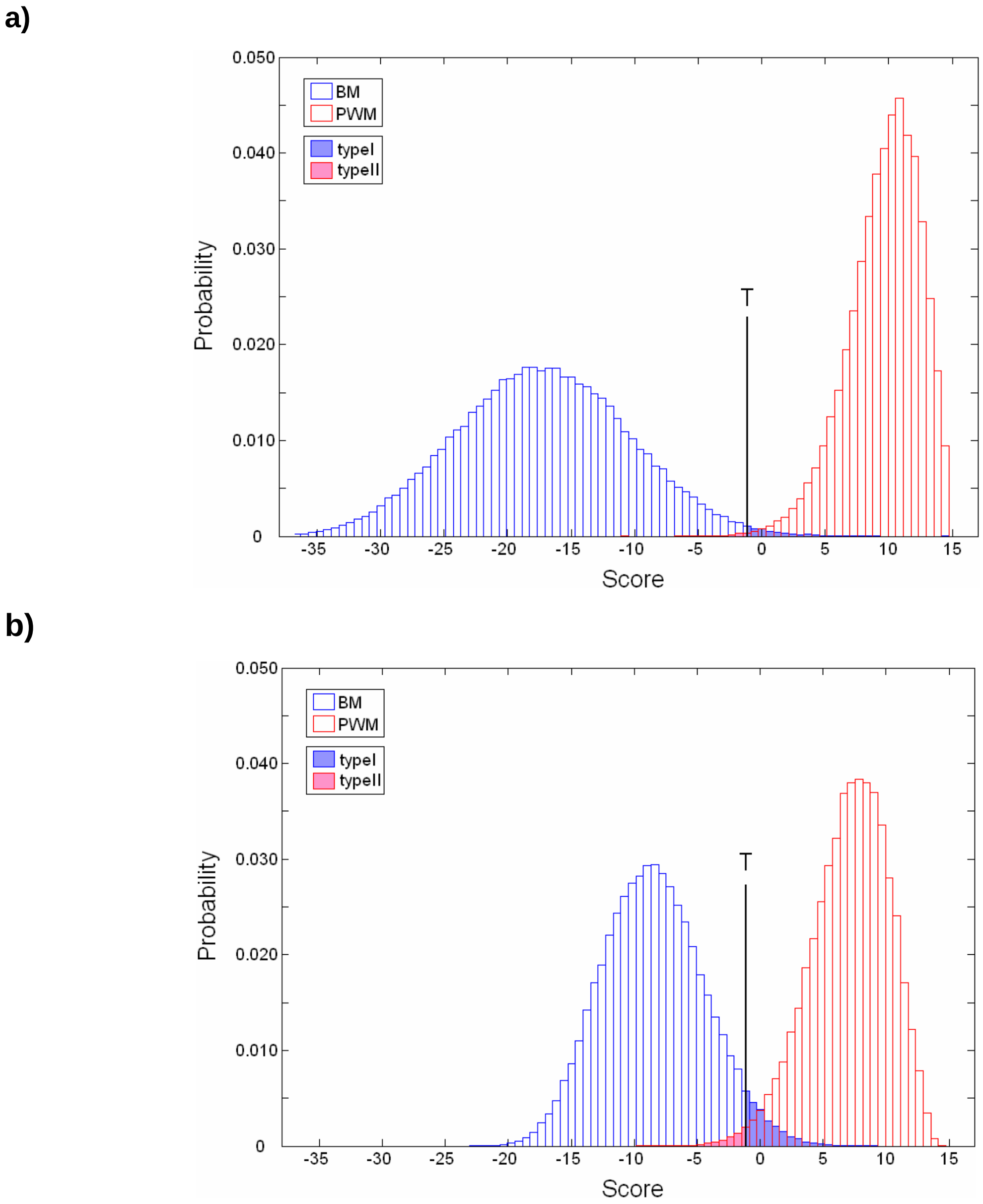
$$\Lambda(S) = \log\left(\frac{P(S \mid PFM)}{P(S \mid BM)}\right) = \sum_{m=1}^{M} \log\left(\frac{v_{m,\alpha}}{b_\alpha}\right) = \sum_{m=1}^{M} w_{m,\alpha} \ . \qquad (3.3)$$

Which score is large enough to indicate a binding site?

Deciding whether a computed score (the log likelihood ratio between PFM and background model) is large enough to suggest a true binding site is traditionally solved by introducing a score threshold *T* that needs to be surpassed. Sites with score > *T* are subsequently considered to be binding sites or hits for the TF while all sites with score < *T* are considered non-binding sites. The choice for the threshold thereby depends on the quality of the PFM and the ultimate goal of the analysis. If a stringent cutoff is chosen then the number of false positive binding site predictions is reduced (minimizing the type I error) while true binding sites might be missed. On the other hand, using a lenient cutoff allows detecting more true positives (minimizing the type II error) but might quickly lead to an overwhelming number of false positive predictions.

