# Models of transcription factor binding: Sensitivity of activation functions to model assumptions

Dominique Chu, Nicolae Radu Zabet, Boris Mitavskiy


www.elsevier.com/locate/yjtbi

# Models of transcription factor binding: Sensitivity of activation functions to model assumptions

Dominique Chu, Nicolae Radu Zabet, Boris Mitavskiy

PII: S0022-5193(08)00631-0
DOI: doi:10.1016/j.jtbi.2008.11.026
Reference: YJTBI 5388

To appear in: *Journal of Theoretical Biology*

Received date: 14 August 2008
Revised date: 18 November 2008
Accepted date: 29 November 2008

Cite this article as: Dominique Chu, Nicolae Radu Zabet and Boris Mitavskiy, Models of transcription factor binding: Sensitivity of activation functions to model assumptions, *Journal of Theoretical Biology* (2008), doi:10.1016/j.jtbi.2008.11.026

# Models of Transcription Factor Binding: Sensitivity of Activation Functions to Model Assumptions

Dominique Chu[1] and Nicolae Radu Zabet[1] and Boris Mitavskiy[2]

[1] *Computing Laboratory, University of Kent, CT2 7NF, Canterbury, UK*
[2] *A-star Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, 138671, Singapore*

**Abstract**

We present 3 models of how transcription-factors bind to their specific binding sites on the DNA: A model based on statistical physics, a Markov-chain model and a computational simulation. Comparison of these models suggests that the effect of non-specific binding can be significant. We also investigate possible mechanisms for cooperativity. The simulation model suggests that direct interactions between transcription-factors are unlikely to be the main source of cooperativity between specific binding sites, because such interactions tend to lead to the formation of clusters on the DNA with undesirable side-effects.



## 1 Introduction

Controlled binding of transcription factors (TF) to one or more specific binding sites is an important mechanism for cells to regulate gene expression. It is therefore a key-challenge for the cell to be able to control the occupation of individual regulatory sites with the respective TFs in response to changes of external conditions. The overall qualitative form of TF binding is well known; if there is a single binding site, then the probability of the binding site to be occupied approaches 1 as the concentration of free TFs in the cell increases. This saturation curve is often modeled using the Michaelis-Menten function. If there are more than one binding sites, then the transition from low to high binding probabilities is more pronounced and often modeled using the so-called Hill-equation (see below). There are a number of models that can

*Email address:* D.F.Chu@kent.ac.uk (Dominique Chu[1] and Nicolae Radu Zabet[1] and Boris Mitavskiy[2] ).

reproduce these overall predictions in a qualitative way. These models are, however, not all equivalent with respect to their key-assumptions. In many practical modeling situations the simplest models will do; however, for other purposes, such as for example model-based parameter estimation, it will be desirable to have a more detailed model. This article will present three different models of TF binding. Its main purpose is to (*i*) explore their properties and predictions (*ii*) show how they are related to one another and (*iii*) show their limitations.

Gene activation and TF binding is commonly modeled using differential equation approaches (see for example Murray (2008); Chu *et al.* (2008); Narang (2006); Narang & Pilyugin (2007); Zhu *et al.* (2007)) or piecewise-linear differential equations (see for example Alon (2006); Batt *et al.* (2005)). Differential equation models are very convenient from a practical point of view because there is a well developed body of theory to either solve them analytically or at least numerically. The fundamental assumption underlying any such model, however, is that variables are continuous—an assumption that is often too far from the truth to be useful in biological systems. The genome is often realized by only a single molecule and TF numbers can be low (several hundreds). Most differential equation approaches assume that TFs and their binding sites are suspended in the well stirred cytoplasm. As we will show in this contribution, the assumption of the cell being a well stirred reactor makes a *qualitative* difference to the behavior of the model when compared to models that take into account a modest amount of spatial organization.

The assumption of well-stirred reactors can be relaxed in simulation based approaches such as discrete event simulation algorithms (Gillespie (1972); Gibson & Bruck (1998); Ramsey *et al.* (2005)), stochastic model checkers (Kwiatkowska *et al.* (2001)) and process algebras (Regev *et al.* (2001)) (to name but a few). For the understanding of TF binding to specific simulation models are possible choices, but not necessarily the most convenient ones, because the representation of a large number of non-specific binding sites can lead to overly complicated models. Another approach are models based on statistical physics. Ackers and coworkers (Ackers *et al.* (1982)) developed a model of the gene regulation of the $\lambda$-phage repressor (also see Ben-Naim (1997, 199)). More recently Bintu *et al* (Bintu *et al.* (2005*a*,*b*)) presented a number of models to calculate the gene activation function of various operator architectures. The idea of these approaches is to take the weighted sum over all states of the system that are of interest and to divide this by the weighted sum over all possible states to calculate the (steady state) probability of the states of interest to be observed.

Such statistical physics models are very useful in that they often lead to formulas to calculate various quantities associated with the model of interest. One usually has to rely on computational algebra systems to compute the results,

but this is normally still much faster than a discrete-event based simulation of the same system. On the downside, for moderately complex systems formulating the partition function (that is the sum over all possible states) can be a taunting exercise in combinatorics. It is presumably for this reason that previous authors made a number of simplifying assumptions to keep their statistical physics models tractable. In particular they ignored non-focal TFs and assumed that there are no intra-species TF-TF interaction with cooperative effect except between the specific binding sites.

In this contribution we are not so much interested in calculating a specific biological scenario, but instead we are interested in the differences between various modeling *ansatzes* to TF-binding. We recognize that modeling *always* requires simplifying assumptions, but it is essential to understand what error any such simplifying assumption causes. In order to come to a better understanding of this, we compare three modeling approaches. Firstly, based on the above mentioned works by Bintu and Ackers, we develop a statistical physics models of the binding of TFs to DNA. Our model is somewhat less concrete than previous work, but allows for an arbitrary number of specific binding sites. In the appendix to this article we also present an extension that models the case of an arbitrary number of TFs, although we do not elaborate this extended model. One of the immediate conclusions we could draw from this statistical physics model is that in order to compute the probability of a certain number of specific binding sites to be occupied only the number of focal TFs are important. Non-specifically binding TFs can be ignored. However, using a discrete-event computational simulation, we can show that for a slight relaxation of these assumptions, this conclusion becomes incorrect.

We also compare the computational and the statistical model with a Markov-chain model. The latter has the advantage that it can easily incorporate cooperativity and leads to some relatively easy-to-compute formulas. However, this comes at the cost of having to neglect the statistical contribution of non-specific binding, and thus leads to a qualitatively different model. Seen from this perspective, it is questionable to what degree one can generalize from conclusions won from models of gene activations that ignore non-specific binding.

Our analysis is itself based on a number of simplifying key-assumptions: We assume that the DNA is a linear string of binding sites (see below for details). For a particular type of TF only some of the binding sites are specific. TFs bind to all sites, but much stronger to specific than to non-specific sites. In all models below we assume that there are 2 types of TFs. The focal TF-type will have a particular set of specific binding sites on the DNA and will be the molecular type we are interested in. Throughout this manuscript, this will be referred to as *type-1 TF*. The second type of TF—*type-2 TFs*—subsumes all other TFs present in the cell. They are of no direct interest other than their possible interference with the binding properties of type-1 TFs. Furthermore,

in this contribution we will assume (except in the Markov-chain model) that all TFs are always bound to the DNA (although possibly to non-specific sites), rather than be freely suspended in the cytoplasm. They find the specific sites through repeated binding and unbinding to/from non-specific sites, rather than through direct attachment from an unbound state. In cells a certain proportion of TFs will normally be freely suspended in the cytoplasm (Kao-Huang *et al.* (1977); Wunderlich & Mirny (2008)). It feels safe to ignore this effect given that it seems unlikely that a TF directly binds to the specific site from a suspended state in the cytoplasm.

