CHAPTER 4
The TRAP method

The following work is based on the biophysical model outlined in section 3.1.3. In particular equation (3.14), page 45, will play a key role as it allows computing the mismatch energy E_i between a given TF and a given site *i* in the genome. In section 3.1.5 it was outlined how the probability of a site *i* to be bound by a TF can be computed from the mismatch energy, when assuming that TFs bind to sites according to Boltzmann statistics. This model, which uses as simplifying assumption that all DNA sites are available for binding, is applicable if the TF concentration is low (ideally one TF molecule per genome) and competition between TFs for the same sites can be neglected. In principle, the molecular partition function of the Boltzmann distribution introduced in equation (3.15) could be computed also for the case of multiple TFs per genome, however, preclusion effects introduced by a TF blocking a certain site for the other TFs quickly makes this an intractable problem.

Because of these limitations I use a different model to predict the binding probability of a TF to a given site. The derivation of this model, which requires next to mismatch parameter λ only one additional parameter to be determined, will be outlined below. The resulting approach for ***T**ranscription factor **A**ffinity **P**redictions* (**TRAP**) not only avoids the assumptions required for applying the simplified Boltzmann distribution but also has a number of advantages over hit-based methods. Most notably, TRAP provides a natural ranking of sequences with respect to a particular transcription factor or conversely the ranking of several TFs with respect to one sequence. It does so by integrating weak and strong binding signals across a longer stretch of DNA such as a promoter region. To quantify the improvements I compare the results from TRAP with traditional hit-based approaches and find that it has higher predictive power over experimental binding data. As will be shown, TRAP improves the correlation between predictions and experimental measurements also in comparison to methods which use Boltzmann statistics for binding energy predictions.

4.1 Deriving the TRAP model

I first derive the probability that a given site in the genome is bound by a TF. While the resulting measure has been used previously for deriving biophysically motivated binding site cutoffs (Shreiman et al., 2005) the derivation below nicely illustrates the physical interpretation for the unknown parameter that arises in the model. The binding probabilities for individual sites will then be used to obtain a measure for the affinity of a TF to a larger

genomic region such as a promoter. Correlating this affinity measure with experimental ChIP-chip data subsequently allows for tuning the model parameters. Finally, I will introduce a general prescription for setting the parameters in the absence of experimental binding data.

4.1.1 Obtaining TF binding probabilities from mismatch energies

For the derivation of binding probabilities we can start by considering a solution containing identical molecules of a given TF and DNA sites of type S_i, with identical sequence and length of the TF motif. The binding and dissociation reactions taking place between the TF molecules and these sites can be described by the chemical reaction:

$$TF + S_i \underset{V_D}{\overset{V_B}{\rightleftharpoons}} TF{\sim}S_i \tag{4.1}$$

where TF, S_i and TF~S_i are the free TF molecules, free DNA molecules and bound TF-DNA complexes, respectively (Atkins, 2007). The speed of the binding reaction, V_B, is given by the laws of chemical kinetics:

$$V_B = [\mathrm{TF}][\mathrm{S_i}] \cdot a_B \cdot e^{-\beta \Delta E_{A,U}} \tag{4.2}$$

where square brackets indicate the concentration (or more accurately, the activity of the molecules), a_B is a reaction specific constant and $\Delta E_{A,U}$ is the energy barrier that the molecules need to overcome in order to bind to each other (see Figure 4.1 for details). In this reaction the energy barrier will be small as no chemical bonds need to be broken and only associated water molecules need to be displaced. Similarly, the speed of the dissociation reaction V_D is given by:

$$V_D = [\mathrm{TF} \sim S_i] \cdot a_D \cdot e^{-\beta \Delta E_{A,i}} \tag{4.3}$$

where a_D is the reaction constant and $\Delta E_{A,i}$ is the kinetic energy required to split the TF-DNA complex (see Figure 4.1). Once equilibrium has been reached the binding and dissociation reactions occur with equal speed, that is:

$$\frac{V_D}{V_B} = \frac{[\mathrm{TF} \sim S_i] \cdot a_D \cdot e^{-\beta \Delta E_{A,i}}}{[\mathrm{TF}][\mathrm{S_i}] \cdot a_B \cdot e^{-\beta \Delta E_{A,U}}} = 1. \tag{4.4}$$

The ratio between bound and unbound molecules is now given by the mass action law:

$$K_i = \frac{[TF \sim S_i]}{[TF][S_i]} = a_i \, e^{\beta \Delta E_{U,i}} \tag{4.5}$$

where K_i is the binding constant of the reaction, which can be approximated by the Arrhenius equation shown on the right hand side (Atkins, 2007). The constant a_i is referred to as the reaction specific Arrhenius constant and $\Delta E_{U,i}$ corresponds to the energy change associated with the binding of the TF to the site. As illustrated in Figure 4.1, the term $e^{-\beta \Delta E_{U,i}}$ approximates how many molecules in the unbound state have enough kinetic energy to cross the energy barrier as compared to molecules in the bound state. We proceed by assigning an energy of 0 to the complex between TF and its consensus site and setting the energy of all other TF-DNA complexes in respect to it (see Figure 4.1). In this case equation (4.4) can be rewritten as:

$$K_i = \frac{[TF \sim S_i]}{[TF][S_i]} = a_i \; e^{\beta(\Delta E_{U,0} - \Delta E_{i,0})} \; . \tag{4.6}$$

where $\Delta E_{U,0}$ is the energy change associated with the complex formation of unbound TF and its consensus site. The second energy term, $\Delta E_{i,0}$, measures how much stronger the TF binds to its consensus site as compared to site *i*. This energy difference corresponds exactly to the mismatch energy given by equation (3.14) on page 45. Using this relation one can now describe the binding reaction of the TF to any site *i* in terms of the binding energetics of the TF-consensus reaction. To this end equation (4.5) is solved for $\Delta E_{U,0}$:

$$e^{\beta \Delta E_{U,0}} = \frac{1}{a_i} \frac{[TF \sim S_i]}{[TF][S_i]} e^{-\beta \Delta E_{i,0}}, \tag{4.7}$$

which simplifies in case of the consensus site to:

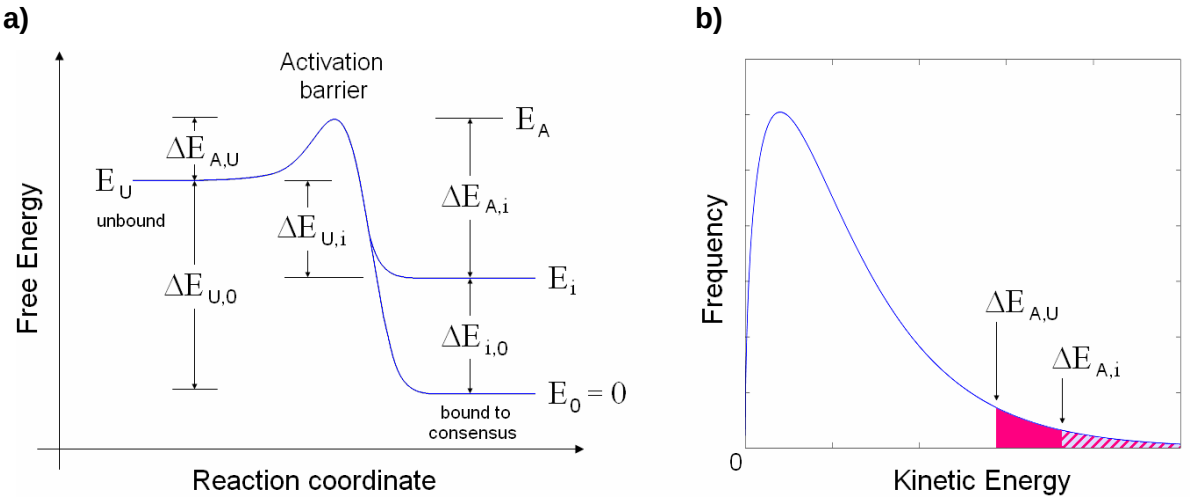
$$e^{\beta \Delta E_{U,0}} = \frac{1}{a_0} \frac{[TF \sim S_0]}{[TF][S_0]}. \tag{4.8}$$

Combining equations (4.7) and (4.8) and assuming that the Arrhenius constants a_i and a_0 have approximately the same value yields:

$$\frac{1}{a_0} \frac{[TF \sim S_0]}{[TF][S_0]} = \frac{1}{a_i} \frac{[TF \sim S_i]}{[TF][S_i]} e^{-\beta \Delta E_{i,0}} \longrightarrow \frac{[TF \sim S_i]}{[S_i]} = e^{-\beta \Delta E_{i,0}} \frac{[TF \sim S_0]}{[S_0]} = e^{-\beta \Delta E_{i,0}} R_0 \tag{4.9}$$

where the ratio between the concentrations of bound consensus sites and free consensus sites has been abbreviated as R_0. Considering a_i and a_0 to be similar is valid because the Arrhenius constants measure how often a collision between molecules with sufficient kinetic

Figure 4.1 – Kinetics of TF-DNA interaction



a) The binding of a TF to a DNA site with sequence *i* goes in hand with a change in the free energy of the system. Binding to the site with consensus sequence thereby causes the largest energy change $\Delta E_{U,0}$. The energy levels of all other TF-DNA complexes can be measured relative to the energy level E_0. The difference between E_0 and E_i hereby corresponds to the mismatch energy from equation (3.14). b) Maxwell-Boltzmann distribution for the kinetic energy of the molecules. The red and red-blue areas under the curve correspond to the number of molecules with energy larger than $\Delta E_{A,U}$ and $\Delta E_{A,i}$, respectively. The area can hereby be estimated according to the Arrhenius law as $e^{-\beta E}$.

energy leads to a reaction (Atkins, 2007). In this respect the requirements are similar for any type of site, therefore $a_i \approx a_0$. It should be noted that the transcription factor concentration cancels out in the above equation. This is so because all sites added to the solution will see on average the same amount of TF molecules. The significance of equation (4.9) is that one can compute the ratio of free versus bound sites for any sequence just from their mismatch energies in respect to the consensus and the ratio R_0 of bound versus free the consensus sites.

The probability of a particular site of type *i* to be bound by the TF is equivalent to the fraction of sites of this type that are bound in solution. This fraction of sites is given by:

$$f_i = \frac{[TF \sim s_i]}{[TF \sim s_i] + [s_i]} \tag{4.10}$$

where the denominator corresponds to the total amount of sites *i* in solution. Dividing each term in equation (4.10) by s_i and substituting the resulting $[TF \sim s_i]/[s_i]$ with the right hand

term of equation (4.9) one obtains the desired relation for the binding probability of the TF to the site:

$$p_i = \frac{e^{-\beta\Delta E_{i,0}}\ R_0}{1 + e^{-\beta\Delta E_{i,0}}\ R_0} = \frac{R_0\ e^{-\frac{1}{\lambda}\sum_{m=1}^{M}\ln\left(\frac{v_{m,\alpha}}{v_{m,0}}\right)}}{1 + R_0 e^{-\frac{1}{\lambda}\sum_{m=1}^{M}\ln\left(\frac{v_{m,\alpha}}{v_{m,0}}\right)}} \tag{4.11}$$

where $\beta E_{i,0}$ is the mismatch energy as computed by equation (3.14). Equation (4.11) can also be derived by assuming that each energy level, represented by a given type of site, can be occupied only once by the TF molecules. Applying this constraint to the Boltzmann distribution yields the Fermi-Dirac equation:

$$p_i = \frac{e^{-\beta(E_i - \mu)}}{1 + e^{-\beta(E_i - \mu)}} \tag{4.12}$$

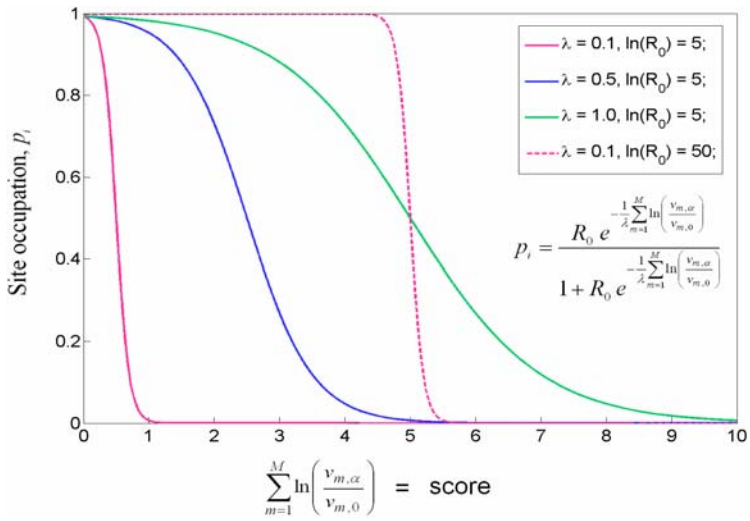
where the chemical potential, μ, takes the role of R_0 in equation (4.11). The physical meaning of R_0 is nicely illustrated by looking at equation (4.8) from which one obtains:

$$R_0\ = \frac{[TF \sim S_0]}{[S_0]} \propto [TF] \cdot e^{-\beta\Delta E_{U,0}}\ . \tag{4.13}$$

Thus R_0 combines the concentration of free TF molecules and the energy difference between the unbound and optimally bound states. In practice, neither the binding strength nor the concentration of a TF is known. R_0 thus constitutes next to λ from equation (3.14) a second unknown parameter that needs to be determined from experimental data in order to predict meaningful binding probabilities.

How the function p for the binding probabilities depends on the values of R_0 and λ is illustrated in Figure 4.2. The parameter λ thereby acts as shape parameter that specifies how smoothly the function p changes when introducing deviations form the consensus sequence, while R_0 acts as a location parameter for the function. Discrete hit based approaches thereby appear as a special case of the continuous binding model when choosing R_0 and λ accordingly. How to optimally set the parameters and how much information is lost when running into the regime of discretized binding probabilities will be outlined in the remainder of this chapter.

Figure 4.2 – Dependence of binding probabilities on R_0 and λ



The binding probabilities, p, of sites depend on the chosen values for R_0 and λ. The x-axis corresponds to the scores (summed log ratios of observed versus consensus base frequencies) of a given PFM. The parameter λ converts these scores into mismatch energies (via equation 3.14). With $\lambda \rightarrow 0$ (resulting in the red dashed curve) the model approaches the hit based schemes, that is, all sites with scores above or below a certain threshold (score $\approx$ 5, for the red dashed curve) are considered unbound ($p_i = 0$) or bound ($p_i = 1$), respectively. The location of the infliction point of the probability curves along the x-axis is thereby determined by the location parameter R_0.