Conveniently, the magnitude of the type I and type II errors, given a certain threshold, can be estimated from the distribution of the scores produced by selecting the elements of the PWM either according to the background or the PFM model (Rahmann et al., 2003). For illustration let us assume the simplest background model with independence between the bases and base frequencies $b = \{0.25, 0.25, 0.25, 0.25\}$. In complete analogy to a coin flipping experiment, any site of length *M* is generated by this model with probability 0.25^M where each occurrence of a site is considered to have arisen by chance. We can now ask how many of these randomly generated sequences will obtain a score higher than the threshold. For visualisation we can plot the probability mass function of the score distribution according to the background model. Figure 3.3 shows an example of this distribution for a PWM from the transcription factor HNF1. The probability that a score larger than the threshold will appear by chance in the genome (type I error) is then given by the area under the curve to the right of the chosen threshold. For the example in Figure 3.3 the background model generates with highest frequency sequences with score around -17 but rarely a sequence with score larger than the indicated threshold. Therefore, the expected type I error

Figure 3.3 – Empirical score distribution based on BM and PFM



a) Empirical score distribution of the scores from a high quality PWM for HNF1 given the background model (blue) or matrix model (red), respectively. The x axis shows the size of the score and the y axis the probability of obtaining the score. The probability of finding a score larger than a certain threshold T given the BM (→ type I error) is shown by the blue area under the curve. Similarly, the probability of a score being smaller than T given the PFM (→ type II error) is indicated by the area in red. b) Empirical score distribution given a low quality matrix for SRF. In this case the two distributions show a larger overlap which complicates the classification of sites within that range. The score distributions can efficiently be derived using a Fourier transform of the PFM or BM.

will be small. Similarly to this procedure, we can generate sequences also according to the PFM model and ask how many of these sequences have a score below the threshold. Such a case means that a sequence has in fact been generated by the PFM model but is not recognized as a TF binding site. Thus we can estimate the number of false negative predictions (type II error) by measuring the area under the probability mass function of the PFM model to the left of the threshold. The higher the quality (information content) of a PFM the more clearly will be the separation of the PFM and BM score distributions.

The score distributions for any PWM according to BM and PFM model can be efficiently be derived via their moment generating or characteristic functions (Staden, 1989; Rahmann et al., 2003). Given the probability mass functions several score cutoffs have been suggested for the annotation of binding sites. One common prescription is to limit the type I error to 5%. A second more sophisticated method aims at balancing the number of false positive and false negative predictions (Rahmann et al., 2003). This approach works well for high information matrices but can cause a large number of false positive predictions if the score distributions between BM and PFM are not clearly separated.

The statistical interpretation of TF binding as outlined above has gained wide popularity with most sequence analysis tools available today relying on similar statistical measures for the predicting of discrete TF binding sites. The following section switches from the statistical to a biophysical interpretation of TF binding and outlines how a set of known binding sites can be used to predict the binding free energy between a TF and a given site.

3.1.4 A biophysical model to predict TF binding energies

While the above considerations are very useful for identifying likely binding sites of a factor they have a number of limitations. Most notably, the strength with which a factor binds to a given site (the binding energy between TF and site) and thus the probability of the TF actually sitting on the sequence is not predicted. In addition, once a cutoff has been chosen all sites with scores above the threshold are considered to be “bound” while all sites below the cutoff are considered “unbound”. This discretization causes a loss of information about the actual divergence from the consensus. Lastly, the number of predicted sites greatly varies with the choice of the score threshold.

In the following I outline the groundbreaking work by Otto Berg and Peter von Hippel, which allows to derive the binding free energy of a TF to any given site in a DNA sequence based on a set of verified binding sites. These binding energies can subsequently be used to find the probability of given TF binding a given DNA site. Much of the logic applied by Berg and von Hippel (1987) stems from a classical derivation of the Boltzmann distribution in statistical mechanics (Atkins, 1978). The reader is referred to Appendix A for an outline of the Boltzmann distribution and a description of the similarities to the present derivation of TF binding energies.

Statistical mechanical theory for TF-DNA interactions

As illustrated in Figure 3.4 the binding of a TF to a piece of DNA with the length of its binding site goes in hand with a favourable change in the energy of the system. In the following we want to derive a measure to estimate the magnitude of the energy change associated with the TF binding to any site *i*. To start with, as in case of the purely statistical analysis of TF binding sites, individual positions in a site are assumed to be independent from each other, i.e., seeing a particular base at position *m* in the site does not affect the probability of observing a particular base elsewhere. Following this prescription, each base pair α at position *m* in a site of length *M* contributes a specific energy $\varepsilon_{m,\alpha}$ to the total binding energy *E*, between the TF and the site:

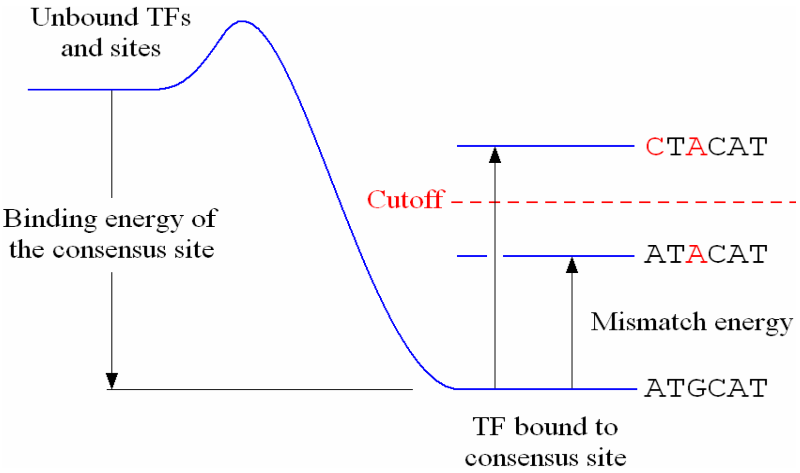
$$E = \sum_{m=1}^{M} \varepsilon_{m,\alpha} \ . \qquad (3.4)$$

The site with strongest possible binding (the optimal binder) is thereby set to have total binding energy $E_0 = 0$, that is, all $\varepsilon_{m,\alpha} = \varepsilon_{m,0} = 0$. For all other sites, changing a base pair in respect to the optimal binder will introduce a specific mismatch energy, $\varepsilon_{m,\alpha} > 0$ (see Figure 3.4). We assume that all experimentally found binding sites for a given TF have a maximal total mismatch energy of less than some critical value, E_C,

$$\sum_{m=1}^{M} \varepsilon_{m,\alpha} \leq E_C \ . \qquad (3.5)$$

Sites which bind weaker to the TF ($E > E_C$) are assumed to not be specifically associated with the factor and are thus not to be found in the set of experimentally known regulatory sites. On the other hand, all sites with $E < E_C$ are assumed to be equally well suited as regulatory sites and are therefore, *a priori*, equally likely to be found in the set.