We compare the models by asking the following question with each of them: Given $S_S$ specific binding sites, $S_{NS}$ non-specific binding sites, $N_1$ TFs of type 1 and $N_2$ type-2 TFs, what is the steady state probability that $0 \leq k \leq S_S$ of specific binding sites are occupied? We will describe our statistical physics model in section 2.1; this model is limited to consider two types of TFs. A more general version of the model including a derivation can be found in the appendix section A. A simple Markov-chain model of the same system will be described in section 2.2. Both of these models will be compared to the computational model described in section 2.3. Section 3 will present simulation results obtained with the computational model and relate the three models to one another. Section 4 concludes this article.



## 2 Models

### *2.1 Statistical Physics Model*

We start with the simplest possible case to illustrate the basic principle of statistical physics-based models of TF binding. We assume that TFs are freely suspended in a perfectly mixed aqueous environment of the cytoplasm. Then following Sneppen & Zocchi (2006) we can write the statistical weight of $l$ TFs being bound as:

$$Z_l = \frac{(2V\sqrt{2}(mk_BT)^{(3/2)})^{N-l}}{(N-l)!} \exp\left(-\frac{lG}{k_BT}\right) \tag{1}$$

Here $V$ is the cell volume, $m$ the mass of the TF and $k_B$ the Boltzmann constant and $T$ the temperature. Setting $F = 2V\sqrt{2}(mk_BT)^{(3/2)}$ one can rewrite this equation as

$$Z_l = \frac{F^N}{(N-l)!} \exp\left(-\frac{lG'}{k_BT}\right) \tag{2}$$

Where $G' = G + \ln F$ is the apparent binding free energy. This form draws all changes from $F$ that depend on the number of binding sites into the apparent binding free energy. If there are altogether three binding sites then the probability of exactly $m < 3$ sites being bound is given by:

$$P(\text{exactly } m) = \frac{Z_{l=m}}{\sum_{i=0}^{3} Z_{l=i}} \tag{3}$$

This model would be valid for a short polymer with relatively few non-specific binding sites. DNA molecules, on the other hand, have a large number of binding sites and non-specific binding of TFs to these sites needs to be taken into account. The simplest case is to consider only a single TF with binding free energy $G_s$ for the specific site and $G_n$ for the non-specific sites. Assume that the DNA is a sequence of non-overlapping binding sites, of which $S_{NS}$ are non-specific and there is exactly one specific binding site (i.e. $S_S = 1$). Depending on the binding strength and the temperature, the TF will spend a certain (stochastic) amount of time bound to a binding site, before detaching and re-attaching to a different site. It thus performs a random walk on the DNA. We are not concerned about the details of the random walk here (though see Wunderlich & Mirny (2008)) although we do assume that over an infinite time the TF will sample every site an infinite number of times. In the long run the cumulative binding time to any particular site will depend on the binding strength. For each of the non-specific sites the statistical weight is given by Ackers *et al.* (1982); Bintu *et al.* (2005*a*):

$$w_{NS} = F exp\Big(-\frac{G_n}{k_B T}\Big) \tag{4}$$

Here, $k_B$ is the Boltzmann constant, $T$ the temperature, and $F$ a factor that takes into account some geometrical particulars of the TF and the DNA sequence and is not further characterized[1]. Note that $w_{NS}$ is the statistical contribution of a single non-specific site only; in order to obtain the total weight of all non-specific sites $w_{NS}$ needs to be multiplied by the number of possible ways to occupy the non-specific sites. For the present case of a single particle this is simply the number of non-specific sites $S_{NS}$; hence the total statistical weight of the TF binding to non-specific sites is given by

$$Z_{NS} = S_{NS} w_{NS} \tag{5}$$

Similarly, the statistical weight of the TF being bound to the specific site can be written as:

$$Z_S = F \exp\Big(-\frac{G_s}{k_B T}\Big) \tag{6}$$

[1] In fact it turns out to be irrelevant for computing the probabilities as long as the apparent binding free energy is known

Since there is only one specific site and one TF this case captures all possible ways to occupy the specific site. In order to calculate the probability $P_b$ of a TF to occupy the specific binding site we divide its statistical weight by the total statistical weight of all possible configurations (compare Gerland *et al.* (2002)):

$$P_b = \frac{Z_S}{Z_S + Z_{NS}} = \frac{1}{1 + S_{NS} F \exp\left(\frac{G_s - G_n}{k_B T}\right)} \tag{7}$$

$P_b$ is a sigmoidal function of $G_s$ making the transition from 1 to 0 as $G_s$ increases (assuming a fixed $G_n$). Hence, for a fixed number of non-specific sites the probability of the specific site to be occupied increases to 1 with $G_s \to -\infty$, i.e. increasing binding strength of the specific site. For a fixed specific free energy, but an increasing number of specific sites, $P_b$ falls exponentially to 0.

This very simple model illustrates the basic behavior of TF binding to specific and non-specific sites, but is by itself rather unrealistic. Firstly, normally there will be more than one TF in the cell, there will potentially be more than one specific binding site, and moreover there will be many types of TFs each with their own specific binding sites. In order to keep the complexity of the model manageable we will assume that the specific binding sites are bound with free energy $G_s$ by the $N_1$ type-1 TFs, whereas all other binding sites have a binding free energy $G_n$; the $N_2$ TFs of type-2 bind all sites with $G_n$. This latter assumption is of course not strictly correct as there will be specific binding sites for every species of TF.

If we again use $F_1$ and $F_2$ as factors that take into account geometric aspects of the system, then setting $F = F_1 = F_2$ one can write the partition function as follows:

$$Z = \sum_{i=0}^{S_S} \binom{S_S}{i} \binom{S_{NS}}{N_1 - i} \binom{S_S + S_{NS} - N_1}{N_2} F^{N_2 + N_1} exp\left(-\frac{N_2 G_n}{k_B T} - \frac{i G_s}{k_B T} + \frac{(N_1 - i) G_n}{k_B T}\right) \tag{8}$$

Analogously to the partition function in eq 6 this partition function sums over all possible configurations of TFs of type-1 and type-2 binding to specific and non-specific sites. In order to calculate the probability of a particular configuration, one needs to normalize the statistical weight of the configuration in question by the partition function $Z$. For example, the probability of exactly one specific binding site being occupied by a TF of type 1 is:

$$P(\text{exactly one}) = \frac{Z_{i=1}}{Z} \tag{9}$$

where $Z_{i=1}$ denotes the summand in eq 8 where $i = 1$. The reader can easily convince herself that this partition function leads to the same binding probabilities as equation 7 for $S_S = 1$, $N_2 = 0$ and $N_1 = 1$. Note that equation 9 is independent of $N_2$, i.e. the number of TFs of type 2 and the geometric factors

$F$. This means that, at least in this simple model, the binding probability of type-1 TFs to their specific sites does not depend on the number of type-2 TFs. Similarly independent of $N_2$ is the probability that all binding sites are occupied:

$$P(\text{all three}) = \frac{Z_{i=3}}{Z} \tag{10}$$

Indeed, it can be easily seen that the binding probability of any configuration of type-1 TFs binding to specific sites is independent of $N_2$. Note however, that this conclusion depends on the simplifying assumption that the binding free energy of type-2 TFs is the same for all binding sites. In general this will not be the case. To illustrate this consider the (extreme) case where type-2 TFs have the exact same binding characteristics as type-1 TFs, i.e. bind to all sites with the same free energy as type-1 TFs. In this case, the probability to find a certain number of specific binding sites occupied by type-1 TFs will crucially depend on the number of type-2 TFs. The partition function of this system can be written as follows:

$$\begin{aligned} Z_s = \sum_{\substack{j=\max(0,i-N_2) \\ i=\max(0,N_1+N_2-S_{NS})}}^{\substack{i=S_S \\ j=i}} & \binom{S_S}{i}\binom{i}{j}\binom{S_{NS}}{N_1-j}\binom{S_{NS}-N_1+j}{N_2-i+j} \times \\ & \times F^{N_1+N_2} \exp\left(-\frac{iG_s}{k_BT} - (N_1+N_2-i)\frac{G_n}{k_BT}\right) \end{aligned} \tag{11}$$