It has to be stressed that this binding model assumes that all sites in the genome see the same effective TF concentration. While this assumption will not hold in all cases, TF molecules have been shown to diffuse quickly through the nucleus (e.g. Sprague et al., 2004, Zadeh et al., 2007) and thus assuming a similar effective concentration across the genome should constitute a reasonable approximation.

The following section will describe how the binding probabilities can be used to define an affinity measure for longer sequence and how the parameters R_0 and λ can subsequently be derived from ChIP-chip binding data.

4.1.2 Deriving a binding measure for longer sequences

Using equation (4.11) one can in principle compute the binding probability of a TF to any site in the genome. However, for many applications such as target gene prediction for a given TF, the question is rather whether or not the factor binds to a larger stretch of regulatory

DNA, such as a promoter region. My measure of choice for the affinity of a TF to a larger sequence is the expected number of TFs bound to the region, ⟨N⟩, which is computed by:

$$\langle \mathrm{N} \rangle = \sum_{i=1}^{L-M+1} p_i \ . \tag{4.14}$$

where L is the length of the region, M is the width of the TF motif and p_i is the binding probability to site i in the region. The term affinity will be used synonymously for ⟨N⟩ throughout the thesis. To account for competitive binding of a given TF to the same site but different strands the binding probabilities of the forward and reverse DNA strand are combined using the following approximation:

$$p_i = p_i^F + p_i^R \left(1 - p_i^F\right) \tag{4.15}$$

where p^F_i and p^R_i are the probability of binding to site i in the forward and reverse strand, respectively, which are given by equation (4.11). Note that the term (1 - p^F_i) corresponds to the probability that site i is free for binding of the TF to the reverse strand. Accounting for competition becomes relevant only if both p^F_i and p^R_i are large, which might be the case especially for palindromic binding motifs.

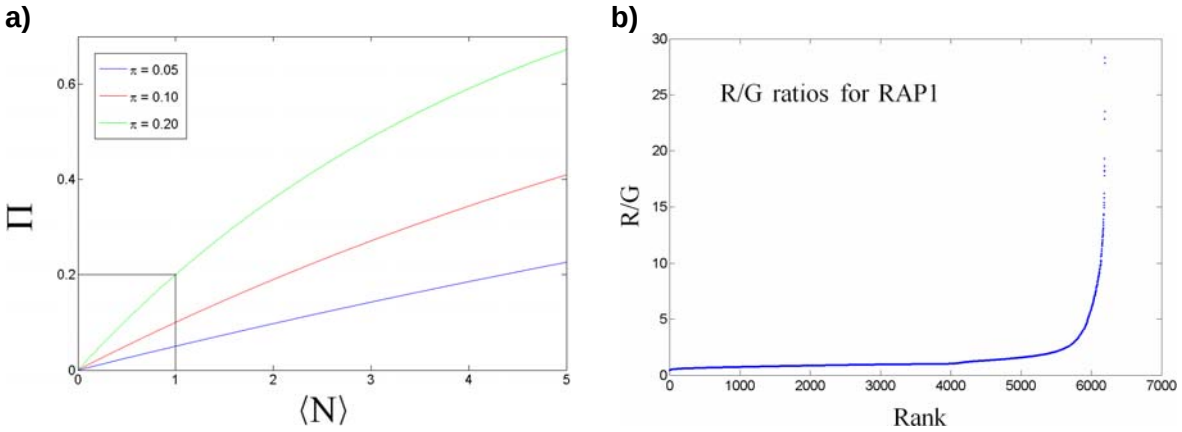
4.1.3 Deriving the TRAP parameters

The following section will outline how the two parameters R_0 and λ can be determined in the presence and subsequently the absence of experimental binding data. The biophysical model outlined above, with the computation of ⟨N⟩ as the affinity measure, together with the general prescription on how to derive the parameters, comprises the TRAP method. TRAP will be used in subsequent chapters to predict the binding affinity of a given TF to a given promoter region.

How are predictions and experimental binding data related?

The binding measure presented in equation (4.14) pertains perfectly to the situation tested with Protein binding microarrays (PBMs). As discussed in Section 2.5.2 such experiments quantify the binding strength between a TF and a longer stretch of DNA by measuring the amount of florescent light that is emitted from the labeled TFs bound to the given spot on the array. In this case the predicted number of TFs, ⟨N⟩, being bound to a certain sequence is expected to correlate linearly with the measured florescent light intensity. In contrast, in ChIP-chip experiments the amount of fluorescent light emitted from a spot on the microarray is expected to grow linearly with the number of sequences that are being pulled down in the

Figure 4.3 – Relationship between ⟨N⟩ and the pulled down efficiency



a) The probability of the antibodies to pull down a sequence with a single bound TF is given by Π = π. If π is small then ⟨N⟩ is expected to correlate nearly linearly with the number of pulled down sequences and subsequently with the observed *R/G* ratios. Curves indicate the theoretical relation between ⟨N⟩ and the probability, Π, of a sequence with ⟨N⟩ bound factors to be pulled down, given different values of π. b) Large *R/G* ratios for the general factor RAP1 from yeast do not show any apparent signs of saturation. Blue dots correspond to the experimentally evaluated sequences sorted according to their measured *R/G* ratios in a ChIP-chip experiment (Harbison et al., 2004). No plateau effect in large *R/G* ratios is observed even though several sequences are known to harbor clusters of RAP1 bind sites (Gilson et al., 1993). Similar inspection of the later used data sets confirmed the absence of any apparent saturation effects for all tested TFs.

antibody precipitation step. If we assume that the antibodies have probability π of pulling down a sequence with a single TF bound to it, then, for a sequence with ⟨N⟩ associated transcription factors the probability, Π, of being pulled down is given by:

$$\Pi = 1 - (1 - \pi)^{\langle N \rangle} \tag{4.17}$$

where $(1-\pi)^{\langle N \rangle}$ is the probability that none of the TF molecules bound to the sequence causes its precipitation (notice, in case of ⟨N⟩ = 1 then $\Pi = \pi$). If π is small, which is supported by the absence of any apparent saturation effects in the measured *R/G* ratios (see Figure 4.3b for an example), then ⟨N⟩ is expected to correlate linearly with Π (see Figure 4.3a) and subsequently with the measured *R/G* ratios. In fact, incorporating π as an additional parameter into the model had no significant effect on the found correlations. Therefore, in the following sections only the results as obtained when assuming that ⟨N⟩ and the *R/G* ratios correlate linearly will be presented.

Experimental binding data

As a first data set to be correlated with the binding predictions I utilized the PBM binding data available for the three yeast transcription factors (Rap1, Mig1 and Abf1) from Mukherjee et al. (2004). In this study the binding between each factor and all 6725 intergenic regions from yeast was tested. The length of the regions varies from ~50 to ~1500bps. It has to be stressed again that this *in vitro* data set constitutes the optimal test case for the binding model as the measured affinity grows linearly with the number of TFs sitting on a given sequence and since competition with other proteins such as histones is excluded. For a more extensive dataset and to investigate the applicability of the model to *in vivo* data I also retrieved the binding data for the comprehensive genome-wide ChIP-chip dataset form Harbison et al., (2004), which provides *R/G*-ratios for > 200 TFs and the 6725 intergenic regions in yeast. For many factors the binding behaviour was thereby not only tested in rich medium condition (YPD) but also in stress conditions such as oxidative stress (induced by H_2O_2) or amino acid starvation (SM). The hybridizations were performed on the same microarrays as in the above PBM experiments.

For each of the data sets the authors compute p-values describing the significance of the experimentally measured *R/G* ratios. A p-value threshold of 10^{-3}, which will be utilized in later sections of this study, has thereby been suggested to indicate binding of the TF.

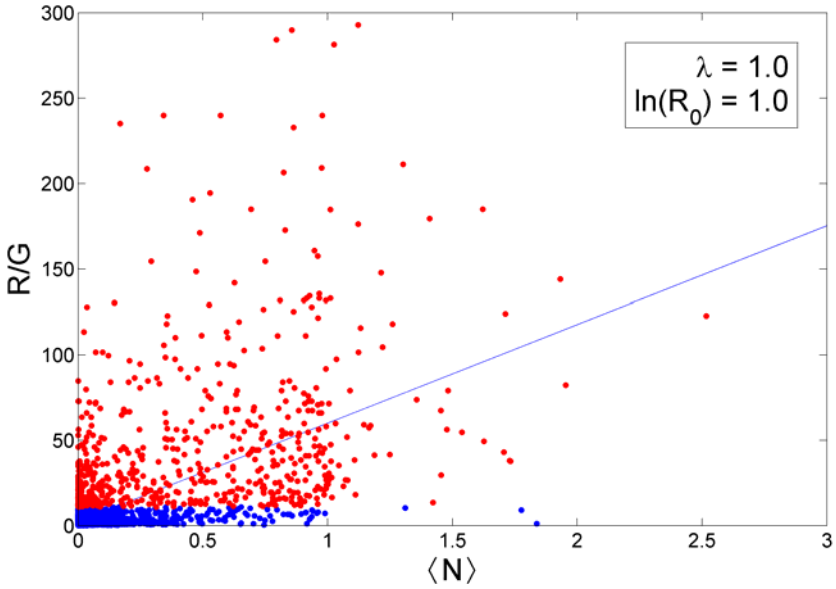
Utilized position frequency matrices

As motif descriptions for the TF binding predictions I used the set of 29 yeast matrices (for 25 different TFs) provided by the TRANSFAC database (Matys et al., 2003) for which PBM or ChIP-chip data is available. A pseudo-count of 1 was added to each element in the count matrices. In the framework of physically motivated binding models, this modification can be interpreted as setting a maximally allowed contribution to the mismatch energies. For comparative purposes also a PC = 0.5 (as recommended by Berg and von Hippel 1987) was used, but the results are unaffected by this change (data not shown). In addition to these matrices I also tested several PFMs derived in the Harbison et al., study (2004) by people from the Fraenkel lab. While many of these matrices yielded good results, they are not use for deriving the general parameter description for TRAP in order to avoid any possible circularity in the argument.

General parameter determination

To find the optimal setting of R_0 and λ for a given transcription factor equation (4.14) was applied to all 6725 intergenic regions in yeast and then correlate the predicted occupancies $\langle N \rangle$, with the measured binding intensities from the above large scale experiments. As an

Figure 4.4 – Correlation between predicted and measured TF affinities



A correlation of $r = 0.41$ is obtained for the factor ABF1 when randomly setting R_0 and λ to unity. Each spot represents an intergenic region whereby red and blue colours indicate whether a sequence is significantly bound by the factor according to experimenters. x and y axis correspond to model predictions and experimentally determined TF binding values, respectively. It has to be stressed that there are ~600 bound sequences (red) versus ~6100 unbound sequences (blue). The spots shifted towards higher values of ⟨N⟩ are thus greatly enriched with bound sequences.

example, Figure 4.4 shows for the transcription factor ABF1 the correlation between the actual binding values from the PBM experiment and the predictions made by the model when arbitrarily setting R_0 = 1 and λ =1. The quality of the correlation is hereby measured by the Pearson correlation coefficient *r*, which is given by:

$$r = \frac{\sum_{j=1}^{J} \left(\langle N \rangle_j - \mu_{\langle N \rangle}\right)\left(R_j / G_j - \mu_{R/G}\right)}{(J-1)\cdot \sigma_{\langle N \rangle} \cdot \sigma_{R/G}} \qquad (4.19)$$

where *J* is the total number of intergenic regions, μ is the population mean of the respective measure and $\sigma_{\langle N \rangle}$ and $\sigma_{R/G}$ denote the standard deviation of the binding predictions and *R/G* ratios, respectively (Magrab et al., 2000). For ABF1 with R_0 and λ both set to unity the observed correlation between binding predictions and real binding values is $r = 0.41$. The p-value for the significance of any measured correlation is estimated by means of a t-test (Magrab et al., 2000). To this end the following *t*-score is computed:

$$t = r\sqrt{\frac{J-2}{1-r^2}} \tag{4.20}$$

where J is again the total number of intergenic regions. Since J is in the order of several thousand the t-distribution follows the normal distribution and the corresponding p-value can be estimated from the area under the Gaussian to the right of the t-score. With this procedure one finds the above correlation to be highly significant (p-value < 10^{-100}).

The correlation coefficient (or alternatively the p-value) was used as quality measure to determine the values for R_0 and λ that optimize the correlation between the predicted number of TFs sitting on a sequence, $\langle N \rangle$, and the experimental measurements for the TF. To this end, for each TF and each data set all pair-wise combinations of the parameter values $\lambda \in \{0.05, 0.10, \ldots, 1.90, 2.00\}$ and $\ln(R_0) \in \{-10, -8, \ldots, 28, 30\}$ were tested. The result of this analysis is exemplarily shown in Figure 4.5a for the factor ABF1 and the PBM *in vitro* binding data set. For ABF1 an optimal correlation of 0.55 is thereby found when setting $\lambda = 0.65$ and $\ln(R_0) = 6.91$. In contrast, setting the parameters to the values that are assumed when applying the simplified Boltzmann statistics introduced in Section 3.14 (unphysiologically low TF concentration) and setting λ to 1 causes a significant drop in the observed correlation to a value of $r = 0.25$. Not surprisingly, nearly identical correlation (0.25) is obtained when using equation (3.25) from PAP (page 52) to compute the affinity scores.

General Features of the parameter space

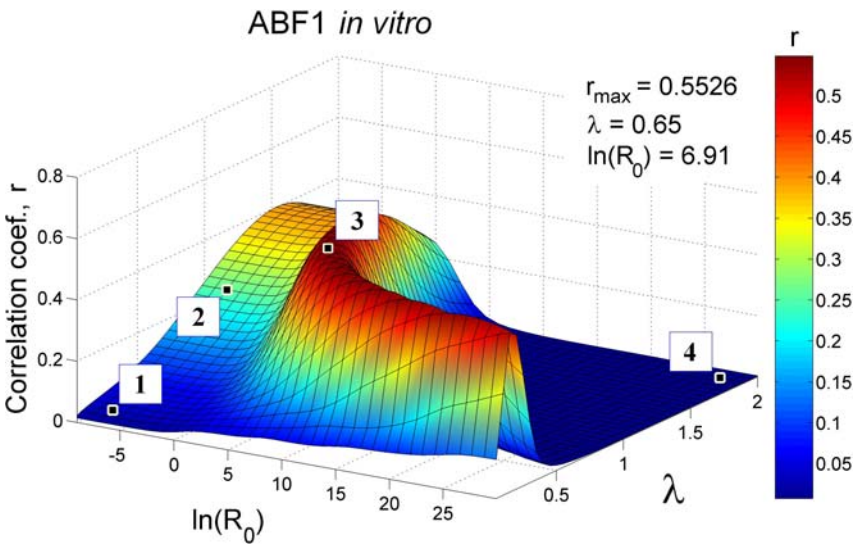
Figure 4.5 visualises the magnitude of correlation coefficient across the parameter space for the transcription factors ABF1 and GAL4. The resulting surface plots nicely illustrate several generic features of the parameter space found for nearly all factors.