Figure 3.4 – Energy changes associated with TF – DNA binding


Unbound TFs and sites
Binding energy of the consensus site
Cutoff
CTACAT
ATACAT
Mismatch energy
ATGCAT
TF bound to consensus site

The binding of a TF to DNA goes in hand with a favourable change in the energy of the system. The largest energy change occurs hereby if the TF binds to its consensus site (e.g. ATGCAT). Changing bases in respect to the consensus site (highlighted in red) introduces a mismatch energy that lowers the binding energy of the TF to the corresponding site. The cutoff introduced by the PWM approach can be viewed as a maximal mismatch energy above which no binding can occur. From the biophysical point of view introducing such a cutoff is unnecessary however.

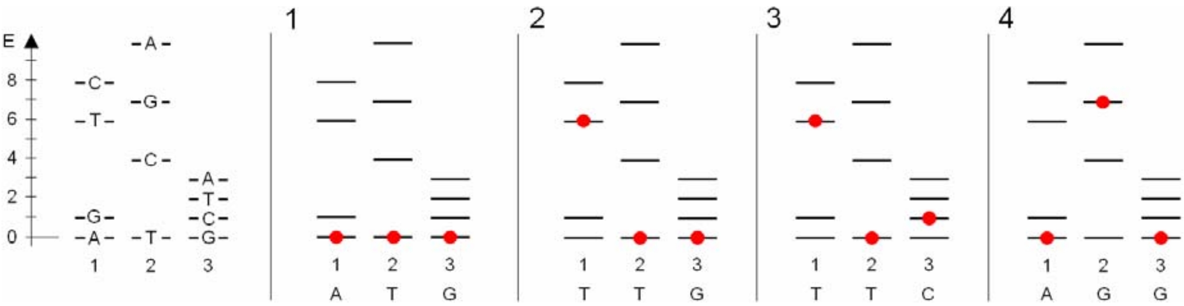
For any given TF with binding sites of length *M* there exist *W(M, E_C)* possible sites which have $E < E_C$. However, choosing $\varepsilon_{j,\alpha} > 0$ at any given position *j* in a site puts a restriction on all other positions,

$$\sum_{m \neq j}^{M-1} \varepsilon_{m,\alpha} \leq E_C - \varepsilon_{j,\alpha}. \qquad (3.6)$$

In general, the larger $\varepsilon_{j,\alpha}$ in a given position the more sequence combinations are disallowed. The situation is illustrated in Figure 3.5a for a hypothetical TF with a binding site of length 3 and E_C = 7.5. For this factor, when putting α = A in position *j* = 1 then there are nine sequence combinations from position 2 and 3 which yield $E < E_C$ (A**TG**, A**TC**, A**TT**, A**TA**, A**CG**, A**CC**, A**CT**, A**CA**, A**GG**). In comparison, putting G in the first position would yield seven allowed combinations while using a T would allow only two possible combinations with $E < E_C$ (T**TG** and T**TC**). Similarly, choosing a G in the second position requires that all other bases match the sequence of the optimal binder to avoid $E > E_C$. It follows that there is only one possible realisation of a binding site with G in the second position. Generally speaking, the number of possible sequences which are allowed, given that we have chosen base α at position *j*, can be written as *W(M - 1, E_C - $\varepsilon_{j,\alpha}$)*. If we assume that all sites of length *M* - 1 occur

Figure 3.5 – Hypothetical energy configurations for a TF with binding site length 3

a)



b)

	1	2	3
A	**9**	0	3
C	0	7	5
G	7	1	**6**
T	2	**10**	4

a) Illustration of the energy configurations of a number of binding sites for a hypothetical TF. The leftmost panel shows the energy levels $\varepsilon_{j,\alpha}$ (horizontal bars) that are associated with putting the indicated base α at position *j* in the shown motif of length 3. Reminiscent of the distribution of particles among available energy levels in an ideal gas, red balls signify which of the energy levels are occupied in a given site. Panel 1 hereby shows the energy configuration for the most strongly bound site, ATG. In this case always the lowest energy level $\varepsilon_{1,A} = \varepsilon_{2,T} = \varepsilon_{3,G} = 0$ is occupied. For a maximal total mismatch energy of $E_C = 7.5$ panels 2 through 4 show the allowed energy configurations resulting from putting a T in the first or a G in the second position. b) Most likely count matrix arising from the energy levels in a).

with the same *a priori* probability in the genome then the fraction of known binding sites with base α at position *j* will be given by:

$$\nu_{j,\alpha} = \frac{W\left(M-1, E_C - \varepsilon_{j,\alpha}\right)}{W\left(M, E_C\right)}. \qquad (3.7)$$

In turn, the number of times one will find a given base α at position *j* with respect to the corresponding base found in the optimal binder (with $\varepsilon_{j,0} = 0$) is:

$$\frac{\nu_{j,\alpha}}{\nu_{j,0}} = \frac{W\left(M-1, E_C - \varepsilon_{j,\alpha}\right)}{W\left(M-1, E_C - \varepsilon_{j,0}\right)} = \frac{W\left(M-1, E_C - \varepsilon_{j,\alpha}\right)}{W\left(M-1, E_C\right)} \qquad (3.8)$$

The observed frequencies of the bases in the set of experimentally identified binding sites will be proportional to these fractions. Figure 3.5b shows the position specific count matrix for the hypothetical TF that would result in the ideal case (ignoring possible sampling effects) from the binding energetics described above. It should be noted that similar counts across a given column in the PSCM indicate that the energy difference between the observed and the optimal base pair are small. In contrast, if only the optimal base pair is found then all other bases must have $\varepsilon_{j,\alpha}$ >> $\varepsilon_{j,0}$.