Here the double index in the summation symbol indicates two nested sums with the inner index indicating a summation for each value of $i$. The statistical weight of the configurations where all $S_S$ specific binding sites are occupied by type-1 TFs is given by:

$$w_a = \binom{S_{NS}}{N_1}\binom{S_{NS}-N_1}{N_2-S_S} F^{N_1+N_2} \exp\left(-\frac{S_S G_s}{k_BT}\right) \exp\left(-(N_1-S_S+N_2)\frac{G_n}{k_BT}\right) \tag{12}$$

Here we make the reasonable assumption that there are more type-2 TFs than specific binding sites, i.e. $N_2 > S_S$. As before, the probability of all $S_S$ binding sites being occupied is $P_s = w_a Z_s^{-1}$. It can be easily seen that $P_s$ reduces to equation 9 for $N_2 = 0$ and $S_S = 1$. More generally, in the case of $N_2 = 0$ then $P_s$ gives the probability that all specific binding sites are occupied when type-2 TFs always bind with $G_n$.

The mathematical model eq 8 leads to the familiar saturation curves that one would expect from gene activation functions (see in this context also Bintu *et al.* (2005*a*,*b*)). In particular, using eq. 8 to calculate the probability that a

TF is bound to a unique specific binding site yields:

$$P(\text{unique specific site bound}) = \frac{\exp\left(-G_s\right)}{\exp\left(-G_s\right) + \left(\frac{S_{NS}}{N_1} - 1\right)} = \frac{N_1}{N_1 + K(S_{NS} - N_1)} \tag{13}$$

Here we assumed $k_B T = 1$ and $G_n = 0$ and $K = \exp(G_s)$. For $S_{NS}$ large compared to $N_1$ this expression is well approximated by a Michaelis-Menten function (which is also frequently used to describe the dynamics of gene activation functions). Whether or not this assumption is indeed met will depend on the binding free energy to the specific site. The approximation is only good when already a small number of TFs will guarantee full occupation of the binding sites.

### *2.2 Markov-chain model*

It is possible to extend the statistical physics model eq 8 to include cooperativity between TFs. This normally requires recursive relations to calculate the probability of a specific micro-state of the system. Developing this is beyond the scope of this contribution. However, a simplified model that assumes perfect mixing in the cytoplasm and ignores the effect of DNA as a reservoir for particles allows more compact modeling of cooperativity. Assume that all non-specifically bound TFs are conceptually concentrated into one single non-specific site that acts as a "reservoir" for TFs. This simplified scenario corresponds to the case where non-specifically bound TFs are suspended in the cytoplasmic solution (note that this is also the assumption behind the Sneppen and Zocchi model eq 2). Binding to the specific sites happens with a specific rate that depends on the affinity of the TF to the binding site and the concentration of the TFs.

A thus simplified system can be described as a $(S_S + 1)$ state continuous time Markov-chain; the individual states of this chain correspond to $0, 1, \ldots, S_S$ specific sites being occupied. Markov-chains are normally represented as $n \times n$ matrices that describe the rate (in the case of continuous time Markov-chains) or probability (in the case of discrete time Markov-chains) of transition between the possible states. We take here as an example the case of 3 specific

binding sites. The transition matrix for this case is then given by

$$\mathbf{Q} = \begin{bmatrix} -3\,N k_b & 3\,N k_b & 0 & 0 \\ k_u\,C_{-2} & -k_u\,C_{-2} - 2\,(N-1)\,k_b\,C_1 & 2\,(N-1)\,k_b\,C_1 & 0 \\ 0 & 2\,k_u\,C_{-1} & -2\,k_u\,C_{-1} - (N-2)\,k_b\,C_2 & (N-2)\,k_b\,C_2 \\ 0 & 0 & 3\,k_u & -3\,k_u \end{bmatrix} \tag{14}$$

Here $Q_{ij}$ is the rate of transition from state $i$ to state $j$; the state $Q_{00}$ is represented by the top left entry of the matrix. $N$ is the number of TFs; $k_b$ and $k_u$ are the phenomenological binding and unbinding rates respectively. The factor of 3 in $Q_{01}$ is due to the fact that there are 3 free binding sites. $C_{\pm l}$ is the cooperativity modifier, i.e. a factor that determines how the forward and backward binding rates are changed when $l$ other TFs are bound to the binding site. If this value is $> 1$ then we deal with positive cooperativity (i.e. once one site is bound binding to further sites is facilitated), otherwise cooperativity is negative. To illustrate the origin of the entries of this matrix, consider as an example $Q_{12}$ (given by $2k_uC_{-1}$); this term describes the transition from a state where 2 specific sites are occupied to a state where only 1 is. The transition rate is given by twice the unbinding rate of a single bound TF, $k_b$, because at any time either of the two could unbind; the $C_{-1}$ term modifies this rate depending on the cooperativity of the system. The entry $Q_{32}$ ($(N-2)k_bC_2$) describes the transition rate from a state where 2 TFs are already bound to a state where all specific binding sites are occupied. In this case, there are only $N-2$ TFs in the cytoplasm (because 2 are bound already); hence the basic rate of binding $k_b$ must be multiplied by the number of TFs that could bind and a cooperativity modifier ($C_2$). The rationale for all other entries is similar.

The steady state distribution vector $\pi$ of such a continuous time Markov-chain is given by the solution to

$$\begin{aligned} \pi \cdot \mathbf{Q} &= 0 \\ \sum_i \pi_i &= 1 \end{aligned}$$

This is a system of equations that can be solved for each of the $\pi_i$. Solving it for $\pi_4$ yields:

$$\pi_4 = \frac{N(N-1)(N-2)}{N^3 - 3N^2 + 2N + K^3 + c^{-1}(3K^2N + 3K^2N^2 - 3KN)} \tag{15}$$

Here we assumed that all cooperativity terms $C_{\pm l} = c$ to simplify the equation and we set $K := k_{ub}/k_b$. In the limit of infinite cooperativity $c \to \infty$ the

parenthesis in the denominator goes to 0, leading to the expression:

$$\lim_{c\to\infty} \pi_4 = \frac{N(N-1)(N-2)}{N(N-1)(N-2)+K^3} \tag{16}$$

This expression is well approximated by a Hill function as long as $N$ is large enough.[2] Equation 16 suggests that the Hill coefficient is limited by the number of binding sites. This indicates that switch like gene activation functions need to include additional mechanisms, simply because the number of TFs controlling one particular gene is limited. In order to achieve thresholding or switching behavior it might be necessary to couple gene activation to transduction pathways (such as the Koshland Goldbeter (Goldbeter & Koshland (1981); Tyson *et al.* (2001)) switch) or to form multi-mer TFs (also see Tyson *et al.* (2003) in this context).