For large λ and R_0 the correlation coefficient drops to ≈ 0. This is expected because setting λ to a value >> 1 corresponds to severely down scaling all mismatch energies and thus treating all sites like the consensus (see Figure 4.6 for a graphical explanation). In addition, using a large value for R_0 indicates high TF concentration and/or strong binding to the consensus (see Figure 4.6). Together this leads to nearly every site in a sequence being occupied by the TF with high probability. As a result $\langle N \rangle$ correlates linearly with the length of the individual regions. Only binding data from proteins that interact completely unspecifically with the DNA are expected to show high correlation in such a parameter setting. I have found only one factor (ROX1) for which this is the case.

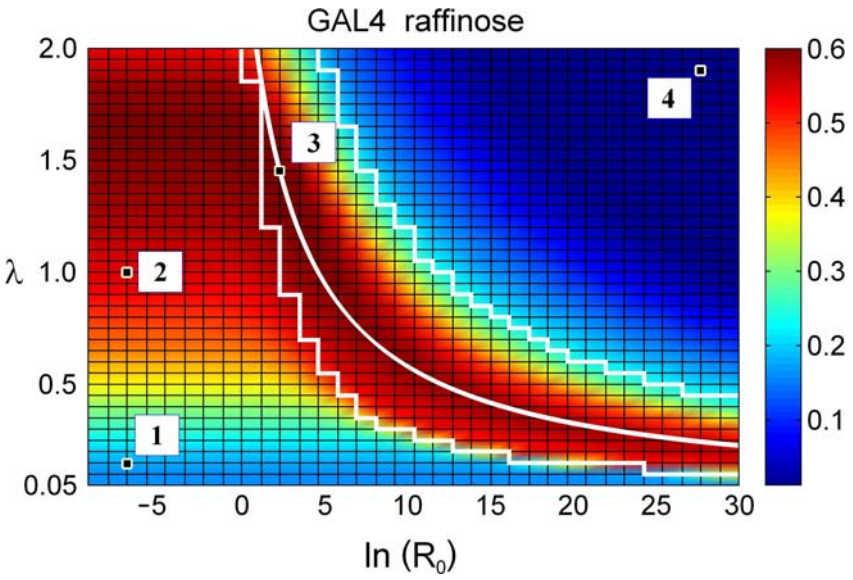
In contrast, setting λ to a value << 1 corresponds to introducing large mismatch energies even for small deviations from the consensus. Setting at the same time R_0 to a small value means that the TF concentration and/or the binding strength to the consensus

Figure 4.5 – Generic features of the parameter.

a)

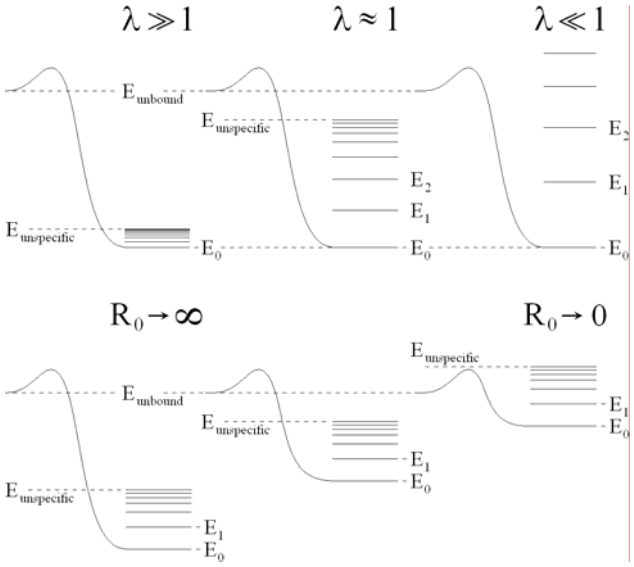


b)



a) The magnitude of the achieved correlation between predictions and binding values for each parameter combination is shown. Point 1 and 4 respectively correspond to the areas in the parameter space where all sequences remain either completely unbound or maximally occupied. Both of these extremes lead to negligible correlation with the binding data. Point 3 shows the parameter combination that results in the optimal correlation (r = 0.55) between predictions and binding values. Lastly, Point 2 indicates the parameters that correspond to assuming very low TF concentration and setting λ to 1. b) Top view on the parameter space for the factor GAL4 in raffinose growth medium condition. The optimal choice of parameters, with the highest correlation coefficient, is again marked by Point 3. The hyperbola highlights a line of parameter combinations with similarly high correlation coefficient. Indicated are also the boundaries (white staggered lines) for which the maximal value of ⟨N⟩ (over all 6725 intergenic regions) lies between 0.5 and 5.

Figure 4.6 – Effects of setting ln(R_0) and λ

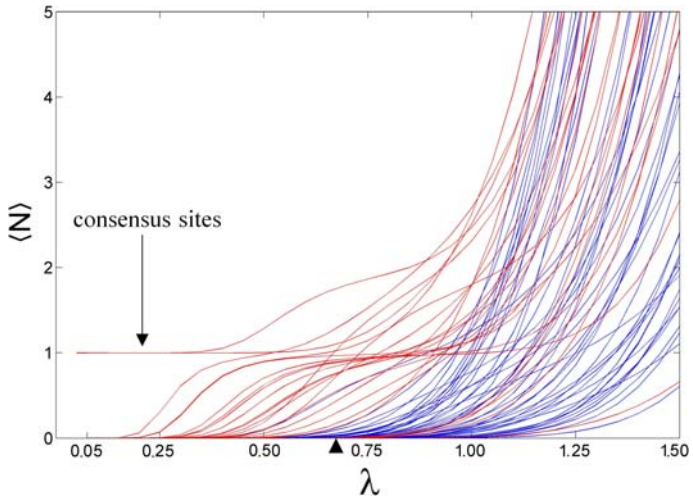


Top panel: λ determines the spacing between mismatch energy levels. For $\lambda >> 1$ mismatches are down weighted and all sites behave like the consensus. This scenario might apply to factors that bind only non-specifically to DNA. In contrast, $\lambda << 1$ introduces large mismatch energies. This can lead to an unrealistic setting where already the smallest deviation from the consensus causes a mismatch energy greatly overshooting the energy level of the unbound state. In any case, most transcription factors can accommodate certain variations in the binding site (Mossing et al., 1985) and therefore such a setting is not expected to yield high correlation between model predictions and binding data. Lower panel: Importantly, aside from TF concentration, R_0 also determines the binding energy between a TF and its consensus site. It thus positions the ground state energy level of the system. Setting R_0 to a too large or too small value causes all sites to be considered as bound or unbound, respectively.

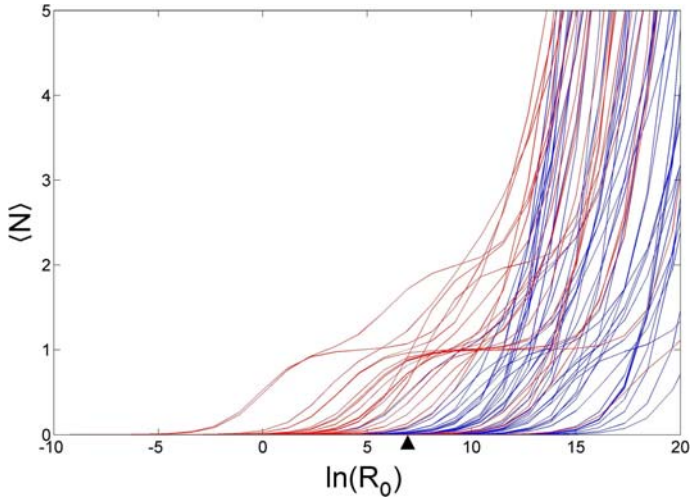
site are assumed to be very low. Thus, in this regime all sequences are predicted to be unbound by the TF which naturally does not reflect the real situation measured in the chip experiments. The obtained correlation from using these parameter settings is thus also ≈ 0 for all factors. As mentioned in context of Figure 4.2, choosing $\lambda \rightarrow 0$ but setting R_0 to an appropriately large value leads back to the regime of hit based methods. The higher correlation values that appear in the bottom right corner of the surface plot thus indicates what might be achieved by hit based methods if the score threshold is chosen properly (recall that the score computation is however different for the PWM and biophysical models as can be seen by comparing equation 3.14 and 3.4). Figure 4.7 illustrates for a group of 50 intergenic regions how the discussed changes of the parameters affect the individual binding affinities.

Figure 4.7 – Changes of ⟨N⟩ with λ and R_0

a)



b)



a) Growth in ⟨N⟩ for the factor ABF1 associated with increasing λ while keeping ln(R_0) fixed at 6.91 ($R_0 = 10^3$). Each line corresponds to an individual intergenic region where red signifies that the region is bound by the factor according to PBM measurements. At very small λ all sites except the consensus have mismatch energies $\rightarrow \infty$. At the same time, given that $R_0 = 10^3$ the consensus site has a binding probability of 1000 / 1001 $\approx$ 1 according to equation 4.11. This is reflected by two intergenic regions both of which contain a consensus site and thus approach a value of ⟨N⟩ $\approx$ 1 for λ < 0.50. Binding sites contained in other regions obtain appreciable binding probabilities only for λ > 0.25. The optimal value for λ (0.65) given ln(R_0) = 6.91 is indicated by a black triangle. b) Shown is the growth in ⟨N⟩ caused by increasing R_0 while keeping λ constant at 0.65. The two regions containing the consensus sites are again the first to obtain larger binding probabilities with growing R_0. Plateaus are reached when the sites obtain saturated binding probabilities. Regions bound according to experimenters (red) in general tend to obtain larger binding probabilities than unbound regions (blue). Optimal correlation is reached when setting ln(R_0) = 6.91 (black triangle).

Another general feature of the parameter space is that for small R_0 the expected number of bound TFs depends linearly on R_0, as can be inferred from a Taylor expansion of equation (4.14) around $R_0 = 0$, which yields:

$$\langle N \rangle \approx R_0 \sum_{i=1}^{L-M+1} e^{-\beta E_i} \,. \tag{4.21}$$

Equation (4.21) illustrates that changes of R_0 in this range only affect the absolute value of ⟨N⟩, but not the correlation of ⟨N⟩ with *R/G* ratios. In Figure 4.5 this is reflected by a constant correlation coefficient for $\ln(R_0) < 0$ and a given λ. Equation (4.21) also highlights the transition of the modelled Fermi-Dirac distribution into the regime of Boltzmann statistics.

Finally, it is evident from equation (4.11) that the affinity of a single site *i* can be kept constant for varying values of R_0 and E_i in such a way that $R_0 e^{-\beta E_i} = c$. With $\beta E \propto 1/\lambda$ there exists a hyperbolic relation:

$$\lambda \propto \frac{1}{\ln(R_0 - c)} \,. \tag{4.22}$$

Interestingly, the characteristic curves of high correlation coefficients seen in Figure 4.5, which can be well described by a hyperbola, suggest that this generic behaviour is effectively reflected in the behaviour of the correlation coefficients for all TFs.

Optimal Parameter Choice Derived from Experimental Data

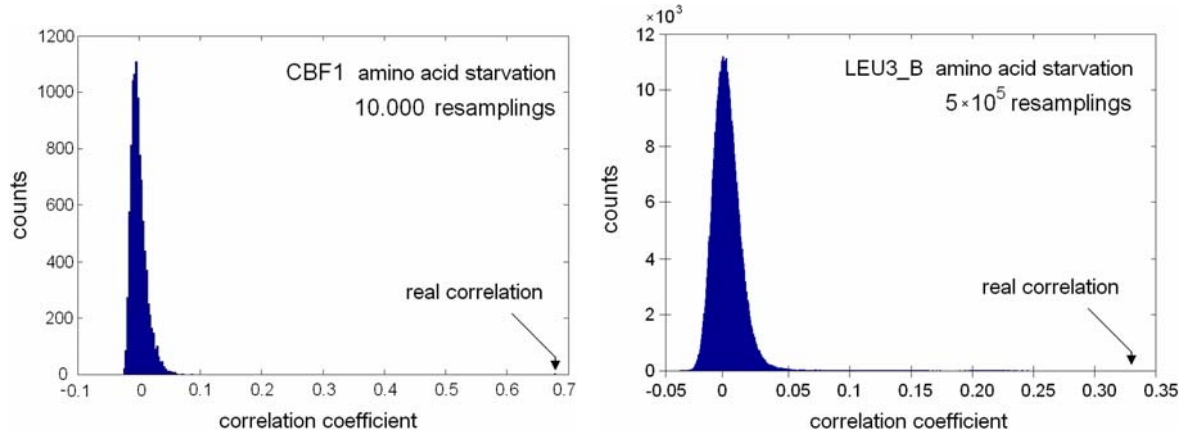
For all three PBM data sets optimal parameters have been found which yield highly significant correlation ($r > 0.5$, p-value < 1e-100), as shown in Table 4.1. This indicates that the binding model can successfully account for a considerable fraction of the observed *in vitro* binding affinities.

Next I analysed the more comprehensive ChIP-chip data set by Harbison et al., 2004. This *in vivo* data corresponds to a more complicated situation, where one cannot always assume that the transcription factor is available for DNA binding and that the DNA is accessible under the tested condition. Therefore, without additional information about competing TFs, chromatin structure and TF concentration, the model cannot always be expected to yield high correlation with the data. Despite these caveats, significant correlations are found for a large number of the *in vivo* data sets.