Given the base utilization frequencies in the set of known binding sites, Berg and von Hippel obtained the underlying $\varepsilon_{j,\alpha}$'s by following the logic of a common derivation of the Boltzmann distribution (see Appendix A). They started by taking the natural logarithm of the ratio in equation (3.8) and expanding the numerator in powers of ε, using the following Taylor expansion:

$$\ln W\left(E_c - \varepsilon_{j,\alpha}\right) = \ln W\left(E_c\right) - \varepsilon_{j,\alpha}\left(\frac{d\ln W}{dE}\right)_{E_c} + \mathrm{O}\left(\varepsilon^2\right)\ldots \tag{3.9}$$

where higher order terms of ε can be neglected. In the resulting equation the first term cancels out, which yields to a first order approximation the desired relation between observed base frequencies (determined from the known binding sites) and the underlying discrimination energies:

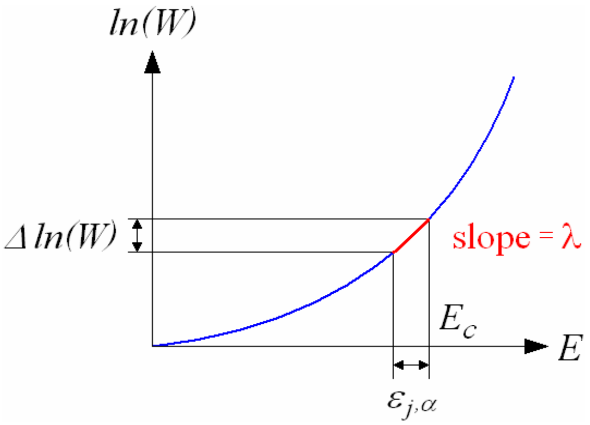
$$\ln\left(\frac{n_{j,\alpha}}{n_{j,0}}\right) \approx \ln\left(\frac{\nu_{j,\alpha}}{\nu_{j,0}}\right) = -\lambda_j \varepsilon_{j,\alpha} \tag{3.10}$$

where $n_{j,\alpha}$ and $n_{j,0}$ correspond to the counts for the observed and the most frequent base in the corresponding PSCM, respectively, and λ has been defined to be:

$$\lambda_j = \frac{d\ \ln\left(W_j\left(M\ \text{-}\ 1,\ E_C\right)\right)}{d\ E} \tag{3.11}$$

(for a graphical illustration see Figure 3.6). As the authors further showed, this approximation is valid as long as $\varepsilon_{j,\alpha} < E_C$ which holds in most cases. In addition, while λ cannot be derived directly from the PSCM it was shown to be largely independent of j and to vary only slowly with changing E_C. While equation (3.10) thus holds best for small $\varepsilon_{j,\alpha}$ it can be used to estimate the binding energy contributions from all position in a binding site. As will become evident in the next chapter λ can be viewed as scaling parameter that adjusts how close the mismatch energy levels are packed together. The above derivation is based on the assumption that all four bases occur with equal frequency in the genome. In general this is

Figure 3.6 – A graphical interpretation for the meaning of λ



The number of ways, *W*, in which a binding site can be realized increases quickly with the maximal critical energy E_c. The slope of the curve at $E = E_c$ corresponds to the value of λ as derived in the main text. With other words, λ measures how quickly the number of possible binding sites decreases when reducing the maximal mismatch energy.

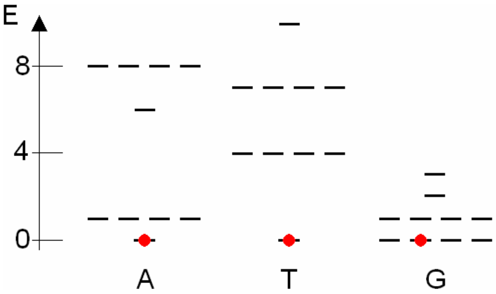
not the case and a correction for the computation of binding energies is advisable. For illustration consider the example shown in Figure 3.5 but this time assuming a genomic GC content of 0.80. In this case, while A is still the base with smallest mismatch energy for position 1, there are four times more sites in the genome containing a G at the first position than an A. The base frequencies observed in the experimentally verified binding sites will thus be biased towards having a G in the first position. This situation can be viewed as energy level $\varepsilon_{1,G}$ being four-fold degenerate in comparison to $\varepsilon_{1,A}$ (Figure 3.7). In general, the fraction of sequences with a given base α at position *j* will therefore be given by:

$$\nu_{j,\alpha} = \frac{b_\alpha \; W(M-1, E_C - \varepsilon_{i,\alpha})}{W(M, E_C)} \tag{3.12}$$

where b_α is the genome wide frequency of base α. Again dividing $\nu_{j,\alpha}$ by $\nu_{j,0}$ and following the above procedure one obtains for the individual mismatch contributions from each base pair the corrected estimate:

$$\varepsilon_{\alpha,j} = -\frac{1}{\lambda}\ln\left(\frac{\nu_{j,\alpha}\; b_0}{\nu_{j,0}\; b_\alpha}\right) = -\frac{1}{\lambda}\ln\left(\frac{n_{j,\alpha}\; b_0}{n_{j,0}\; b_\alpha}\right) \tag{3.13}$$