Turning our attention now to the case of no cooperativity, i.e. $c = 1$ equation 15 becomes:

$$\pi_4 = \frac{N(N-1)(N-2)}{N^3-3N^2+2N+K^3+3K^2N+3KN^2-3KN} \tag{17}$$

This model can be related to the statistical physics model eq 8. In particular, looking at the probability of three binding sites being occupied in eq 10 if one ignores the contribution from the non-specific sites to the statistical weight, i.e. $S_{NS}/(S_{NS}-N1+i) = 1$, then the statistical model corresponds to the Markov-chain model in eq 14. In this case eq 10 can be expanded as:

$$P(\text{all three}) = \frac{3!\binom{N}{3}\exp(-3G)}{\sum_{i=0}^{3} i!\binom{N}{i}\exp(-iG)} \tag{18}$$

Here again we set $k_BT = 1$ for notational convenience. Expanding the binomial coefficients in equation 18 and setting $K = e^G$ yields after some simple yet tedious manipulations:

$$P(\text{all three}) = \frac{N(N-1)(N-2)}{K^3+3NK^2+3N^2K-3NK+N^3-3N^2+2N} \tag{19}$$

[2] In Biochemistry cooperativity is often associated with the value of the Hill coefficient, i.e. the value of $h$ regulating the sharpness of the transition from the minimal to the maximal value in the so-called Hill-function:

$$P(x) = \frac{x^h}{K^h+x^h}$$

Here $P(x)$ is the probability that all $S_S$ operator sites are occupied; $K$ is a parameter indicating the number (or concentration) of TFs where $P(K) = \frac{1}{2}$. Generally, the higher the value of $h$ the more switch-like the function $P(x)$.

This is the same as eq 17 showing the equivalence of the Markov model with the statistical physics model when the statistical contribution of the non-specific binding sites is ignored in the latter. A similar exercise shows that the Sneppen and Zocchi model eq 2 can be brought into the same form.

### *2.3 Computational Model*

This section describes a computational simulation model of TF binding. This model explicitly represents the DNA sequence and the TFs populating the sequence. The DNA sequence is a random string of length $l$ composed of a 4 letter alphabet. Its length and composition bias can be set arbitrarily by the user. TFs bind to the DNA string with free energy $G(\mathbf{s})$, where $\mathbf{s}$ is the particular sequence to which the TF attaches; when the sequence $\mathbf{s}$ of the string coincides with the specific binding site, then we call $\mathbf{s}$ the *binding motif*. The length of $\mathbf{s}$ is equal for all TFs and fixed during a simulation and equal for all TFs in the model (in all simulations considered here the length was kept fixed at the arbitrary value of 9). We used two different rules to determine the binding energy:

(1) There are only two binding energies, namely a specific binding energy and a non-specific one. TFs bind with the non-specific free energy $G_n$ unless they are of type 1 **and** the binding sequence exactly matches the binding motif, in which case the TF binds with energy $G_s$. This rule corresponds exactly to the above described theoretical model eq 8, but is an approximation with respect to the real biological case. It is more realistic to assume that the non-specific binding free energy is sequence dependent.
(2) For TFs of type 1 the binding energy is calculated as $G(\mathbf{s}) = \sum_{\mathbf{s}} \epsilon \cdot \delta_{s_i,s_i^t}$ where $\mathbf{s}^t$ denotes the binding motif and $\epsilon$ a factor representing the contribution from each matching nucleotide and $\delta_{x,y} = 1$ for $x = y$ and 0 otherwise. So, for example, if the binding motif is `aatc` and the actual sequence is `atgc` then $G(\texttt{atgc}) = \epsilon_1 + 0 + 0 + \epsilon_4 = 2\epsilon$. TFs of type 2 bind either with a user determined fixed energy $G_2$ or with the same motif matching rule as type-1 TF.

The second scenario is generally held to be a good approximation to the biochemical reality (see Gerland *et al.* (2002)); the first binding rule is still of interest here because the binary distinction between specific and non-specific binding sites allows a direct comparison of the computational model with the statistical physics model. Note, however, that even the first rule is somewhat different to the above mathematical model 8 that assumed the DNA to be partitioned into separated binding sites. In the computational model each TF has a binding motif of length $l$ and thus occupies at least $l$ nucleotides. On

circular DNA strands there will still be $L$ binding sites if the DNA is composed of $L$ nucleotides.

The update algorithm of the model is as follows:

(1) The simulation is initialized by populating the DNA string with a user-determined number of TFs of type-1 and type-2. The time is set to zero and each TF is assigned a binding time drawn from an exponential distribution with mean $G(\mathbf{s})$, where $\mathbf{s}$ indicates the particular binding site and $G(\mathbf{s})$ the free binding energy appropriately calculated for TFs of different types (see above).
(2) The TFs are placed in a list ordered with respect to the remaining binding time; the TF with the lowest remaining binding time is the top element.
(3) The top TF of this list is updated, i.e. removed from its current binding site and moved to a randomly chosen new position.
(4) The system time is set to $T_b + T_u$ where $T_b$ is the system time when the top TF attached to its current site and $T_u$ is the total time it was bound to this site.
(5) A new binding position is determined for the TF as in step 1 and it is assumed that the time required for TF to move from a position to the next is negligible compared to the time they spend bounded to the DNA.
(6) A new binding time is determined for this TF and it is added to the ordered list at the appropriate position.
(7) The procedure continues with step 3.

The location of the specific binding sites on the DNA can be determined by the user. During the simulation the cumulative occupation time of each of the specific binding sites is recorded. After a user-specified system-time (i.e. not real simulation time) the simulation is stopped and estimates for the relevant binding probabilities are calculated by dividing the actual occupation time by the total system time. In the limit of an infinite system time this would give an exact value for the binding probabilities, however, at the expense of an infinite simulation time. In practice we found a system simulation time of 100000 time units to be sufficient to give fairly accurate estimates of the binding probabilities (as indicated by the scatter of the results) while allowing reasonable simulation times.

The model also allows cooperativity. In real systems there are at least two different possible sources for cooperativity. One way to think about it is that attachment of a TF leads to a local conformational change of the DNA which in turn leads to an increased (or decreased) binding affinity of other TFs. Another possible source of cooperativity is direct interaction between TFs: If two TFs bind close to one another and form bonds between them as well as with the polynucleotide then this will result in an increased period of residency of the individual TFs on their respective binding sites. Which one of those

mechanisms is biologically more important is unclear; indeed there might be other causes of cooperativity. Corresponding to those two possible mechanisms in the model the user can choose between two types of cooperativity. **Cooperativity-1** is assumed to be effective at the specific binding sites only; one can think of it as being mainly caused by indirect effects (such as conformational changes of the DNA). **Cooperativity-2** is effective between any pair of adjacent type-1 TFs whether or not they are bound to specific sites. Biologically, cooperativity-2 can be thought of as being due to direct TF-TF interactions.

Cooperativity-1 is implemented as follows:

- Upon binding to a specific site $\Sigma$ the total number of occupied specific binding sites (other than the current) is determined. This number is $n$.
- Assuming $n > 0$, two random binding times are drawn from an exponential distribution. Firstly, $T1$ is the binding time for the specific binding site in absence of cooperativity; $T2$ is the binding time taking into account cooperativity and is drawn from an exponential distribution with mean $G_s + nc^M$ where $c^M$ is the cooperativity parameter specified by the user. Note that $c^M$ is different from the cooperativity parameter $c$ of the Markov-chain model. In the computational model, the case of no cooperativity is realized by $c^M = 0$ whereas in the Markov-chains model it is $c = 1$. In general there is no simple relationship between $c$ and $c^M$.
- The TF at $\Sigma$ binds for a period of $T2$; the other TF at specific sites have the value $\max(0, T2 - T1)$ added to their binding time.

The algorithm for cooperativity-2 is similar, yet instead of taking into account all TFs bound to specific sites all TF modify their binding properties according to the number of TF of the same type that bind to immediately adjacent sites.

## 3 Results and Discussion

We first check that the mathematical model 8 indeed matches the predictions of the computational model given the same parameters. Figure 1 confirms for a specific set of parameters (see figure caption) that there is good agreement between the model and the theoretical predictions; note that figure 1 shows results obtained from simulations with overlaping binding sites, whereas the model eq 8 assumes non-overlapping sites[3]. The simulation results show a certain degree of noise; this noise could be reduced by increasing the time

[3] Strictly speaking figure 1 only confirms the agreement between model and simulation for the particular parameter set used. However, we found similar (or better) agreement for all parameter sets we tested (data not shown).