The t-score statistics outlined in equation (4.20) thereby produces significant p-values also for small correlation coefficients of ~0.05 suggesting that the assumptions underlying the t-test are violated in some cases. To more rigorously assess the quality of the optimal correlations I therefore performed a resampling test for each TF and each condition. To this

Figure 4.8 – Resampling verifies the significance of obtained optimal correlations

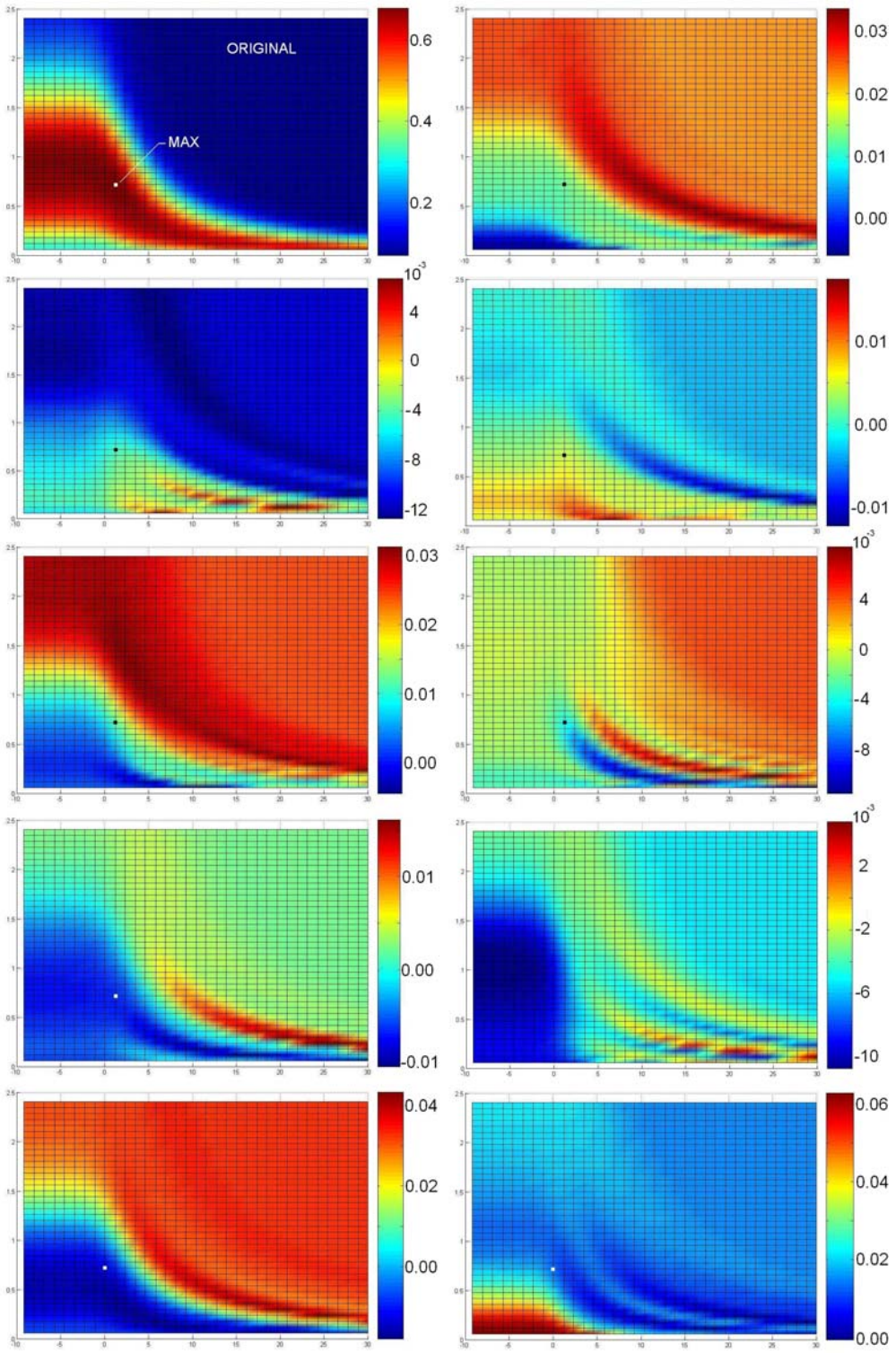
a) **b)**



a) The correlation coefficients obtained from binding data where the *R/G* values for CBF1 have been randomly shuffled over the 6725 intergenic regions never exceed a value of 0.08. The correlation of $r \approx 0.67$ obtained for the original ChIP-chip binding values (denoted as "real correlation") thus appears highly significant and is assigned a p-value < 1e-4. b) For LEU3, 500.000 reshufflings never yielded a correlation as high as the original ChIP-chip data set. The p-value of the correlation is thus $< 2\times10^{-6}$.

end the experimental *R/G* binding values were randomly shuffled between the 6725 intergenic regions. Subsequently the correlation between model and shuffled data was computed using the optimal parameters from the original data set. This procedure was repeated a minimum of ten thousand times for each factor. Figure 4.8 shows the results of this test exemplarily for the factors CBF1 and LEU3 in amino acid starvation condition. As the histogram shows, the correlation coefficients for the resampled data sets cluster around $r = 0$ and never produce a correlation as high as the original data. Similar results are found for all factors and conditions where the correlation with the original is larger 0.3. In these cases none of the resamplings achieved a correlation as high as the one obtained from the original data. In contrast, many correlations with $r < 0.2$ were found to be of low significance (p-value $> 10^{-4}$) by the resampling technique. The shuffling procedure was also performed across the entire parameter space, i.e. without using the optimal parameters, in order to exclude the possibility that other parameter pairs could yield a higher correlation by chance. The parameter space for ten such resampled data sets is shown in Figure 4.9 for the factor CBF1. In accordance with equation (4.22) the correlation coefficients obtained from the resampled data also show hyperbolic curves across the parameter space, however, the magnitude of the correlations never exceeds a value of 0.06.

Figure 4.9 – Correlations across the parameter space obtained by resampling



The top left plot shows the correlation coefficients obtained from the real data set while the other plot shows the correlation coefficients obtained from a particular reshuffling of the *R/G* ratios over al 6725 intergenic regions. While hyperbolic curves are discernable in all plots the magnitude of the correlations never exceeds a value of 0.06.

Table 4.1 summarizes the results of the correlation analysis for a group of 15 PFMs for which the affinity predictions show high correlation (Pearson $r > 0.3$) with the experimentally observed *R/G* ratios and the resampling p-values are < 1e-4 (that is, no resampling yielded higher correlation than the actual data). Remarkably, the optimal parameters for all factors and conditions yield maximal values of $\langle N \rangle$ (over all 6725 intergenic regions) in the range of 0.5 to 5, that is, each factor recognizes at least one intergenic region with high probability. This is biologically reasonable assuming that each transcription factor should strongly bind some promoter region, in at least one condition while at the same time the factor should not cover nearly all sites in the sequence. $\langle N \rangle_{max}$ falls outside of the meaningful range only in two cases. In the case of Hap1 the ''optimal'' R_0 is small and thus poorly defined in the sense explained in context of equation (4.21). Conversely, for Rap1 several sequences have large clusters of neighbouring Rap1 binding sites (Gilson et al., 1993) which yields a $\langle N \rangle_{max}$ of ≈ 12.5. In general only a limited area of the parameter space allows $\langle N \rangle_{max}$ to lie within the biologically meaningful bounds. This area is indicated by the white staggered lines in Figure 4.5 for the transcription factor GAL4 while a generic picture of how $\langle N \rangle_{max}$ changes across the parameter space is shown in Figure 4.10 for the factor CBF1. The fact that the optimal parameters yield meaningful values for $\langle N \rangle_{max}$ thus strongly underlines the validity of the found correlations and the location of the optimal parameters.

Parameter Choice in the Absence of Experimental Data

While it is possible to determine the optimal coefficients R_0 and λ in the presence of sufficient binding data, it is clearly desirable to have some prescription, which would allow the parameter determination on general grounds also for factors for which no binding data is available.

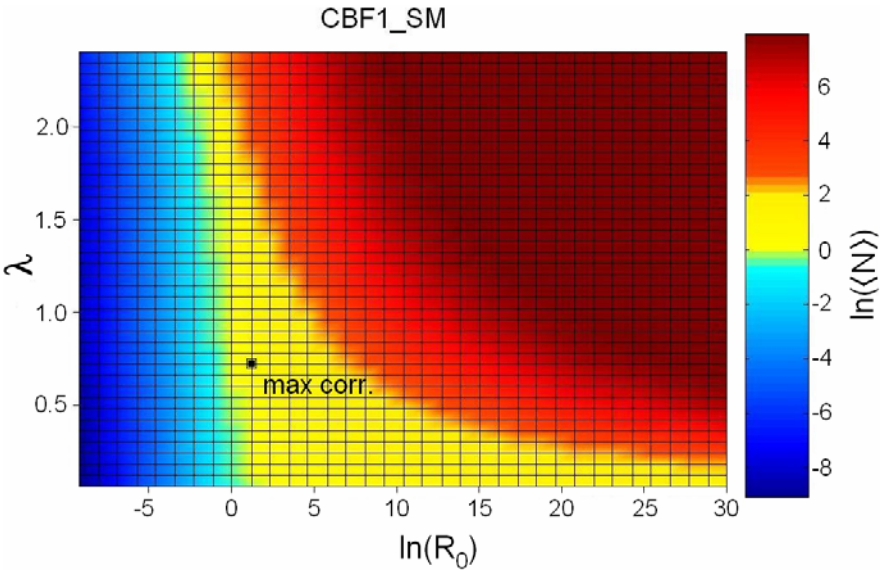
By investigating the parameter space for many different TFs it became apparent that the location of the hyperbolic lines of high correlation across the parameter space is dependent on the motif length *M* of the respective TF. The hyperbolic lines thereby appear shifted towards larger values of R_0 if *M* is large. This is indicated in Figure 4.11 (size of white arrows) for four factors with binding motifs of varying length. Astonishingly, when setting λ to a fixed value for all TFs then there exists a linear correlation between the optimal values of $\ln(R_0)$ and the motif length of the respective TF. In addition, it is evident that the optimal correlation coefficients are rather insensitive to small changes in the parameters, particularly when following the hyperbolic lines of high correlation (notice the width of the hyperbolas representing high correlations in Figures 4.5b and 4.11). Based on these observations and on the optimal values for λ shown in Table 4.1 I decided to fix λ to an average value of 0.7 for all transcription factors and all conditions. This fixation reduces the parameter space to only

Table 4.1 – Summary of the correlation analysis for 15 TFs

Matrix	Condition	M	λ	$\ln(R_0)$	$\langle N \rangle_{max}$	r	r_{pred}
ABF1_01	YPD	22	0.60	8.11	2.95	0.567	0.563
	in vitro	22	0.65	6.91	2.52	0.553	0.545
ABF_C	YPD	15	0.45	4.61	3.08	0.586	0.562
	in vitro	15	0.50	3.51	2.37	0.569	0.543
CBF1_B	YPD	10	0.75	0.00	1.23	0.427	0.427
	SM	10	0.45	3.51	2.90	0.684	0.674
GAL4_01	YPD	23	0.40	13.82	2.99	0.559	0.557
	galactose	23	0.25	25.33	3.00	0.336	0.326
	raffinose	23	1.45	2.30	1.67	0.605	0.590
GAL4_C	YPD	22	0.65	8.11	3.33	0.573	0.572
	galactose	22	0.25	26.53	4.13	0.340	0.315
	raffinose	22	1.30	3.51	2.53	0.624	0.601
GCN4_01	SM	27	0.50	15.02	2.31	0.341	0.150
	rapamycin	27	0.60	15.02	2.31	0.312	0.142
GCN4_C	SM	10	0.50	0.00	1.35	0.352	0.321
	rapamycin	10	0.50	0.00	1.35	0.351	0.313
HAP1_B	YPD	14	0.75	-9.21	0.004	0.350	0.319
HSF_04	H2O2Hi	15	0.90	4.61	2.77	0.488	0.417
	H2O2Lo	15	0.80	4.61	2.66	0.480	0.438
LEU3_B	SM	14	1.20	0.00	0.68	0.335	0.310
MCM1_02	YPD	27	1.70	3.51	0.93	0.316	0.309
	α factor	27	1.45	4.61	0.97	0.368	0.356
MIG1_01	*in vitro*	17	0.90	2.30	1.23	0.596	0.591
RAP1_C	YPD	14	0.60	6.91	12.51	0.382	0.337
	SM	14	0.35	12.72	12.99	0.440	0.370
	in vitro	14	0.15	13.82	5.23	0.445	0.432
RCS1_Q2	H2O2Lo	13	0.05	21.93	1.01	0.388	0.321
REB1_B	YPD	9	0.45	1.20	1.25	0.515	0.506
	H2O2Hi	9	0.55	2.30	2.26	0.325	0.312
	H2O2Lo	9	0.50	1.20	1.37	0.597	0.596
MBP_F	YPD	7	0.50	3.51	5.50	0.314	0.278
	H2O2Hi	7	0.50	4.61	7.11	0.579	0.478
	H2O2Lo	7	0.30	5.81	5.80	0.260	0.235
FHL_F	YPD	11	0.65	2.30	2.40	0.481	0.453
	SM	11	0.85	1.20	1.80	0.434	0.413
	rapamycin	11	0.80	1.20	1.70	0.466	0.449
	H2O2Hi	11	0.75	2.30	2.67	0.368	0.331

The first column denotes the matrix identifier (TRANSFAC version 11.1) of those TFs, for which the affinity predictions yield highly significant correlation ($r > 0.3$, p-value << 1e-6) with the genome-wide *R/G* binding ratios from ChIP-chip in at least on of the tested conditions (second column). The motif length (M) is given in the third column followed by the parameter settings for λ and R_0 that result in the maximal correlation coefficient r and some maximal value $\langle N \rangle_{max}$ over all intergenic regions. The last column denotes the correlation coefficient that is achieved by setting $\lambda = 0.7$ and deriving R_0 from the regression analysis of Figure 4.12. It is apparent that in all but one case the differences between the correlation obtained from optimal and predicted parameters are small. The only exception is GNC4_01 whose motif has been arbitrarily extended by many unspecific positions. Stars indicate PFMs obtained from Frankel et al., 2004. These matrices were not included for the derivation of the generic parameter prescription yielding r_{pred} in the last column.