Figure 3.7 – Effect of background base frequencies on energy predictions



Having four times as many G's in the genome as A's can be viewed as $\varepsilon_{j,G}$ being four fold degenerate in comparison to $\varepsilon_{j,A}$. The probability of a binding site starting with a G thus increases by a factor of four in comparison to the case shown in Figure 3.5.

and for the total mismatch energy, *E*, of a given site of length *M*:

$$E = -\frac{1}{\lambda}\sum_{m=1}^{M}\ln\left(\frac{\nu_{m,\alpha}}{\nu_{m,0}}\frac{b_0}{b_\alpha}\right) = -\frac{1}{\lambda}\sum_{m=1}^{M}\ln\left(\frac{n_{m,\alpha}}{n_{m,0}}\frac{b_0}{b_\alpha}\right) \qquad (3.14)$$

where the genome wide frequencies b_0 and b_α have to be taken from the organism from which the corresponding PFM has been derived (for SELEX derived matrices b_α = 0.25).

Several physically motivated models exist that convert predicted mismatch energies into probabilities of a TF occupying a given site in the genome. Each of these models makes a number of simplifying assumption about the situation found in cells. The most frequent of these approaches, which is based on assuming a Boltzmann distribution for the binding probabilities, will be outlined in the next section.

3.1.5 From the biophysical approach to PWMs

This section describes how the mismatch energies derived above can be used to predict relative TF-DNA binding probabilities by assuming that binding sites are occupied by a TF molecule according to Boltzmann statistics (see Appendix A). I will thereby focus on the simplified model outlined by Heumann et al. (1994), which forms the argumentative basis for many current bioinformatics applications (among which are PAP and Clover, two methods described in Section 3.2). The main claim hereby is that the exponential of PWM scores derived in Section 3.1.2 is directly proportional to the binding probability of a TF to DNA. Combining the results of the previous section with the model of Heumann et al. (1994) will highlight under what special circumstances this claim holds and the likelihood ratios and binding probabilities converge.

For the derivation of the binding probabilities from mismatch energies imagine a single TF molecule together with a mixture of all 4^M types of sequences of length *M*, where *M* is the length of the binding motif of the TF (Gerland et al., 2002; Heumann et al., 1994; Fields et al., 1998; Stormo 2000). This system can be viewed as a greatly simplified version of a real genome. The energy scale of the system is thereby set so that the TF bound to its consensus site resides on energy level $E_0 = 0$. Without any competition between TF molecules for the sites (since there is only one TF molecule in the system) and since in contrast to a real genome the sites do not overlap, the probability of a site *i* to be bound by the factor is given by the Boltzmann distribution:

$$p_i = \frac{e^{-\beta E_t}}{\sum_{j=1}^{N} e^{-\beta E_j}} = \frac{e^{-\beta E_t}}{Z} \tag{3.15}$$

where E_t is the mismatch energy between the TF and a site of type *t* and $\beta = 1$ over Boltzmann constant times absolute temperature. The probability of a site to be bound by the TF is here given by its Boltzmann factor $e^{-\beta E_t}$ scaled by *Z*, the so called molecular partition function of the system. In this case *Z* is simply the sum over the Boltzmann factors of all *N* sites in the system. If each type of site occurs exactly *n* times then $N = n\ 4^M$. Assuming the presence of only one TF molecule is important in this context as it allows us to view all sites as being available for binding. Otherwise certain sites would be occupied and not accessible for other factors. In turn, *Z* would need to be adjusted accordingly, which would make the following calculations hard to carry out.

The connection between PWM scores and the binding probabilities derived by the above simplified biophysical approach will be illustrated below. To this end I follow the analytical derivation of *Z* as outlined by Heumann et al. (2004) but substituting their undefined binding energy term with the binding energy as given by equation (3.14). To arrive at the notation used by PWMs we can start by noting that each type *t* of sites contributes to the partition function a certain value $N\ f_t e^{-\beta E_t}$, where f_t is the fraction of sites of the type *t* in the system. If all types of sites occur equally often then f_t can be expressed in terms of the individual base frequencies, b_α as:

$$f_{\mathrm{t}} = \prod_{m=1}^{M} \prod_{\alpha=A,C,G,T} b_{\alpha}^{s_t(m,\alpha)} \qquad (3.16)$$

where $s_t(m,\alpha)$ is a site specific selector function that is either 1 if α matches the observed base in the site or 0 otherwise. By assuming that all sites occur equally often the base frequencies have to be 0.25 for all four bases. After replacing E_i with the mismatch energy description of Berg and von Hippel (equation 3.14), the contribution of each type of sites to the partition function is now:

$$N\, f_{\mathrm{t}} \mathrm{e}^{-\beta \mathrm{E_i}} = N \prod_{m=1}^{M} \prod_{\alpha=A,C,G,T} \left(b_{\alpha} \mathrm{e}^{-\beta \varepsilon_{\alpha}} \right)^{s_t(m,\alpha)} = N \prod_{m=1}^{M} \prod_{\alpha=A,C,G,T} \left(\mathrm{b}_{\alpha} \left(\frac{n_{m,\alpha}}{n_{m,0}} \right)^{\frac{1}{\lambda}} \right)^{s_t(m,\alpha)} \qquad (3.17)$$

where $n_{m,\alpha}$ and $n_{m,0}$ correspond to the counts for the observed base α and the most frequent base in the PSCM of the corresponding TF and λ is the scaling parameter for the mismatch energies. With adding up the individual contributions from all 4^M types of sites, the partition function becomes:

$$Z = N \sum_{t=1}^{4^M} \prod_{m=1}^{M} \prod_{\alpha=A,C,G,T} \left(\mathrm{b}_{\alpha} \left(\frac{n_{m,\alpha}}{n_{m,0}} \right)^{\frac{1}{\lambda}} \right)^{s_t(m,\alpha)} . \qquad (3.18)$$

For a hypothetical factor with binding site length 3 there exist 4^3 different sites:

AAA
AAC
AAG
. . .
TTT.