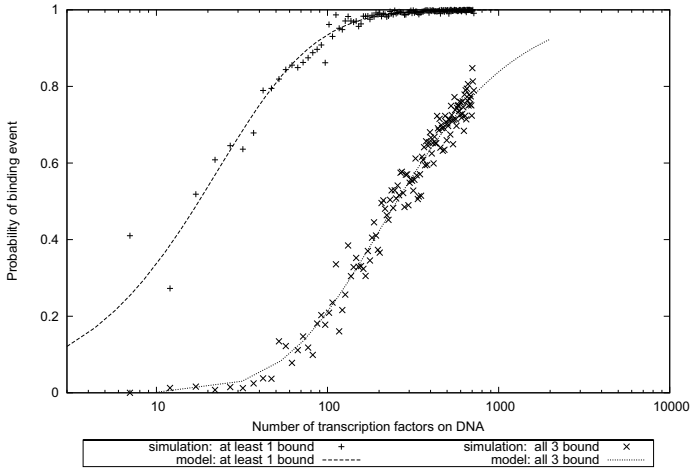


Fig. 1. Comparison of the mathematical model with the simulation. The solid curves indicate the prediction of the statistical physics model (eq 8) and the dots are results of simulations. Each dot represents the result from a single run of the model. The label "at least 1 bound" indicates the probability that at least one of the 3 specific binding sites is occupied. The label "all 3 bound" indicates the probability of all 3 specific binding sites being occupied. A logarithmic scale for the x-axis was used to improve readability. The parameters used in this figure are: $G_s = -5k_BT$, DNA size: 10000, length of motif: 9; the probability of TF binding in the simulation was obtained by averaging over 10000 time units.

over which results are averaged, although only at the expense of increased computational costs. We could also confirm the prediction of the theoretical model that the probability of specific sites being occupied is independent of the number of type-2 TFs (data not shown).

While the statistical physics model eq 8 does agree with the simulation results, the Markov-chain model eq 15 does not (data not shown). As discussed above, the Markov-chain model is in general not equivalent to the statistical physics model eq 8 even for the case of $c = 1$. It is therefore not surprising that it does not reproduce the data for the same value of $K$.

Figure 2 shows a fit of the Markov-chain model to two simulations of the computational model with identical parameters but cooperativities of $c^M = 1$ and $c^M = 5$ respectively (see legend of figure 2). In both cases a good fit can be obtained and the fit correctly assigns a higher cooperativity factor to the $c^M = 5$ simulation. It also assigns a different $K$ to both simulations, which is incorrect, as these simulations are only distinguished by their different cooperativities but not by their $K$. Attempting to fit the Markov-chain model with the $K$ obtained for the $c^M = 1$ case to the simulation with $c^M = 5$ however is not successful, as can be seen in figure 2.

Figure 3 shows example simulations of the system with cooperativity-2 (i.e. local TF-TF interactions) enabled. If there is only one TF-species in the sim-

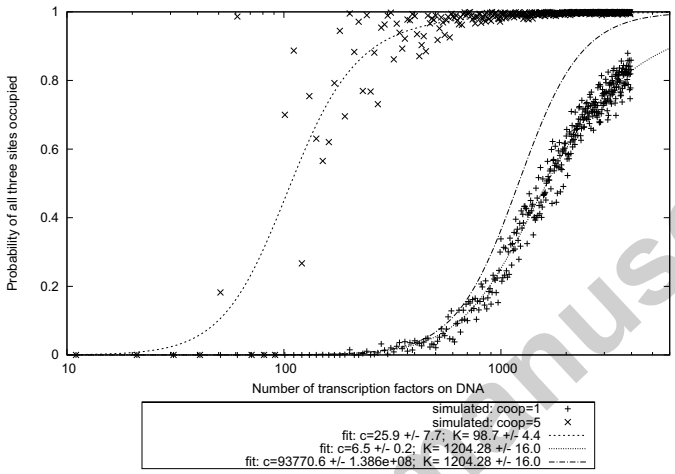


Fig. 2. Comparing the computational model with cooperativity-1 with the Markov-chain model eq 15; we use logscale to improve readibility of the graph. The dots show simulations of the same model with two different cooperativity factors $c^M = 1$ and $c^M = 5$. It is possible to fit the Markov-chain model to both simulations. As expected the fit leads to a higher cooperativity $c$ for the simulation with higher cooperativity $c^M = 5$. However, the $K$ values of the two fits do not co-incide: Using the $K$ values obtained from one fit (to the $c^M = 1$ simulation) and attempting to fit the model to the $c^M = 5$ simulation is not successful. As can be seen in the graph, this leads to a very bad fit (the dashed line). The simulation used the following parameters: DNA size: 100000, number of specific binding sites: 3, length of motif $l = 9$, contribution per correct base $\epsilon = 0.3333$.

ulation then, for medium and high numbers of TFs and medium to high cooperativity the occupation probabilities obtained from simulation runs fell into two distinct classes: During each particular run the specific binding sites were either (close to) permanently or (nearly) never occupied with very little in between. Which of those outcomes is realised in a particular run is a probabilistic choice of the system (with some bias; see below). The splitting of outcomes is clearly visible in figure 3 for the simulation with only 1 species; for more than about 150 TF the occupation probabilities are either very high or very low. Note that this effect is an artifact of the limited averaging time to estimate the occupation probability. More accurate estimates of the true occupation probabilities are possible but would become increasingly expensive in terms of computing time. Note, however, that biological cells are limited by a similar time constraint. They themselves do not "see" the true steady state of a system but must average over some finite time-period. The above failure of the computational model to give an accurate estimate of the steady state probabilities, although an artifact, is therefore likely to be of some biological relevance.

The underlying cause for this effect is the formation of clusters of adjacent TFs forming strong cooperative bonds to the DNA. Once such a cluster forms it would be very stable over time. If it happens to cover the specific binding sites, then these will be stably occupied for a long time. If, however, they are not covered, then the stability of the clusters means that the waiting time before they are covered will be very long. The transition between situations where the TFs cover their specific sites and where they are not is very slow. For moderately strong cooperativities, TF-TF interactions provide a much larger contribution to the binding strength than the specific sites. One would thus expect the importance of (i.e. frequency of binding to) the specific binding sites to diminish relative to the clustering effect.

This is confirmed by figure 4 which provides another perspective of the same phenomenon. It shows a histogram of the observed fraction of times of the specific sites being occupied for many re-runs of the model with identical parameter settings (see figure caption for details). For high coperativity ($c^M = 5$) the bars of the histogram concentrate at the extreme ends near the occupation probabilities of 1 or 0 indicating that the specific sites are either always covered or never. Lowering the cooperativity from 5 to 1 reveals a different picture. There is still the possibility that the specific binding sites are never covered or (nearly) always; in addition there is another maximum around 0.3 showing that in some simulations the specific sites are sometimes covered; this suggests that there are a number of runs where the formation of clusters does not happen or happens to a lesser degree. Lowering the coperativity even further to below 1 (data not shown) will lead to a single maximum around a specific probability.