Figure 4.10 – Changes of $\langle N \rangle_{max}$ across the parameter space



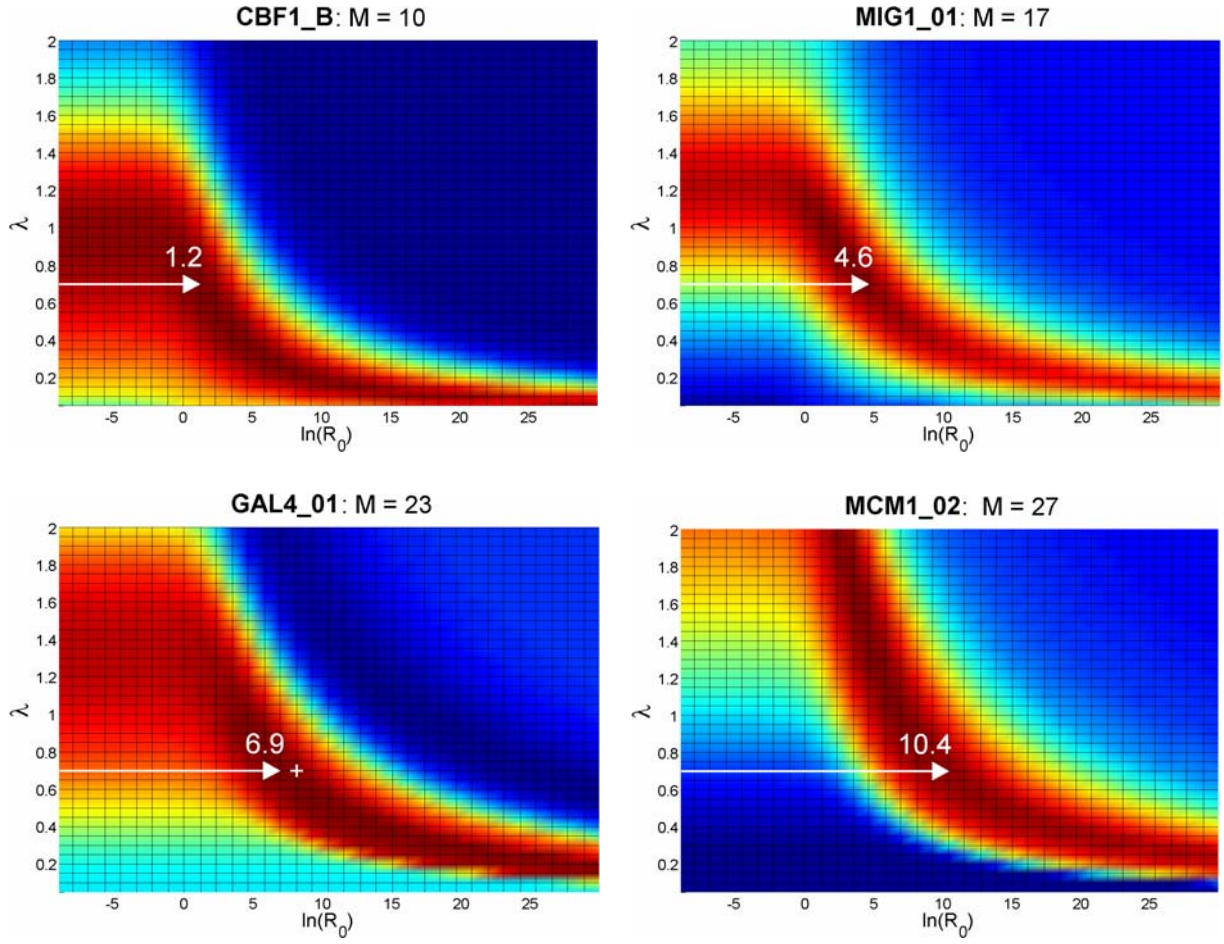
How $\ln(\langle N \rangle_{max})$ changes across the parameter space for CBF1 is in indicated by colour shades. The yellow area thereby highlights the part of the parameter space in which $\langle N \rangle_{max}$ stays within biologically meaningful bounds of 0.5 to 5.0. These bounds are based on the assumption that a given TF will bind at least some sequence with probability > 0.50 in some cellular condition in which the factor is actually expressed. As indicated for CBF1 the optimal correlation between model predictions and experimental ChIP-chip data lies in fact within this area. Assuming $\ln(R_0) < 0$ causes a $\langle N \rangle_{max}$ to quickly decline to unrealistically small values (given that CBF1 is expressed in the shown condition). In contrast, assuming a large value for λ and R_0 causes all sites of a given intergenic region to be occupied. It follows that $\langle N \rangle_{max}$ will take on the value of the longest intergenic region (disregarding the preclusion effects of neighbouring TFs).

R_0, which can now be described as a function of M. Figure 4.12 shows the regression analysis of $\ln(R_0)$ against M for all TRANSFAC matrices with significant correlation between model predictions and experimental data. The resulting regression formula:

$$\ln(R_0) = 6 \cdot M - 6.0 \qquad (4.23)$$

allows to determine R_0 for any motif length. This finding can be understood if one remembers that R_0 determines the binding energy between a TF and its consensus site. This binding energy likely grows with the width of the TF motif through an increasing number of protein-DNA contacts (see Figure 2.5). Thus, since R_0 grows exponentially with the binding energy, $\ln(R_0)$ grows linearly with M (see equation 4.13). On the other hand, R_0 also depends on the

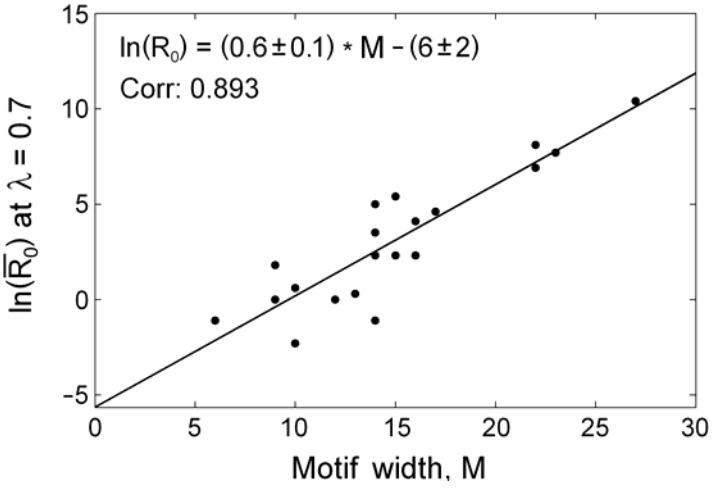
Figure 4.11 – R_0 grows with the length of the TF motif



Comparing the parameter space of different TFs reveals that the optimal ln(R_0) grows in an almost perfect linear fashion with the length M of the TF motif for any given λ (compare length of arrows with M, highlighted for the case $\lambda = 0.7$). The effect of changes in TF concentration is illustrated for GAL4. The optimal ln(R_0) changes for this factor from 6.9 to 8.1 (indicated by a white cross) when analysing ChIP-chip data from cells grown in galactose instead of glucose containing medium. This is in accordance with experiments which showed GAL4 to be 5 times higher expressed under this condition.

concentration of free TF molecules. Its value should thus vary between different cellular conditions. Indeed such changes are observed for instance in case of GAL4, which is known to be fivefold overexpressed in galactose compared to glucose containing medium. Accordingly, its optimal ln(R_0) shifts form 6.9 to 8.1 between the corresponding ChIP-chip data set (indicated by a white cross in Figure 4.11). However, this shift amounts to such a small change in R_0 that good correlation with the experimental data is obtained when deriving ln(R_0) based on the regression formula from Figure 4.12. Over all data sets the changes in ln(R_0) due to varying conditions are found to be much smaller (±2 around the average) than

Figure 4.12 – Deriving a general prescription for setting R_0



For each matrix the optimal value of $\ln(R_0)$ for a fixed value of $\lambda = 0.7$ is plotted. For factors that were tested in various cell culture conditions the average of the optimal $\ln(R_0)$ for the corresponding matrix is plotted. Deviations from this value, due to condition-dependent (TF-concentration-dependent) variation, are generally small (maximally $\ln(R_0) \pm 2$). The p-value of the resulting regression is $1.2 \cdot 10^{-7}$. The errors in the formula denote the 95% confidence interval on the regression parameters.

the overall dependence on the motif length of the respective TF. This robustness against changes in TF concentration is explained by R_0 depending only linearly on the concentration but exponentially on the optimal binding energy. In turn, the motif length is found to strongly dominate the behaviour of R_0.

It should be noted that instead of setting λ to a fixed value one could instead fix R_0 in which case λ would grow linearly with M. While λ might indeed vary between factors, it is difficult to find a meaningful biophysical explanation as to why interrupting a certain amino acid - base pair interaction would cause a much higher mismatch energy in a TF with long binding motif as compared to one with a short motif. For a more mathematical explanation consider the following scenario where all mismatch energies $\varepsilon_{j,\alpha}$ have the same value ε' . In this case the number of ways, W, in which a binding site for a given factor can be realized without exceeding a critical mismatch energy, E_C, can be analytically derived. With a maximal number of K deviations from the consensus ($\rightarrow E_C = K\,\varepsilon'$) one obtains (von Hippel et al., 1986) as a first order approximation for W:

$$W(M,E_C) = \sum_{k=0}^{K} \binom{M}{k} 3^k \quad \approx \quad \binom{M}{K} 3^K \quad \approx \quad \frac{(3M/K-3)^K}{(1-K/M)^M \sqrt{2\pi K(1-K/M)}}. \tag{4.24}$$

Applying equation (3.11) for deriving λ yields:

$$\lambda = \frac{d\ln(W)}{dE} \approx \ln\left(3M\varepsilon'/E_C - 3\right)/\varepsilon' . \qquad (4.25)$$

If E_C is considered a constant then λ grows logarithmically with M. This behaviour is expected to carry over also to other more general situations where the different $\varepsilon_{j,\alpha}$'s may have individual values. A linear correlation between λ and M would thus be difficult to explain. Moreover, if we assume that the binding energy, and with it the maximal mismatch energy E_C, increases with the motif length then $E_C \propto M$ and thus, λ should be largely independent of M. From these considerations, it appears biologically meaningful to assume R_0 instead of λ to be dependent on M. For purely practical purposes it is however irrelevant which of the two parameters is considered to be invariant.

Computing $\ln(R_0)$ for different TFs based on the regression formula shown in Figure 4.12 completes the derivation of the TRAP method and provides the basis for the subsequent analyses. As shown in the last column of Table 4.1 this approach yields indeed correlations with the *R/G*-ratios that are almost as high as the optimal correlations. Importantly, using this prescription for setting R_0 yields values for $\langle N \rangle_{max}$ that always lie in the meaningful range of 0.5 to 5.0. In later sections this way of obtaining R_0 and λ will therefore be used to predict relative binding affinities for transcription factors with known motifs for which no genome-wide binding data is available that could be used to obtain the optimal parameters. In this context it has to be stressed that the above prescription for determining the parameters yields as good results for the matrices directly derived by Harbison et al., (2004) as it does for the TRANSFAC matrices. This is important as the Harbison matrices were not included in deriving the regression model for obtaining $\ln(R_0)$, equation (4.23), and thus serve as a control for its general validity.

Choosing the proper motif length

As mention in context of Figure 2.5 and 3.2 matrices may contain unspecific positions which then define an arbitrary consensus site. In the above situation such a consensus obtains a spuriously low binding energy through the dependence of R_0 on M as given by equation (4.23). The resulting overestimation of the optimal R_0 in turn causes the prediction of arbitrarily high binding affinities. For instance, for the factor GCN4 there exist two matrices in TRANSFAC namely, GCN4_01 with M = 27 bp and GCN4_C with M = 10 bp (sequence LOGOs shown in Figure 4.13). For identical λ, GCN4_01 therefore results in a vastly larger estimate of R_0 as compared to GCN4_C. That the estimate for R_0 based on GCN4_01 is

Figure 4.13 – Sequence LOGO of GCN4_01 and GCN4_C

a)

b)

The TRANSFAC motif GCN4_01 a) is derived from a SELEX experiment and shares the informative core with the matrix GCN4_C b). From the sequence LOGO in a) it is apparent that the PFM has been arbitrarily extended beyond the actual binding motif. The uninformative tails of the PFM can be truncated by applying a cutoff to the information content of the matrix positions. (Sequence LOGOs were obtained from TRANSFAC online.)

indeed inaccurate can be seen in Table 4.1, where GCN4_01 is the only matrix with a large difference in the correlation obtained from the optimal and estimated model parameters (r = 0.34 vs. 0.15). The problem of overestimating R_0 due to arbitrarily added bases in a matrix can be avoided by restricting the motif to positions with higher information content. Following equation (3.20) the information content of a single position i in the PFM can be computed by the Kullback-Leibler entropy (Kullback et al., 1951) difference between the PFM and background base frequencies:

$$I_i = \sum_{\alpha=A,C,G,T} v_{i,\alpha} \log_2\left(\frac{v_{i,\alpha}}{b_\alpha}\right) \tag{4.26}$$

where I_i can range between 0 and 2 bits. Using an entropy cutoff of 0.1 bits for bases near the motif core and a cutoff of 0.2 for distant flanking bases reduces the motif length of GCN4_01 to 11 bases and in turn greatly improves the obtained correlation coefficients between predictions and ChIP-chip data (r = 0.56 vs. r = 0.15). The results could likely be further improved by first applying a sophisticated method for adding pseudo counts (Rahmann et al., 2003) as outlined on page 34. It should be noted that the regression line in Figure 4.12 is not sensitive to such influences.

4.2 Results

4.2.1 Comparison of TRAP with Hit-Based Methods

As outlined in section 3.1.3 traditionally, computational target predictions have focused on the identification of individual binding sites for a given TF. This is usually done by scanning a position weight matrix along the sequence and assigning a TF hit, whenever the log likelihood score exceeds the pre-defined threshold (Wasserman et al., 2004, Rahmann et al., 2003). Such traditional methods may suffer from the arbitrariness of the score threshold and perhaps also from the subsequent discretization of the binding site scores. In the following I will investigate how well hit based approaches can account for the experimental binding values as compared to TRAP (using predefined parameters according to equation 4.23). This analysis will show how much information about the relative binding strength of a TF to a sequence is contained in the continuous affinity predictions made by TRAP as compared to the binary values assigned by the hit based approaches.