For these 4^3 sequences we can write down the terms from equation (3.18) by denoting $b_A \left(\frac{n_{1,A}}{n_{1,0}} \right)^{\frac{1}{\lambda}}$ as A_1, $b_A \left(\frac{n_{2,A}}{n_{2,0}} \right)^{\frac{1}{\lambda}}$ as A_2 and so forth. Using this annotation and assuming that every type of site occurs only once we get for Z:

$$\begin{aligned}
&A_1\, A_2\, A_3 \;+\; A_1\, A_2\, C_3 \;+\; A_1\, A_2\, G_3 \;+\; A_1\, A_2\, T_3 \;+\\
&A_1\, C_2\, A_3 \;+\; A_1\, C_2\, C_3 \;+\; A_1\, C_2\, G_3 \;+\; A_1\, C_2\, T_3 \;+\\
&A_1\, G_2\, A_3 \;+\; A_1\, G_2\, C_3 \;+\; A_1\, G_2\, G_3 \;+\; A_1\, G_2\, T_3 \;+\\
&\ldots\\
&T_1\, T_2\, A_3 \;+\; T_1\, T_2\, C_3 \;+\; T_1\, T_2\, G_3 \;+\; T_1\, T_2\, T_3 \;=
\end{aligned}$$

$$Z = \left(\mathrm{A}_1 + \mathrm{C}_1 + \mathrm{G}_1 + \mathrm{T}_1\right)\cdot\left(\mathrm{A}_2 + \mathrm{C}_2 + \mathrm{G}_2 + \mathrm{T}_2\right)\cdot\left(\mathrm{A}_3 + \mathrm{C}_3 + \mathrm{G}_3 + \mathrm{T}_3\right).$$

By generalizing for any TF with binding motif of length *M* we obtain:

$$Z = N\prod_{m=1}^{M}\sum_{\alpha} b_{\alpha}\left(\frac{n_{m,\alpha}}{n_{m,0}}\right)^{\frac{1}{\lambda}} . \tag{3.19}$$

By replacing *Z* in equation (3.15) we can now compute the probability, p_i, of a particular site being bound:

$$p_i = \frac{1}{Z}\left(\prod_{m=1}^{M}\frac{n_{m,\alpha}}{n_{m,0}}\right)^{\frac{1}{\lambda}} = \left(\prod_{m=1}^{M}\frac{n_{m,\alpha}}{n_{m,0}}\right)^{\frac{1}{\lambda}}\left[\frac{1}{N}\prod_{m=1}^{M}\sum_{\alpha} b_{\alpha}\left(\frac{n_{m,\alpha}}{n_{m,0}}\right)^{\frac{1}{\lambda}}\right]^{-1} = \frac{1}{N}\prod_{m=1}^{M}\frac{\left(n_{m,\alpha}\right)^{\frac{1}{\lambda}}}{b_{\alpha}\sum_{\alpha}\left(n_{m,\alpha}\right)^{\frac{1}{\lambda}}} \xrightarrow{\text{with } \lambda=1}$$

$$p_i = \frac{1}{N}\prod_{m=1}^{M}\frac{\nu_{m,\alpha}}{b_{\alpha}} . \tag{3.20}$$

Except for the scaling factor 1/*N* this quantity is identical to the result obtained by Heumann et al. (1994) for their optimal position specific weights and the likelihood ratio described in Section 3.1.3. To see that the above result is intuitively correct imagine a PFM with $v_{m,0}$ = 1 for one base in each column and $v_{m,\alpha}$ = 0 for all other elements. If we assume that all types of sites occur exactly once then it follows that $N = 4^M$ and b_α = 0.25. The binding probability is thus computed as:

$$p_i = \frac{1}{4^M}\prod_{m=1}^{M}\frac{1}{0.25} = 1 .$$

From the physical point of view, assuming $v_{m,0}$ = 1 for every position in the matrix means mismatch energy *E* = 0 for the consensus and $E \to \infty$ for all other sites. The consensus site is therefore the only site that could be bound by the TF molecule. In turn, the probability of the consensus to be bound is 1 in agreement with the above computation.

When computing *Z* according to the simplifying assumptions made above then the binding probabilities as derived by this biophysical model are identical to the exponential of the PWM scores introduced in Section 3.1.3, except for the scaling factor 1/*N*. In turn, the ranking of sites according to the two approaches is identical. This relationship is frequently used to motivate a biophysical interpretation of PWM scores (Frith et al., 2004; Fields et al., 1998; Stormo, 2000). However, it is important to recall the assumptions required for the two approaches to converge. Most notably, it requires that there is a single TF molecule in the cell. Only in this case is the probability of a given site *i* (energy state *i*) to be occupied by a TF molecule equivalent to its Boltzmann factor normalized by the simple molecular partition function used above. This is so because the partition function used implicitly assumes that all energy levels (DNA sites) can be occupied by an infinite number of molecules (Appendix A). With DNA sites representing the available energy levels of the system this is not the case as every site can be occupied by at most one factor. Secondly, the scaling parameter *λ* for the mismatch energies was set to 1, which for most TFs will not be the optimal value. Lastly, the genome in which the sites are located was assumed to contain an equal fraction of all possible sites. In Chapter 4 I will show the derivation and application of a more realistic binding model that avoids all of the above assumptions. This model will in turn be shown to yield higher correlation between predicted and actual binding probabilities than the above model. In conclusion, while the exponential of PWM scores (the likelihood ratio between PFM and background model) can be viewed as related to binding probabilities the statistical and biophysical approach do converge only under special conditions. The last section of this chapter will introduce several state of the art bioinformatics methods that use either the statistical or the above simplified biophysical approximation in order to detect TFs regulating groups of genes.