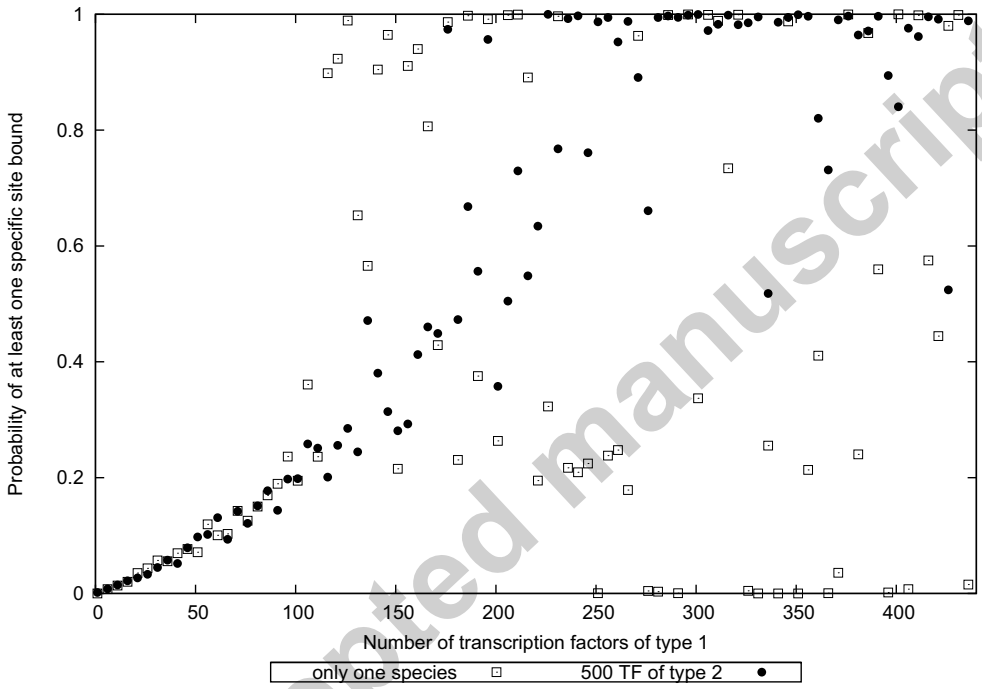


Fig. 3. If cooperativity-2 is turned on, then clusters with high stability may build up on the DNA. In this case, the probability of the specific binding site being occupied is reduced. This will, however, not be a problem if there are also other transcription factors available that do not cooperate with the type-1 TFs.

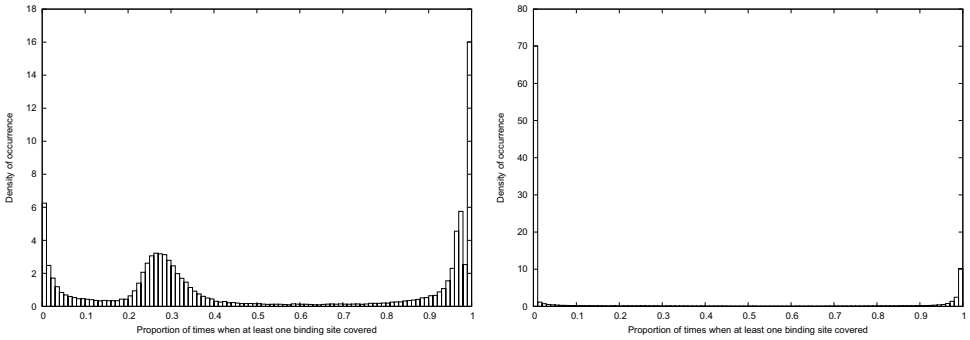


Fig. 4. Histograms showing the distribution of probabilities that at least one of the three binding sites is covered by a TF. The left graph shows the case of $c^M = 1$ and the right graph $c^M = 5$. These graphs were produced by repeating a simulation with fixed parameter setting and then recording the proportion of time the condition "At least one specific site covered" was met. The figure shows that for high cooperativity the solution is "all or nothing," whereas for a lower cooperativity ($c^M = 1$) there are also cases where the binding sites are occupied with a given probability (here around 0.3). The parameters used to generate these graphs are as follows: DNA-size: 10000, $\epsilon = 1/2$

Assuming there is no cooperativity between TFs of different species one is led to conjecture that introducing type-2 TFs will tend to reduce the clustering, because the second species would occupy space and prevent type-1 TFs from forming too long chains. Figure 3 shows that introducing a second species of TF on the DNA indeed restores the (statistical) predictability of TF-binding, at least up to a certain point (about 300 type-1 TFs in the case of this graph). Clustering, can only be reversed for low cooperativities. This can be understood by considering that for high cooperativities a single pair will tend to have long resident times on the DNA compared to adjacent type-2 TFs. They will thus sample a larger number of neighboring TFs, which increases the probability that one of those is of type-1 and thus increases the size of the cluster. Hence a higher cooperativity parameter $c^M$ tends to lead to longer chains.

This suggests that cooperativity based on direct TF-TF interactions is mechanistically problematic and for this reason possibly selected against over evolutionary times. At least, we would expect it to be of subordinate importance only, because it would lead to TF clustering on the DNA if the cooperative interactions are too strong. On the other hand if they were only weak then they would lead to a relatively minor modification of the probability of binding. So either way the conclusion from this is that localised effects at the specific site, possibly mediated through conformational changes upon binding of TFs is biologicaly more plausible as a mechanism for cooperativity; this is in line with previous empirical findings (Ben-Naim (199)).

So far we assumed that the type-2 TFs do not discriminate between binding sites and generally have a low affinity to the DNA. The qualitative conclusion

from the basic model eq 8 was that in this case the number of type-2 TFs are immaterial for the occupation probability of the specific binding sites. In the context of clustering due to TF-TF interactions, it became apparent that this conclusion is not always correct. To the extent that TF-TF cooperativity does exist, it seems that the presence of type-2 TFs plays a certain role in avoiding the above mentioned clustering.

We now extend the basic model and assume that also type-2 TFs have their own binding profile in the sense that their binding affinities are determined according to the same rules as those of type-1 TFs. (We do not drop the assumption that the cooperativity between TFs of different types are negligible.) Figure 5 shows 3 scenarios of this modified system: In the first two scenarios the binding motifs of type-1 and type-2 TFs are non-overlapping, i.e. their respective binding motifs do not share a single position. In this case one type of TF has a miminal binding free energy for the specific binding sites of the other type and thus spends minimal time on the specific sites of the other type. For the parameters used in the example simulation in figure 5 the binding probability increases near linearily and reaches about 0.7 for 50 TFs.

The situation changes very much in the other extreme case when the binding motifs of both types of TFs completely overlap. This case can be treated mathematically and is given by equation 12. Both types of TFs have equal binding times (on average) and there will thus be direct competitive binding to the sites; the probability of type-1 TFs to bind the specific sites will then strongly depend on the number of type-2 TFs. Figure 5 illustrates this scenario and shows that the (comparatively low number of) type-1 TFs are crowded out from the specific site by the much higher number of type-2 TFs. This would make control of the occupation of the binding sites inefficient.

Control over the operator region can be restored by cooperativity. Figure 5 shows a simulation of TF-binding to the specific sites when both type-1 and type-2 binding motifs are identical, but only type-1 TFs show cooperativity. Already a moderate number of TFs (about 50) leads to binding probabilities even higher than in the case of non-overlaping motifs. Interestingly, the increase of the binding probability as a function of the number of type-I TFs is very steep. The graph in figure 5 uses cooperativity-2, resulting in a rather noisy transition; cooperativity-1 leads to a similar result but considerably less noise (data not shown).