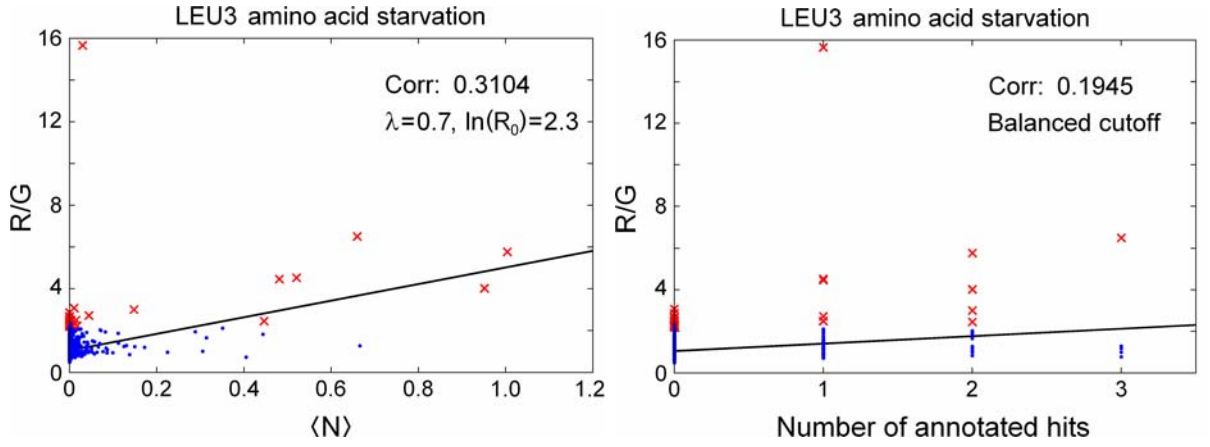
Comparison of achieved correlations

For the comparison I consider two commonly used hit-based methods (see Section 3.1.3 for details). The first approach is referred to as “balanced cutoff” and constitutes a state of the art method that invokes a likelihood score threshold, which is chosen in such a way that the expected number of false positive hits is balanced by the expected number of false negatives (Rahmann et al., 2003). For each sequence this method predicts a discrete number of hits, which can be compared to experimental binding ratios and the TRAP predictions for $\langle N \rangle$. The difference between the predictions made by TRAP and the balanced method is illustrated for the factor Leu3 in Figure 4.14a. The TRAP approach thereby leads to improved correlation with experimental data.

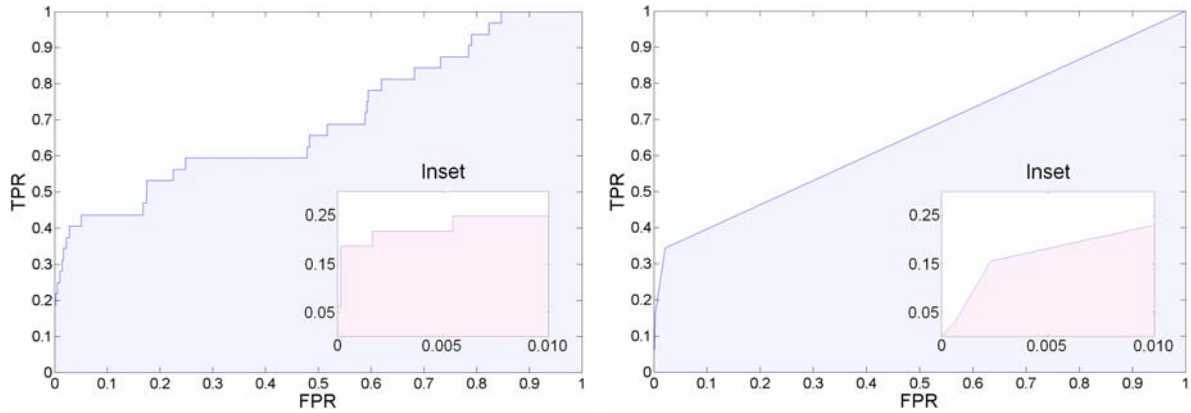
For a second comparison, I applied a different threshold prescription, referred to as 5-FP, in which the expected false-positive rate is arbitrarily set to 5%. Table 4.2 provides a complete comparison of the results obtained from all three methods for those experimentally tested TF-condition pairs that yielded a Pearson correlation $r > 0.3$ for at least on of the methods. It can be seen that in ≈ 75% of cases TRAP results in better correlations with experimental binding ratios than the hit-based methods. This percentage is increased to 90% if uninformative matrix positions are removed from the matrices before computing affinities and deriving $\ln(R_0)$.

Figure 4.14 – Comparison of TRAP to a standard hit based method

a)



b)



a) As an example, the results for Leu3 (in amino acid starved condition) from TRAP (left panel) are compared with the results obtained from the balanced cutoff method (right panel). Sequences with significant *R/G* ratios (ChIP-chip p-value < 0.001) are shown as red stars. It is apparent that TRAP improves the correlation with *R/G* ratios and in particular also the ranking of the significant ChIP-chip targets. b) The improved ranking of LEU3 targets is signified by the larger area under the ROC curve (AUC = 0.70) obtained by TRAP as compared to the balanced cutoff method (AUC = 0.66). Importantly, as the inset shows, especially the start of the ROC curve, which refers to the sequences with highest predicted affinity and thus to the most relevant range for experimentalists, is improved.

Comparison of target gene rankings

In addition to improving the correlation with experimental data, TRAP also improves the ranking of the sequences considered to be bound in the ChIP-chip or PBM experiments. As shown in Figure 4.14a, out of the top seven ranking sequences according to affinity six were indeed bound by LEU3 in the chip experiment. In contrast, only one out of the five sequences

with a maximum of three annotated hits according to the balanced cutoff method was bound by the TF according to experimentalists (ChIP-chip p-value < 10^{-3}).

Given a set of "true" target sequences (ChIP-chip p-value < 10^{-3}) the quality of a particular ranking scheme can be evaluated by means of a ROC curve analysis. To this end one passes sequentially through the list of all intergenic regions, ranked according to the affinity measure, starting with highest predicted affinity. Then to generate the ROC curve, beginning at the origin of the plot, each time a true target is encountered in the list one moves a step upwards and if a non-target is encountered one move a step to the right. The step size is thereby chosen in such a way that all vertical and horizontal moves sum to 1 separately. The quality of the ranking is then measured by the area under the generated curve (AUC) which can range from 0 to 1. An AUC of 1 thereby corresponds to a perfect ranking, i.e. 100% sensitivity (all true positives are found on top) and 100% specificity (no false positives is ranked above a true positive), while an AUC of 0.5 (corresponding to a diagonal line across the plot) indicates an arbitrary ranking that does not provide discrimination between experimentally bound and unbound sequences. The ROC curve corresponding to the ranking of LEU3 targets according to TRAP is shown in the left panel of Figure 4.14b.

Table 4.2 shows the comprehensive results of the ROC-curve analysis. Most AUCs are much larger than 0.5, indicating a strong predictive power of TRAP over the experimental binding data. To compare again with the hit-based methods I took the number of hits as measure to rank the intergenic regions and computed the ROC curve area based on this ranking. The result of this analysis for LEU3 is shown in the right panel of Figure 4.14b. As shown in Table 4.2 TRAP again performs consistently better than the hit-based approaches. On the entire set of 29 TRANSFAC matrices TRAP yields a ROC curve area of ≥ 0.7 for 22 matrices in at least one of the experimentally tested conditions as opposed to only 16 and 14 matrices for the balanced and 5-FP cutoff methods, respectively.

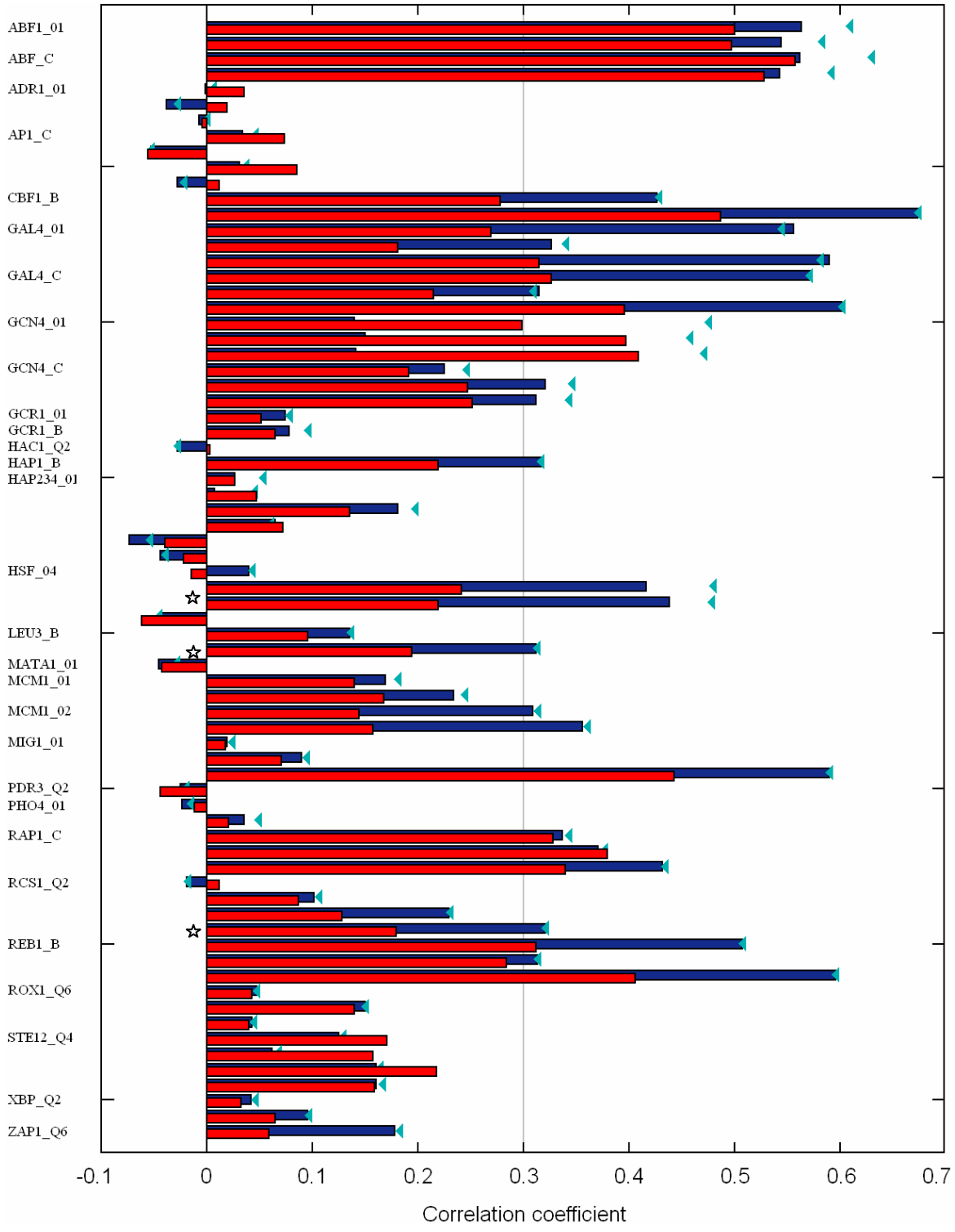
The correlation and ROC AUC results for all experimentally tested pairs of TF and cell culture conditions are shown in Figure 4.15 and 4.16 for TRAP using unmodified or modified matrices (according to a simplified entropy cutoff of 0.1 bits for all positions in the matrix) and for the balanced cutoff method. For the entire list of 25 factors and 13 conditions (61 experimentally tested combinations), the TRAP predictions result in highly significant correlations (r > 0.3) for 23 of these combinations. In addition, for 36 combinations TRAP yielded a ROC curve area ≥ 0.7 including cases such as PHO4, which yielded low correlation likely due only to high noise in the experimental data. Removing uninformative positions from the matrices resulted in better correlation and larger ROC curve areas in almost all cases. Particularly the matrices with many arbitrarily included positions (GCN4_01 and to a lesser extent ABF1_01) profited from applying the entropy cutoff.

Table 4.2 – Comparison of TRAP to the balanced cutoff and 5-FP methods

		Correlation Coeff.			ROC-curve AUC		
Matrix	Condition	TRAP	5-FP	Bal	TRAP	5-FP	Bal
ABF1_01	YPD	0.563	0.511	0.501	0.924	0.868	0.871
	in vitro	0.545	0.506	0.497	0.894	0.848	0.851
ABF_C	YPD	0.562	0.580	0.558	0.932	0.921	0.920
	in vitro	0.543	0.544	0.528	0.896	0.854	0.869
CBF1_B	YPD	0.427	0.303	0.278	0.994	0.978	0.975
	SM	0.674	0.524	0.487	0.886	0.830	0.833
GAL4_01	YPD	0.557	0.287	0.270	0.678	0.634	0.632
	galactose	0.326	0.191	0.180	0.584	0.641	0.639
	raffinose	0.590	0.332	0.315	0.716	0.657	0.655
GAL4_C	YPD	0.572	0.327	0.327	0.677	0.636	0.636
	galactose	0.315	0.214	0.214	0.576	0.661	0.661
	raffinose	0.601	0.395	0.395	0.726	0.677	0.677
GCN4_01	SM	0.150	0.381	0.397	0.801	0.750	0.691
	rapamycin	0.142	0.391	0.408	0.807	0.802	0.730
GCN4_C	SM	0.321	0.209	0.248	0.771	0.649	0.720
	rapamycin	0.313	0.214	0.251	0.784	0.662	0.764
HAP1_B	YPD	0.319	0.251	0.219	0.808	0.656	0.687
HSF_04	H2O2hi	0.417	0.260	0.242	0.753	0.656	0.662
	H2O2lo	0.438	0.232	0.219	0.789	0.695	0.701
LEU3_B	SM	0.310	0.209	0.195	0.698	0.649	0.662
MCM1_02	YPD	0.309	0.109	0.144	0.807	0.716	0.634
	α Factor	0.356	0.100	0.157	0.861	0.771	0.701
MIG1_01	*in vitro*	0.591	0.463	0.443	0.879	0.698	0.704
RAP1_C	YPD	0.337	0.351	0.328	0.909	0.781	0.777
	SM	0.370	0.402	0.380	N/A	N/A	N/A
	in vitro	0.432	0.365	0.340	0.886	0.698	0.716
RCS1_Q2	H2O2lo	0.321	0.158	0.179	0.547	0.499	0.504
REB1_B	YPD	0.506	0.420	0.312	0.929	0.897	0.875
	H2O2hi	0.312	0.339	0.284	0.844	0.853	0.819
	H2O2lo	0.596	0.516	0.406	N/A	N/A	N/A