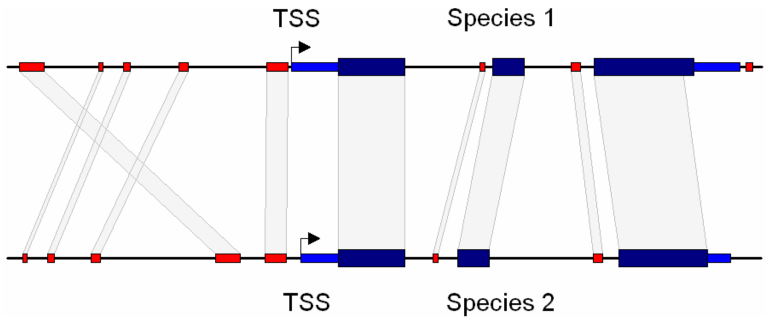
3.2 Predicting regulating TF

The previous section focussed on methodologies used to identify likely TF binding sites in the genome. Here we will discuss how such predictions can be used to infer which TFs likely regulate a group of genes. Such gene groups arise in many different experimental settings. For instance, treating cultured cells with cortisol might cause the upregulation of a number of genes. Given these genes, a natural question is then which TF was mediating the signal and finally causing the altered gene expression. Similarly, one might asked which TFs are responsible for tissue specific gene expression or the activation of stress response genes. There are a number of different approaches concerned with identifying TF that regulate cohorts of genes. Three frequently cited methods addressing this question are **oPOSSUM** (Sui et al., 2005), **PAP** (Chang et al., 2007) and **CLOVER** (Frith et al., 2004). oPOSSUM relies on statistical TF binding site predictions while CLOVER and PAP use a more physical motivated model. To reduce false binding site predictions oPOSSUM and PAP both make heavy use of phylogenetic footprinting a technique used to restrict the sequence space to regions with likely regulatory function. After a brief review of phylogenetic footprinting the main characteristics of the above three methods will be outlined.

3.2.1 Phylogenetic footprinting

The concept of phylogenetic footprinting is based on the idea that important regulatory sequences are retained during evolution while all other non-coding sequences are free to mutate. To identify such sequences the open reading frame, typically the most strongly conserved part of a gene, is used to first find likely orthologs (genes in different species stemming from the same common ancestral gene) in other species. Given a pair of orthologs their promoter regions can subsequently be aligned and conserved sequence blocks can be identified (see Figure 3.8) using dedicated alignment programs such as BLASTZ. The higher the conservation the more likely it appears that a given sequence block plays an important role in the gene regulation. It has been estimated that about 70% of all functional TF binding sites are conserved between mouse and human (Lenhard et al., 2003, Dermitzakis et al., 2002). At the same time on average only 20% of the promoter regions are conserved. Thus the number of non-functional binding site predictions will be cut down by a factor of ~5 while the number of false negative predictions increases only slightly when restricting the search to conserved blocks. Given the high prevalence of predicting non-functional binding sites that occur by chance in the genome most methods thus scan only the conserved sequence elements found in a given promoter for TF binding sites.

Figure 3.8 – Phylogenetic footprinting



Shown is the alignment between orthologous genes from two related species such as human and mouse. Dark blue boxes indicate the open reading frame of the genes, narrow light blue boxes the 5' and 3' untranslated regions and red boxes indicate non-coding evolutionary conserved sequence blocks. The ORF is usually the best conserved part of a gene and can be used to identify orthologs between species. Conserved non-coding sequence blocks located near the ORF indicate the presence of important regulatory elements. As shown such elements can relocate in respect to the TSS.

3.2.2 oPOSSUM

oPOSSUM (Sui et al., 2005) uses vertebrate PWMs stored in the JASPAR database (Sandelin et al., 2004) to scan all highly conserved non-coding regions located within 5kb upstream and downstream of all orthologous human and mouse genes. In order to reduce spurious predictions only PWMs with information content > 8 bits are utilized. The information content of a PWM is thereby defined as the total Kullback-Leibler entropy distance (Kullback et al., 1951) between the underlying PFM and a background model $\{b_\alpha\}$, which can be computed by the following equation:

$$I = \sum_{m=1}^{M} \sum_{\alpha=A,C,G,T} v_{m,\alpha} \log\left(\frac{v_{m,\alpha}}{b_\alpha}\right) \tag{3.20}$$

where $v_{m,\alpha}$ is the frequency of base α at position *m* in the PFM of length *M*. In the following, a TF binding site is annotate if its PWM score from Section 3.1.3 lies in the top 15% of scores and the position of the site is conserved between mouse and human. Once the binding sites have been annotated for all human or mouse genes oPOSSUM uses two statistical tests to check for the overrepresentation of binding sites of a given TF in the user defined input gene list in comparison to the background set of all mouse or human genes. The first test computes the following z-score statistics:

$$z = \frac{x - \mu - 0.5}{\sigma} \tag{3.21}$$

with *x* being the number of binding sites for a given TF in the input gene set, *µ* being the expected number of binding sites for the input set and *σ* being the standard deviation for the binding site predictions assuming a binomial distribution. *µ* is thereby given as:

$$\mu = B\frac{n}{N} \tag{3.22}$$

where *B* is the number of predicted binding sites in the background gene set, and *n* and *N* are the number of nucleotides scanned in the input set and background set, respectively. Finally, due to the binomial assumption the standard deviation can be computed by:

$$\sigma = \sqrt{nP(1-P)} \tag{3.23}$$

where *P* is the probability of a hit for the TF, which is given by *B / N*.