The conclusion from this is similar as above: Once one takes into account that type-2 TFs have themselves a specific binding profile, then a whole new range of potential interferences between types of TFs can arise. These effects were not visible in the statistical physics model eq 8 and they are certainly absent from the simpler differential equation models in the literature or the Markov-chain model eq 14. The question that arises now is to what degree one

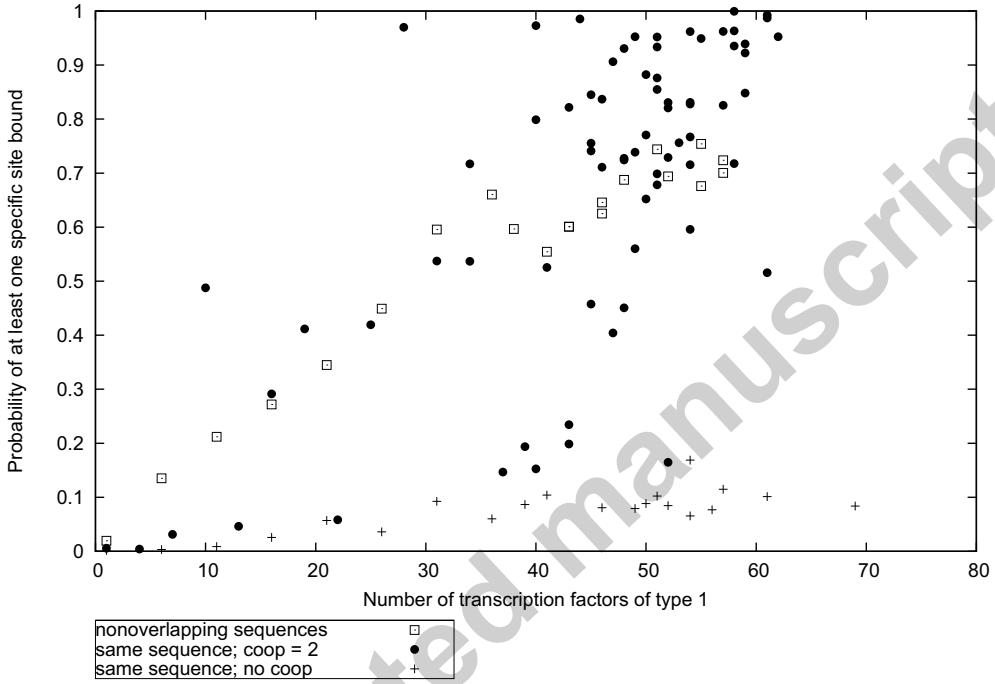


Fig. 5. The probability of at least one specific binding site being bound. The label "same sequence" refers to a run where the type 2 transcription factors have the same binding motif as the type-1 transcription factors and there is cooperativity between type-1 TFs; correspondingly, the points labelled "nonoverlapping sequences" refer to simulations where the binding motif of the type-2 transcription factors does not overlap that of the type-1 transcription factors and there is no cooperativity. In all runs there were 900 type-2 TFs on the DNA of length 10000, $\epsilon = 0.6$ and $c^M = 2$ (where applicable).

can still trust any conclusions that are drawn from the elegant but potentially over-simplifying models in the literature?

## 4 Conclusion

We have presented a statistical physics model of TF binding to DNA and compared it to a simpler Markov-chain model and a computational simulation model. The statistical physics based model (eq 8) predicted that the probability to find a specific binding site occupied depends only on the number of focal TFs (i.e. the number of TFs for which this site is specific). This conclusion, however, is only justified if at least the following two conditions are met:

- There are no TF-TF interactions: Once one allows for direct interactions between adjacent TFs, then this can lead to clustering of TFs on the DNA. The number of non-specific TFs is then crucial to retain cellular control over the specific sites.
- Type-2 TF have equal affinity to all sites: This assumption is certainly not correct and changing it can have noticeable consequences for the binding of TFs to the DNA.

One has to assume that in real genomes, these conditions are normally not met; hence the statistical physics model above (eq 8) is incorrect in a qualitative way. Formulating more realistic partition functions by taking into account sequence dependent binding or cooperativity is possible in principle; in practice it quickly leads to very complex and intractable models. The main reason for this explosion of model complexity is the necessity to count over all possible states and all possible binding sites.

Models that assume that the cell is a perfectly mixed solution of TFs (such as the above Markov-chain model eq 14) are much simpler to formulate and compute. The essential simplification of these models compared to statistical physics models is that they do not take into account that DNA acts as a reservoir for TFs. As such they ignore a potentially important spatial aspect of the system. Our comparison indicates that ignoring this effect leads to models that are fundamentally different to statistical physics models. The Markov-chain model eq 14 can in general not be fitted to the statistical physics model that takes into account non-specific binding. Figure 2 shows a case where it could be fitted to simulation results of the computational model with cooperativity-1; in this case it led to a wrong estimate for the binding free energy of the system.

For theoretical investigations, the increased tractability of the simplified models will in many cases compensate for their relative inaccuracy. When they are used to estimate system parameters (for example via fitting) then these models are normally not suitable because they would lead to incorrect paramter values.

## References


Ackers, G. K., Johnson, A. D. & Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Science USA,* **79** (4), 1129–1133.

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall.

Batt, G., Ropers, D., deJong, H., Geiselmann, J., Mateescu, R., Page, M. & Schneider, D. (2005). Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli. *Bioinformatics,* **21**, i19–28.

Ben-Naim, A. (199). Cooperativity in binding of proteins to DNA. ii. vinding of bacteriophage *lambda* repressor to the left and right operators. *The Journal of Chemical Physics,* **108** (16), 10242–10252.

Ben-Naim, A. (1997). Cooperativity in binding of proteins to DNA. *The Journal of Chemical Physics,* **107** (23), 10242–10252.

Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T. & Phillips, R. (2005*a*). Transcriptional regulation by the numbers: applications. *Current Opinion in Genetics and Development,* **15** (2), 125–135.

Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J. & Phillips, R. (2005*b*). Transcriptional regulation by the numbers: models. *Current Opinion in Genetics and development,* **15** (2), 116–124.

Chu, D., Roobol, J. & Blomfield, I. (2008). A theoretical interpretation of the transient sialic acid toxicity of a *nanR* mutant of *Escherichia coli*. *Journal of Molecular Biology,* **375**, 875–889.

Gerland, U., Moroz, J. & Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-dna interaction. *Proceedings of the National Academy of Science USA,* **99** (19), 12015–12020.

Gibson, M. & Bruck, J. (1998). An efficient algorithm for generating trajectories of stochastic gene regulation reactions. Technical Report CaltechPARADISE:1998.ETR026 California Institute of Technology.

Gillespie, D. (1972). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry,* **81**, 2340–2361.

Goldbeter, A. & Koshland, D. E. (1981). An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Science USA,* **78** (11), 6840–6844.

Kao-Huang, Y., Revzin, A., Butler, A. P., O'Conner, P., Noble, D. W. & von

Hippel, P. H. (1977). Nonspecific dna binding of genome-regulating proteins as a biological control mechanism: measurement of dna-bound escherichia coli lac repressor in vivo. *Proceedings of the National Academy of Science USA,* **74** (10), 4228–4232.

Kwiatkowska, M., Norman, G. & Parker, D. (2001). PRISM: probabilistic symbolic model checker. In *Proc. Tools Session of Aachen 2001 International Multiconference on Measurement, Modelling and Evaluation of Computer-Communication Systems*, (Kemper, P., ed.), pp. 7–12,. Available as Technical Report 760/2001, University of Dortmund.

Murray, J. (2008). *Mathematical Biology: An Introduction: Pt. 1.* Springer-Verlag.

Narang, A. (2006). Comparative analysis of some models of gene regulation in mixed-substrate microbial growth. *Journal of Theoretical Biology,* **242** (2), 489–501.

Narang, A. & Pilyugin, S. (2007). Bacterial gene regulation in diauxic and non-diauxic growth. *Journal of Theoretical Biology,* **244** (2), 326–348.

Ramsey, S., Orrell, D. & Bolouri, H. (2005). Dizzy: stochastic simulation of large-scale genetic regulatory networks. *Journal of Bioinformatics and Computational Biology,* **3**, 415–436.

Regev, A., Silverman, W. & Shapiro, E. (2001). Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Pacific Symposium on Biocomputing 2001* pp. 459–470,.

Sneppen, K. & Zocchi, G. (2006). *Physics in Molecular Biology.* Cambridge University Press.

Tyson, J., KChen & Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current Opinion Cell Biology,* **15** (2), 221–231.