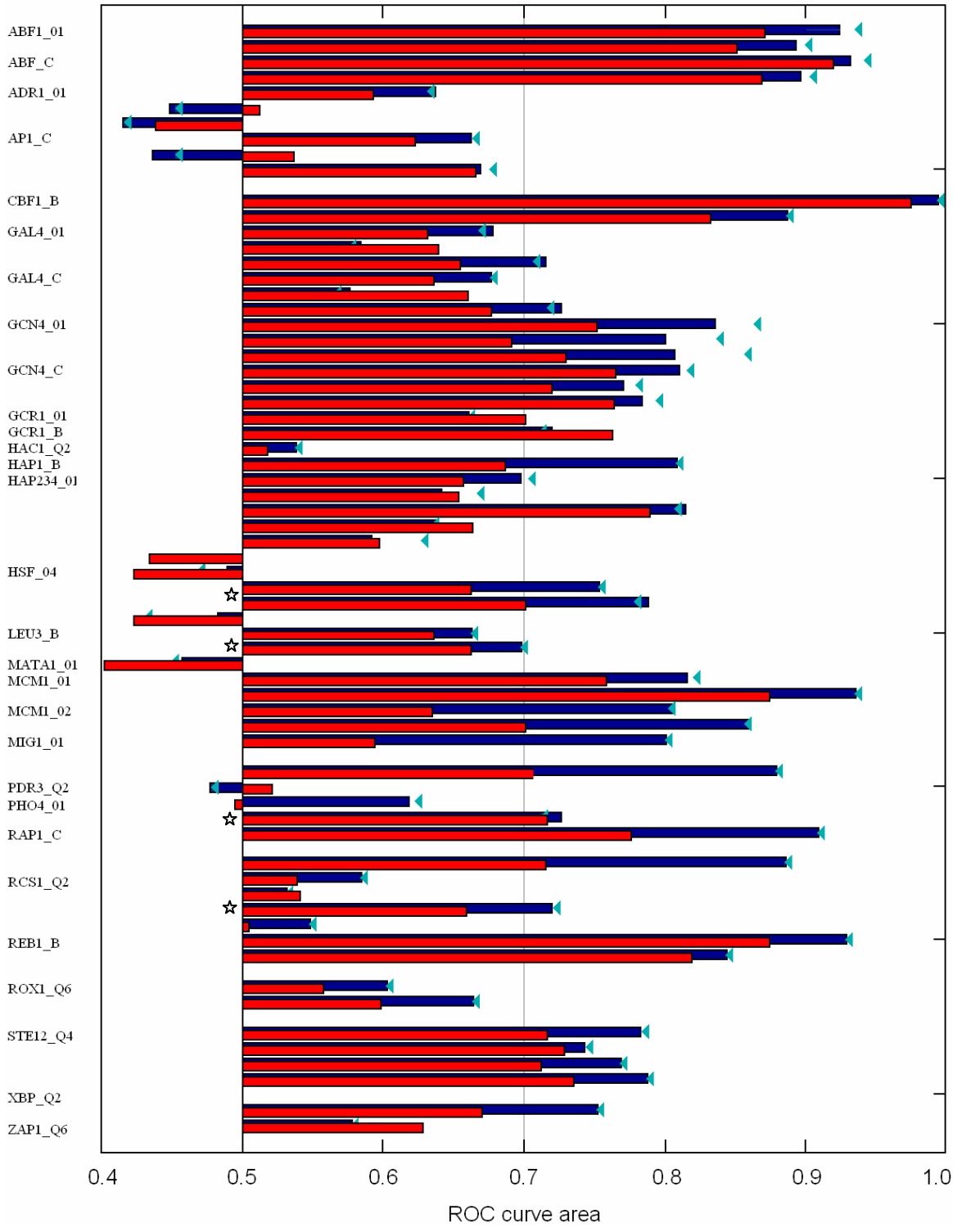
Results from the correlation and ROC curve analysis. N/A denotes those cases for which the transcription factor does not have any true associations in the specified condition according to experimenters. Even when using the unmodified matrices TRAP yields the highest correlation in 23 out of 30 TF-condition pairs as well as the largest ROC curve area in 25 out of 28 cases. Between the two hit based methods the balanced cutoff approach shows slight advantages in the ROC curve analysis while the 5-FP methods tends to yield somewhat higher correlations. However, overall the differences are small between the two hit based methods.

Figure 4.15 – Correlations obtained for all TF-condition pairs



Each bar represents the correlation obtained between a given experimental data set and TRAP (Blue bars) or the balanced cutoff method (red bars). TRAP improves the correlation in virtually all cases with a significant correlation of $r > 0.3$. Light blue triangles point to the correlation obtained when restricting the motifs to positions with information content > 0.1 bits indicating in general further improved correlations. Stars refer to TFs primarily active in the indicated stress condition.

Figure 4.16 – Comparison of ROC curve areas



Each bar represents the ROC AUC obtained for a given data set by TRAP (blue) or the balanced cutoff method (red). TRAP yields larger ROC AUCs for nearly all TF-conditions pairs. Results are usually further improved if only informative motif positions are considered for the affinity computation (triangles). Blank lines correspond to cases where no TF targets were found in the chip experiments (all p-values > 0.001). Stars refer to TFs which are active primarily in the indicated stress condition.

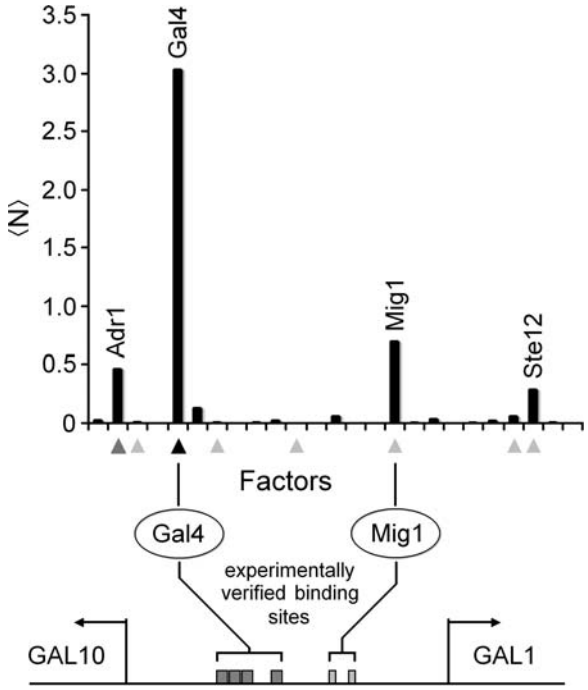
4.2.2 Predicted affinities are comparable between factors

The above analysis shows that TRAP successfully ranks sequences for a given transcription factor. Here I address the complementary question: given a certain intergenic region, can TRAP successfully predict which TFs bind to the sequence? In general, factors bound to a given intergenic region in the ChIP experiments should have higher predicted values of $\langle N \rangle$ than unbound factors. Since experimental *R/G* ratios for different factors are not directly comparable, I follow again the binding prescription as given in (Harbison et al., 2004, Mukherjee et al., 2004) and distinguish binding TFs from non-binders according to the p-value threshold of 10^{-3}.

Figure 4.17 shows as an example the intergenic region between the genes GAL1 and GAL10 with its experimentally verified high affinity sites for Gal4 and Mig1 (Selleck et al., 1987, Frolova et al., 1999). This region is also significantly enriched in the ChIP-chip experiment of Gal4 and in the PBM experiment of Mig1. In contrast, none of the other 23 factors were bound to the region according to ChIP chip or PBM data. In accordance with these experimental findings, TRAP predicts the highest affinities for Gal4 followed by Mig1, Adr1 and Ste12. All other factors have only negligible predicted affinities in good agreement with ChIP-chip data. Interestingly, independent chromatin precipitation experiments have shown that Ste12 has also weak but measurable affinity to the GAL1 - GAL10 intergenic region (Reeves et al., 2005). The balanced cutoff method also predicts these 4 factors as potential binders but in addition 4 others (Ap1, Gcr1, Hsf1 and Rox1). If one ranks traditional annotations according to the number of hits, then Gal4 is ranked highest with seven annotated hits followed by Adr1 with two while Mig1 and Ste12 with one binding site each are assigned a tied rank with Ap1, Gcr1, Hsf1 and Rox1.

This analysis was carried out on the entire set of 4451 intergenic sequences which have a ChIP-chip p-value assigned for all the 25 TFs for which TRANSFAC contains a matrix. In total this set yields 2388 (~2% of all 25×4451 interactions) TF-DNA interactions with chip p-value $< 10^{-3}$ (true positives). For each intergenic region the TFs are ranked according to predicted affinity in case of TRAP or the number of annotated binding site hits in case of the balanced and 5-FP methods. Subsequently, all cases are counted in which a true positive is ranked above all true negatives. For the case of the GAL1 – GAL10 intergenic region this yields a count of two as both Gal4 and Mig1 are true positives and are ranked above all other TFs. For the standard methods, an ambiguous ranking can arise if two factors have the same number of hits annotated. In cases where the ambiguous rank involves true

Figure 4.17 – Affinities for the upstream region of GAL1 and GAL10



The histogram shows the affinity scores as predicted by TRAP. Triangles indicate the factors that have hits annotated according to the balanced cutoff method (black: 7 binding sites, dark grey: 2 binding sites, light grey: 1 binding site). The lower panel indicates the experimentally verified binding sites are indicated (Selleck et al.,1987, Frolova et al., 1999).

positives and a true negatives I count $\frac{\text{true positives}}{\text{true positives} + \text{true negatives}}$, which corresponds to the probability of picking a true positive among the TFs with same number of hits. The analysis shows that 643 (27%) of the significant interactions are correctly ranked on top according to TRAP as compared to 343 (14%) in case of the balanced cutoff and 551 (23%) in case of the 5-FP method. These findings show that in a considerable number of cases the ranking of TFs according to TRAP is in accordance with ChIP-chip data and again overall better than what is obtained by traditional hit-based methods. Due to using unmodified matrices (for the purpose of comparing with the hit based methods) especially GCN4_01 yielded artificially high affinities for many regions. Refining matrices by retaining only positions with higher information content would thus further improve the obtained results. It is important to note that using binding site scores obtained from the log likelihood method or the simplified Boltzmann model yield meaningless results in the above setting, as the scores obtained by these methods for different PFMs differ by many orders of magnitude.

These results together with the findings from Section 4.2.1, which showed that TRAP improves the correlation with the data as well as the ranking of sequence, suggest that the affinity scores contain information that is lost by discretizing the individual binding site scores. One possibility is hereby that it is important to retain the relative binding strength between the higher affinity sites (the sites that are predicted as hits by the balanced or 5-FP methods). The other possibility is that integrating over low affinity sites (which are considered non-hits by the classical methods) may yield significant overall affinities that contribute strongly to the correlation coefficients. Which of these scenarios plays a larger role for the improvement of binding predictions is addressed below.

4.2.3 Contributions from Low Affinity Sites to ⟨N⟩

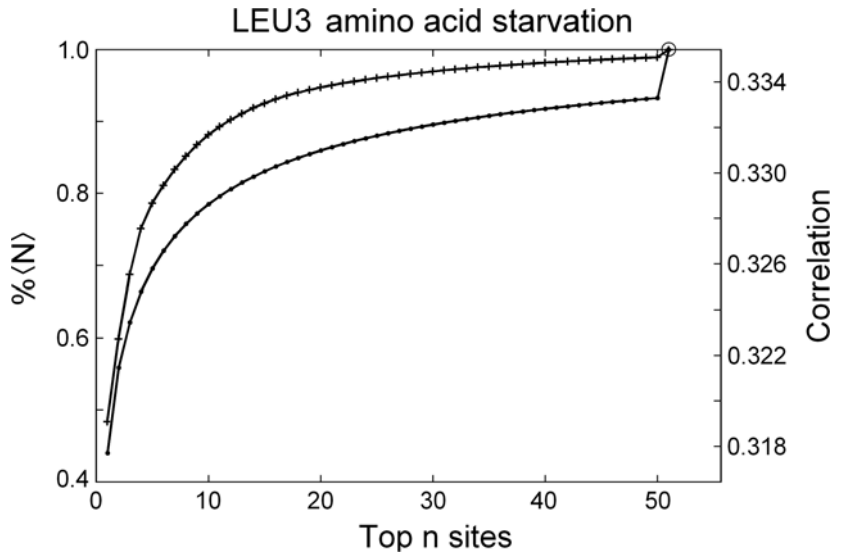
While TRAP predicts the overall affinity of a transcription factor to a sequence region, it is interesting to ask which sites add most significantly to ⟨N⟩ and to what extent strong and weak sites contribute. To address this question I studied in more detail the relative contribution of different sites to the total expected count, ⟨N⟩, and therefore to the correlation of ⟨N⟩ with the observed binding ratios. To this end, first all sites in a given sequence are ranked according to their probability of being bound, *p*, as given by equation (4.11). Then the expected number of TFs bound to the sequence is approximated by the sum of its *n* top-ranking sites,

$$\langle N \rangle_n = \sum_{i=1}^{n} p_i \ . \tag{4.26}$$

The resulting quantities $\langle N \rangle_n$ for all intergenic regions and a given value of *n* are then correlated with the experimental binding values. The result of this analysis is illustrated in Figure 4.18 for Leu3. Averaged over all 6725 intergenic regions one sees that the strongest binding sites contribute about 50% to the total affinities and the second strongest sites about 10%. All other sites together add up to only about one third of the affinities. These results apply also to other factors where the strongest bound sites sometimes contribute over 90% of the total affinity. This is in line with what was observed in Figure 4.7 for ABF1 where the total affinity of many intergenic regions with apparent plateaus in ⟨N⟩ was determined primarily by a few sites with higher affinity. Nevertheless, for the majority of matrices a better correlation is obtained when all sites are taken into account rather than a single strongest site. This suggests that the relative binding affinities for a given intergenic region are well modelled by taking the total sum over all sites in the region, and supports the claim that a mechanistic description of binding data is possible without imposing any threshold. Table 4.3 provides the results of this analysis for all factors and conditions for which a significant

correlation has been obtained. The arbitrary focus on only a few binding sites ($n = n^*$) has little effect on the overall correlation with the data and in many of cases performs worse. On the other hand the analysis suggests that the measured Pearson correlations are primarily determined by the higher affinity sites in each sequence.

Figure 4.18 – Contribution of sites with lower affinity



When ⟨N⟩ is arbitrarily constrained to only the top n scoring sites then the expected number of bound TFs is reduced, which in turn affects the correlation with the experimental R/G ratio. The upper line shows the changes in the correlation coefficient with varying n, the lower line the changes in ⟨N⟩. The right-most circled dots denote the values when all sites are taken into account. The increase in the correlation coefficient suggests that the inclusion is biologically meaningful until the correlation coefficient saturates (optimal $r = 0.335$) as more and more sites with vanishing affinity are taken into account. This demonstrates that integrating the contributions from all sites provides a more robust approach than limiting the annotation to a few best sites determined by some arbitrary cutoff.