As a second statistical test oPOSSUM computes the Fisher exact probability that k genes in the input set have at least one TF binding site as compared to the number of genes with at least one binding site in the background set. TFs with large z-score or small Fisher exact p-value are subsequently considered as likely regulators of the input gene set. As we will see in Chapter 4 a likely shortcoming of the method is that it relies on predefined score cutoffs. These cutoffs do not allow to distinguish between a consensus binding site and a site just surpassing the threshold. In addition, for the Fisher exact test, due to the large number of false positive predictions assuming that any hit in a promoter indicates a TF target makes it difficult to find a meaningful target enrichment in the input gene set.

3.2.3 PAP

PAP (Chang et al., 2007) starts by computing binding site scores for a given PWM and a set of sequences as described in section 3.1.3. Scores above a given threshold are in the following considered to be directly proportional to the probability of the TF being bound to the corresponding site *i* via the relation:

$$P(site_i = bound \mid score_i) \propto e^{score_i} \tag{3.24}$$

while all scores below the threshold are ignored. For a sequence with several hits for the same factor the probability score, P_{Seq}, for the entire sequence is then calculated as:

$$P_{Seq} = \sum_{i=1}^{L} e^{score_i} \ . \tag{3.25}$$

The resulting probability scores are used to rank all *N* sequences from highest to lowest P_{Seq}. Subsequently the ranks are converted into rank-order values, *R*, in the following way:

$$R = \ln(N) - \ln(rank) \tag{3.26}$$

Finally, to detect TFs associated with a given input gene set the average rank value across all genes in the set is computed:

$$\langle R \rangle = \frac{1}{n} \sum_{seqs} R \ . \tag{3.27}$$

The TFs with largest average ⟨*R*⟩ are considered likely regulators of the input gene set. As mentioned in section 3.1.5 while the log likelihood scores are related to the binding free energy the value of e^{score} is directly proportional to the binding probability only under special assumptions. Since these assumptions are presumably not met in general the magnitude of the contributions in equation (3.25) from the individual sites to the total affinity score of a longer sequence will be distorted.

Like oPOSSUM, PAP scans only sequence blocks conserved between orthologous genes from various species. It thereby includes UTRs, introns and exons. Promoter regions are extended to 10kb upstream of the TSS and 5kb downstream of the respective gene. It should be noted that scanning the ORFs is somewhat contrary to the concept of using phylogenetic footprinting since the coding region of a gene is likely conserved only to preserve the amino acid sequence of the protein and not because of the presence of TF binding sites. The web interface through which one can access PAP does not allow to select the size of the upstream and downstream regions to be scanned but searches the conserved blocks in the defined range around the TSS. The user thus only has to specify gene identifiers to start the search for associated TFs.

3.2.4 CLOVER

As input, Clover (Frith et al., 2004) requires a set of PFMs, a set of DNA sequences for which one seeks regulating TFs and in addition, a set of background sequences. Clover starts, in complete agreement with PAP, by computing the binding site scores, called *LR1* in this application, according to the PWM model outlined in section 3.1.2. The obtained scores are again viewed as being directly proportional to the probability of the TF occupying a site. However, in contrast to equation (3.25) here the average likelihood ratio of a TF sitting anywhere on a sequence of length *L* is computed as:

$$LR2 = \frac{1}{L}\sum_{i}^{L} e^{score_i} \tag{3.28}$$

where *score_i* is the PWM score obtained for site *i* in the sequence. The authors interpret this average likelihood ratio as the probability of the TF binding to the sequence. However, the quantity *LR2* rather represents the average expected number of TFs bound per site in a given sequence as the authors perform a sum over individual binding probabilities from all sites in a sequence and divide by the length of the sequence. Nevertheless, by reasoning that it cannot be expected that a given TF regulates all genes in an input gene set of total size *N* the authors proceed by computing the “likelihood ratio” for a motif being present across only a subset containing *n* genes. This “likelihood ratio” is calculated as:

$$LR3a = \prod_{k=1}^{n} LR2_k \ . \tag{3.29}$$

Since there are C_n ways of selecting *n* genes out of a set of size *N* the program subsequently computes the average likelihood ratio of a TF binding all sequences in any subset of size *n*. This average is given by:

$$LR3b = \frac{1}{C_n} \sum_{l} LR3a \tag{3.30}$$

where the index *i* runs over all possible subsets of size *n*. The final score is computed as the log of the average over all possible values of *n* $(1 \leq n \leq N)$:

$$LR4 = \ln\left(\frac{1}{N} \sum_{n=1}^{N} LR3b \right). \tag{3.31}$$

The magnitude of the final score is dependent on the size and quality of the applied PWM. Therefore the scores obtained for a given input gene set are compared to scores obtained from a resampling procedure that randomly picks sequences of identical length from the control sequence set. As with PAP, the interpretation of *LR2* in equation (3.31) as a binding probability is valid only under special circumstances. Particularly *LR3a* relies on adequate binding probabilities however, as spurious probabilities might strongly effect the product. PAP in contrast to Clover largely avoids this problem by using its P_{Seq} scores to simply rank all sequences (and ignores all scores < 0). In addition, although the authors have found an efficient way to compute the scores of all possible subsets of size *n* the required resampling procedure is computationally expensive and causes long runtimes.

Having outlined the major concepts for TF binding site predictions and the detection of TFs that regulate groups of genes the next chapter will introduce the TRAP binding model and its application to yeast.