Tyson, J. J., Chen, K. & Novak, B. (2001). Network dynamics and cell physiology. *Nat Rev Mol Cell Biol,* **2** (12), 908–916.

Wunderlich, Z. & Mirny, L. (2008). Spatial effects on the speed and reliability of protein-DNA search. *Nucleic Acids Research,* **36** (11), 3570–3578.

Zhu, R., Ribeiro, A., Salahub, D. & Kauffman, S. (2007). Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. *Journal of Theoretical Biology,* **246** (4), 725–745.



## A Generalisation of the statistical physics model for the case of more than two types of TF

In section 2.1 we only considered the case of 1 type of TF with specific binding sites and a second type that has no specific binding sites. It is possible to write down a partition function for the more general case of $n$ types of TFs (denoted as $\alpha_i$ each with its own number of specific sites $A_i$. The problem boils down

to correctly counting all possible ways to distribute all TFs of different types over the possible specific and non-specific sites. One way to do this is to count all the ways that TFs can be distributed over the DNA as follows:

(1) Distribute each of the $\alpha_i$ over their specific binding sites $A_i$.
(2) Then distribute the remaining $\alpha_i$ over the specific binding sites of other types of TFs, i.e. all remaining free $A_j$, where $j \neq i$.
(3) Finally distribute all remaining TFs over all non-specific sites.

The first step, i.e. distributing each TF of type $\alpha_i$ over their $A_i$ specific binding sites. This leads to a binomial coefficient $\binom{A_i}{x_i}$ for each type of TF.

The next step is to distribute the TFs over the remaining free specific binding sites of other types of TFs. For this we need to choose for each type of TF $j$ a number of molecules to distribute over the specific binding sites of all other types of TFs, i.e. we must select a number $y_r^k$ of TFs of type $k$ among the binding sites of type $r$ (here $r = q$ or $r = q+1$ (if $q = k$)). At every step in the second round the total number of TFs already placed on binding sites of type $r$ is $x_r + \sum_{j<k \text{ and } j\neq r} y_r^j$ so that the number of available binding sites of type $r$ is $A_r - x_r - \sum_{j<k \text{ and } j\neq r} y_r^j$; the sum represents the binding sites occupied in the first round (i.e. the specific sites occupied by their native TFs) and at every previous step in this second round (i.e. the specific binding sites that have so far been nonspecifically bound). The number of ways to place these $y_r^k$ TFs of type $k$ on the remaining binding sites of type $r$ is then

$$\binom{A_r - x_r - \sum_{j<k \text{ and } j\neq r} y_r^j}{y_r^k}.$$

Care must be taken to select the indices correctly, i.e. $0 \leq y_r^k \leq \min\{A_r - x_r - \sum_{j<k \text{ and } j\neq r} y_r^j,\, a_k - x_k - \sum_{1\leq h\leq r \text{ and } h\neq k} y_h^k\}$. Finally, having positioned the appropriate numbers of TFs on the specific binding sites, the still remaining $a_i - x_i - \sum_{0\leq j\leq n \text{ and } j\neq i} y_j^i$ TFs of type $i$ must be distributed among the nonspecific binding sites for every one of the types $i$. The number of ways to do this is given by the multinomial coefficient

$$\binom{W}{z_1,\, z_2,\, \ldots,\, z_i, \ldots, z_n,\, W - \sum_{i=1}^n z_i} = \frac{W!}{z_1! z_2! \ldots z_n! (W - \sum_{i=1}^n z_i)!}$$

where $z_i = a_i - x_i - \sum_{0\leq j\leq n \text{ and } j\neq i} y_j^i$. To reduce the notational complexity, we introduce the sets of allowable indices

$$S_j^i = \left\{ h \in \mathbb{N} \,|\, 0 \leq h \leq \min\left\{ A_r - x_r - \sum_{j<k \text{ and } j\neq r} y_r^j,\, a_k - x_k - \sum_{1\leq h\leq r \text{ and } h\neq k} y_h^k \right\} \right\}$$

and also a shorthand notation for the binomial coefficients

$$\mathcal{Y}_r^k = \binom{A_r - x_r - \sum_{j<k \text{ and } j\neq r} y_r^j}{y_r^k}$$

(keep in mind that $r \neq k$). Our final counting formula is then as follows:

$$\mathcal{Z}_{\mathbf{x}} = \prod_{i=1}^{n} \binom{A_i}{x_i} \sum_{y_2^1 \in S_2^1} \mathcal{Y}_2^1 \sum_{y_3^1 \in S_3^1} \mathcal{Y}_3^1 \dots \sum_{y_n^1 \in S_n^1} \mathcal{Y}_n^1 \sum_{y_1^2 \in S_1^2} \mathcal{Y}_1^2 \times$$
$$\times \sum_{y_3^2 \in S_3^2} \mathcal{Y}_3^2 \sum_{y_4^2 \in S_4^2} \mathcal{Y}_4^2 \dots \sum_{y_{n-1}^n \in S_{n-1}^n} \mathcal{Y}_{n-1}^n \frac{W!}{z_1! z_2! \dots z_n! (W - \sum_{i=1}^n z_i)!} \quad \text{(A.1)}$$

This counts the total number of ways to distribute the TFs over the various binding sites for a particular assignment of TFs to their specific binding sites corresponding to the statistical weight

$$w_{\mathbf{x}} = \sum_{i=1}^{n} F_i^{\alpha_i} \exp\left( - \sum_j \left( \frac{x_j G_s^j}{k_B T} + \frac{x_j G_n^j}{k_B T} \right) \right).$$

In order to obtain the entire partition function $Z$ must be summed over all feasible values of $\mathbf{x}$.

$$Z = \sum_{\mathbf{x}} \mathcal{Z}_{\mathbf{x}} w_{\mathbf{x}}$$

## B Dependence on nucleotide composition

Throughout this article it was assumed that the DNA is not biased with respect to its base composition. In real bacteria, this assumption is not correct and needs to be taken into account for quantitatively correct models of TF-binding. However, our simulations indicate that the base composition only plays a relatively minor role. Figure B.1 shows a number of simulations for DNA strings with extreme nucleotide biases.

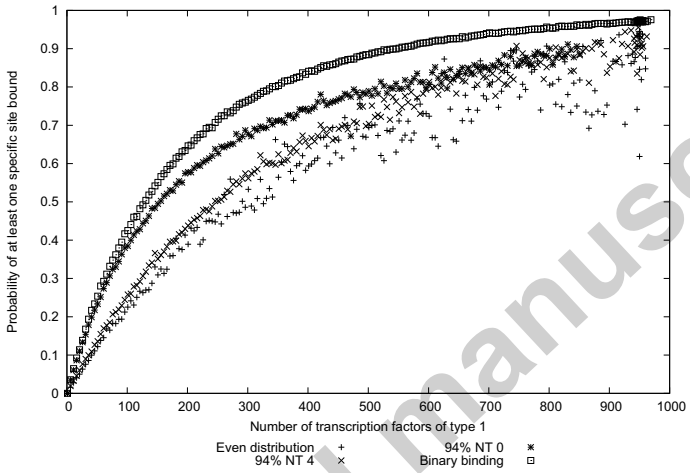


Fig. B.1. This graph shows simulations with identical parameters, but different nucleotide compositions. The motif sequence in these runs was `221331111`. In the curve labeled "94% NT 0" the nucleotides of the genome were chosen with probability 0.94 to be 0; all other bases were chosen with equal probability. The label "Even distribution" indicates that all nucleotides were chosen with probability 0.25. "Binary binding" means that perfect binding sequences have a strong affinity ($G_s = -3$) and all other sites have a low affinity ($G_n = 0$). The length of the DNA was 10000 and there was no cooperativity, $\epsilon = 0.33333$ (where applicable).