An important assumption in the model by Berg and von Hippel is that different base-pairs contribute independently from each other to the overall binding energy. This assumption also entails that mismatch energies for large deviations from the consensus sequence are not calculated differently from small deviations, which likely does not reflect the behaviour of real TFs. This is because TF-DNA complexes can, presumably, compensate for mismatches through, for instance, conformational changes. Such conformational changes would also likely go in hand with changes in the mismatch energy levels $\varepsilon_{i,\alpha}$ that are assumed to be

Table 4.3 – Contributions of top-ranking sites

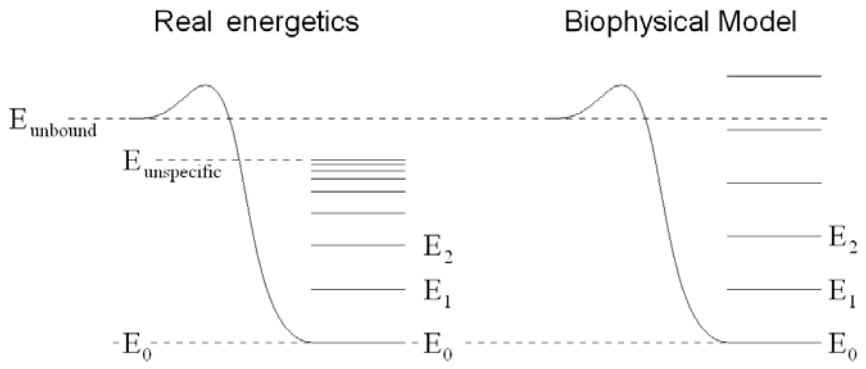
Matrix	Condition	$r(n=1)$	$r(n=\text{ALL})$	$r(n=n^*)$	n^*
ABF1_01	YPD	0.592	0.586	0.593	2
	in vitro	0.562	0.558	0.562	1
ABF_C	YPD	0.629	0.637	0.643	2
CBF1_B	YPD	0.396	0.423	0.425	5
	SM	0.600	0.680	0.681	7
GAL4_01	YPD	0.514	0.581	0.581	ALL
	galactose	0.325	0.338	0.345	2
	raffinose	0.604	0.614	0.614	ALL
GAL4_C	YPD	0.509	0.580	0.580	ALL
	galactose	0.316	0.349	0.352	3
	raffinose	0.595	0.615	0.620	3
GCN4_01	YPD	0.320	0.328	0.328	ALL
	SM	0.378	0.393	0.393	ALL
	rapamycin	0.392	0.408	0.408	ALL
GCN4_C	YPD	0.282	0.301	0.301	ALL
	SM	0.360	0.389	0.389	10
	rapamycin	0.370	0.395	0.395	ALL
HAP1_01	YPD	0.334	0.357	0.357	12
HSF_04	H2O2Lo	0.427	0.491	0.491	ALL
	H2O2Hi	0.407	0.483	0.483	ALL
LEU3_B	SM	0.324	0.333	0.333	ALL
MCM1_02	YPD	0.314	0.315	0.315	ALL
	α Factor	0.369	0.370	0.370	7
MIG1_01	*in vitro*	0.583	0.591	0.592	8
RAP1_C	YPD	0.419	0.394	0.422	2
	SM	0.457	0.444	0.476	2
	in vitro	0.553	0.480	0.553	1
RCS1_Q2	H2O2Lo	0.320	0.338	0.338	4
REB1_B	YPD	0.537	0.545	0.545	2
	H2O2Hi	0.356	0.344	0.356	1
	H2O2Lo	0.624	0.624	0.625	2

This table illustrates that in the majority of cases the best correlations can be obtained when all sites are taken into account, rather than just the best site. Choosing an optimal number, n^*, of best binding sites sometimes results in a minor improvement in the correlation at the expense of an additional and arbitrary parameter. This suggests that the relative binding affinities for a given intergenic region are well modeled by taking the total sum over all sites in the region.

invariant in the Berg and von Hippel model. Thus while real TFs can bind to DNA in an unspecific fashion the model may predict mismatch energies that would result in binding probabilities far below the unspecific binding probability (see Figure 4.19 for an illustration). The contribution of low affinity sites to ⟨N⟩ might therefore be larger than suggested by the above analysis. To see whether the correlation with the data can be improved by assuming a maximal mismatch energy for completely unspecific sites I introduced an additional parameter ξ into the TRAP model representing the minimal binding probability for a given TF to DNA. This measure thus adds a term that grows linearly with the length of the intergenic sequences to ⟨N⟩. Surprisingly, neither the correlation with the experimental data nor the ROC AUCs could thereby be improved. This suggests that the unspecific binding probability is indeed very low and not detectable given the noisy ChIP-chip binding data.

Together the results indicate that the improved correlations and ROC AUCs obtained from TRAP as compared to the hit-based methods are mainly due to retaining the relative binding probabilities among the higher affinity sites in each intergenic region and to a lesser extent due to integrating over weak sites, which normally lie below the hit-based cutoffs.

Figure 4.19 – Assumptions of the biophysical model.



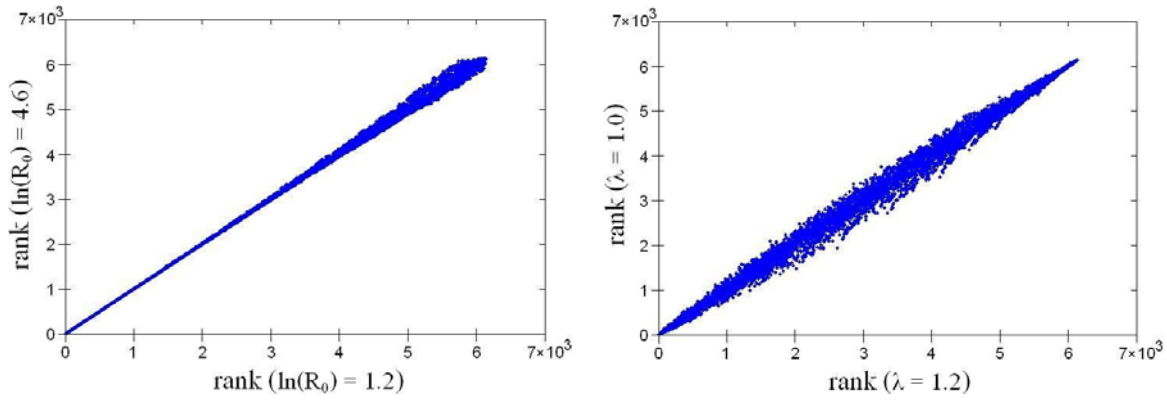
The left panel shows energy levels as they are likely to be found for real TFs in which subsequent mismatches cannot exceed the maximal mismatch energy $E_{unspecific}$. If $E_{unspecifc}$ is reached the factor will simply bind in an unspecific fashion. In addition the real mismatch energy introduced by several deviations from the consensus is likely smaller than the sum over the individual mismatch energies $\varepsilon_{i,\alpha}$ as computed by equation (3.20) since every mismatch will weaken the entire protein – DNA interaction and thus reduce the contributions from the other mismatches. These issues are not reflected in the biophysical model (right panel) which assumes independent mismatch contributions and invariant energy levels for the mismatch energies.

4.2.4 Ranking of Intergenic Regions is Robust

In Section 4.1.3 I showed that the correlations obtained from the biophysical model are relatively insensitive to changes in λ and R_0, which allowed to derive a general prescription for setting the parameters. Given that much of the binding analysis revolves around adequate ranking of TFs for a given sequence or promoters for a given transcription factor I also analyse to what extent the ranking of intergenic regions according to the affinity measure $\langle N \rangle$ is affected by changes in the parameters. Figure 4.20 presents a rank correlation analysis for Leu3 under two different parameter settings representing: a) a large change in $\ln(R_0)$ from 1.2 to 4.6, which corresponds to ≈ 30-fold change in TF-concentration, b) a 20% change in λ. Even for such large changes the ranking of intergenic regions is hardly affected. Correlating the ranks of the intergenic sequences obtained from one setting of λ versus the other results in Spearman rank correlation coefficients of ≈ 0.98. Similarly, the change in R_0 gave a Spearman correlation coefficient of ≈ 0.99.

Figure 4.20 – Robust ranking of affinities

a) **b)**



Comparison of the ranking of intergenic regions according to predicted affinities for Leu3 using different parameter settings. The rank correlation plot for different values of $\ln(R_0)$, is shown in a). The corresponding plot for a 20% change in λ is shown in b).

4.3. Discussion

In order to derive a method that pertains to the gradual binding behaviour of TFs and at the same time allows for the robust prediction of TF targets I have applied a biophysical model that predicts the relative binding affinities of transcription factors to promoter regions in yeast. In contrast to the traditional search for discrete binding sites, I do not impose any threshold,

but integrate the contributions from individual strong sites and weak sites to calculate the expected number of transcription factors bound to a given sequence. In contrast to the ranking provided by hit-based methods, which often varies strongly with the choice of the cutoff, the ranking of sequence fragments according to this affinity measure is robust with respect to sizable variations in the space of two model parameters. Using recent *in vitro* and *in vivo* data from budding yeast, I find that the mismatch scaling parameter λ lies in the range of 0.4 to 1.5 for most factors, in agreement with original predictions made by Berg and von Hippel 1987. The second parameter R_0 is largely determined by the number of informative positions in the binding motif of a given TF, and to a much lesser extent by the transcription factor concentration. This likely reveals a general tendency of TFs with long motif to bind stronger to the DNA than factors with short motifs through an increased number of amino acid base pair interactions. These observations together with the relative insensitivity of the model to small changes in the parameters allowed to provide a simple parameterization of the model with $\lambda = 0.7$ and $R_0 = R_0(M)$ for all factors and conditions. The resulting TRAP method is able to account for a highly significant part of the experimentally measured *R/G* ratios in one or more cellular conditions. Importantly, the parameterization proved to be applicable not only for the TRANSFAC matrices but also for the matrices derived directly from the ChIP-chip data by (Harbison et al., 2004), which had not been included in the derivation of the generic prescription of how to obtain λ and R_0.

The results of the comparison between TRAP and hit-based methods indicate that TRAP can better predict relative binding affinities than any of the hit-based approaches. This improvement is largely due to accounting for differences in the binding strength of sites, which are traditionally only reported as hits and to a lesser extent also due to integrating over weak sites, which likely fall under the threshold of the traditional methods. These findings are complementary to an analysis by Tanay (2006) where it was shown that predicted PWM scores allow to distinguish not only high but also intermediate *R/G* ratios from low ones, indicating that a large range of affinities contribute to the observed binding behaviour of many TFs.

The success of the TRAP approach is not only reflected in better correlation but also in a more accurate and robust ranking of transcription factors. TRAP also performs consistently better than the simple Boltzmann model outlined in Section 3.1.4 with always yielding larger correlations with experimental data. In particular, correlations for PFMs with longer motifs (ABF1_01, GAL4_C, GCN4_01, MCM1_02) were improved by some 100% compared to what is achieved by assuming very low TF concentration and setting $\lambda = 1.0$.

Interestingly, the TRAP predictions also match what is known about the involvement of transcription factors in various cellular conditions. For example, Hsf1, Rcs1 and Leu3 are known to be involved in several aspects of stress response (Raitt et al., 2000, Blaiseau et al., 2001, Zhou et al., 1987) and their predicted affinities show high correlation with *R/G* ratios only in conditions of oxidative stress (H_2O_2) and amino acid starvation (SM), but not in rich medium (marked by a star in Figure 4.15). This also suggests why the physical model cannot be expected to predict binding affinities *in vivo* for certain factors and cellular conditions. For 9 factor-condition pairs with only small correlation ($r < 0.3$) the transcription factors may indeed not be expressed or available for binding under the condition tested. These include Adr1, Hac1, Mata1, Pdr3, Pho4, Xbp1, Yap1, Rox1 and Zap1 in rich medium and Mig1 in medium with galactose as carbon source. For example, Mig1 is known to be located only in the cytoplasm in the presence of galactose, and hence it is not available for DNA-binding in the nucleus (Vit et al., 1997). However, the predictions for Mig1 do show a high correlation ($r = 0.60$) with *in vitro* binding data.

Despite this success the TRAP approach appears to fail for a number of matrices and conditions, even though there is no indication that the corresponding TF is absent from the nucleus in the tested cell condition. In this context it has to be stressed that the affinity approach requires the definition of matrices which can be used as good approximation for mismatch energies in the physical model. There are several cases where one can suspect that the matrix description may be inappropriate. For example, for Hsf1 TRANSFAC lists four matrices, but only one of them (an alternating trimer motif HSF1_04) yields good correlations with the experimental binding ratios. Interestingly, the trimer combination of this matrix has been described as the site with highest affinity for Hsf1 (Sorger et al., 1987, Xiao et al., 1991). Other technical problems may affect the quality of the found correlations. For instance, as mentioned before, ChIP-chip data is very noisy and many large *R/G* ratios are not supported by a significant binding p-value. On the computational side a problem arises when a high affinity site is located just outside of a particular intergenic region. In this case while longer fragments containing the site are able to hybridize to the corresponding spot on the microarray the annotation programs cannot detect the site and thus reports a spuriously low affinity score. Such cases exist for several factors including LEU3 and PHO4. In these cases the correlations can be improved by extending the sequences by some 100 bases beyond the fragments spotted on the microarrays. Finally, it is possible that better predictions can be achieved by improving given PFMs or deriving matrices directly from the ChIP-chip data (Foat et al., 2006, Tanay 2006, Kinney et al., 2007). The focus of this work, however, is to explain ChIP-chip data in a biophysical framework that is applicable to all PFMs. Hence only publicly available matrices were utilized. This stays also in contrast to a study by Granek

et al., (2005), who employed a biophysical model for RAP1, but did not provide a rationale for choosing the parameters of their model.

As mentioned in Section 4.2.3 the model of Berg and von Hippel makes a number of assumptions about the binding energetics that are likely not reflected by the real binding behaviour of many TFs. The fact that using such a simplified model allows to account for a highly significant fraction of yeast binding data is all but obvious given the noisy data and the complicated binding mechanisms in eukaryotes. In the next chapter I will investigate to what extent the observations made for yeast carry over to multi-cellular organisms.