How does the DNA sequence affect the Hill curve of transcriptional response?

M Sheinman and Y Kafri

1 Department of Physics and Astronomy, Vrije Universiteit, Amsterdam, The Netherlands
2 Department of Physics, Technion, Haifa 32000, Israel

E-mail: mishashe@gmail.com

Received 18 May 2012
Accepted for publication 13 July 2012
Published 29 August 2012
Online at stacks.iop.org/PhysBio/9/056006

Abstract

The Hill coefficient is often used as a direct measure of the cooperativity of binding processes. It is an essential tool for probing properties of reactions in many biochemical systems. Here, we analyze existing experimental data and demonstrate that the Hill coefficient characterizing the binding of transcription factors to their cognate sites can in fact be larger than one—the standard indication of cooperativity—even in the absence of any standard cooperative binding mechanism. We demonstrate that this effect occurs due to the disordered binding energy of transcription factors to the DNA molecule and the steric repulsion between the different copies of the transcription factor. We show that the enhanced Hill coefficient implies a significant reduction in the number of copies of the transcription factors which is needed to occupy a cognate site and, in many cases, can explain existing estimates for number of copies of the transcription factors in cells.

1. Introduction

Molecular recognition plays an important role in biological systems ranging from antigen–antibody identification to protein–protein binding [1]. In many cases, the recognition process is driven by free-energy differences between a desired reaction and many competing undesired reactions [2–5]. One example, of particular importance, is that of protein–DNA interactions. Appreciation of their role in transcriptional regulation [6] has led to a large experimental effort which, in particular, aims to map the binding energy of transcription factors (TFs) to different subsequences on the DNA [7]. It is known that in many cases the energy can be written to a good approximation as a sum of energies representing the binding energy of a nucleotide on the DNA to the region on the protein with which it is aligned [8, 9]. Specifically, the binding energy of a nucleotide \( s = A, C, G, T \) to position \( j = 1, 2, \ldots, L \), where \( L \) is the number of basepairs occupied by a DNA binding domain (typically in the range of 10–30), is usually described by a \( 4 \times L \) position weight matrix (PWM), \( \epsilon_{s,j} \). By now, the PWM is known for many TFs and, together with knowledge of the genomic sequence, it specifies the binding energy landscape of a TF to the DNA.

Irrespective of the energy landscape properties, the activation of a cognate site—a specific location on the DNA—by a TF is usually described by a Hill curve [10]. Namely, if we consider a DNA molecule inside a container representing, say, a prokaryotic cell, then the activation probability of an operator by a TF is given by

\[
P_T = \frac{1}{1 + \left( \frac{m}{m_1/2} \right)^n}.
\]

Here, \( m \) is the copy number of the TF in the cell and at \( m = m_{1/2} \) the occupation probability is one-half (the conversion to concentrations is trivial). The Hill coefficient (HC), \( n \), governs the steepness of the curve and is widely used to extract qualitative information about the regulation of genes from experimental data [11–14]. In the simplest cases, when there is no cooperative binding involved, one expects \( n = 1 \). In the presence of cooperative interactions, \( n \) is different than 1. For example, in the case of activation by dimers, one expects \( n = 2 \).

It is well known that the HC contains limited information [15, 16] and one cannot deduce all the stoichiometric details of the system from it. Nevertheless, the current assumption is that its deviation from 1 clearly indicates some sort of cooperativity. In this paper, we demonstrate that even this...
simple intuitive picture for the Hill curve can fail and a
disorder enhanced HC which is larger than 1 can appear
even in the absence of any cooperative binding of TFs to
the operator or to each other, apart from a steric repulsion
which prevents different copies of the TF to be bound to the
same location along the DNA. We show that this is a direct
consequence of a non-trivial combination of variations in the
binding energy of TFs to different sites along the DNA and
the steric repulsion between them. In addition to the increased
HC, this also leads to a dramatic increase in the occupation
probability of the cognate site as compared to a system with
no steric repulsion between the TFs or/and a constant non-
cognate binding energy. Importantly, we show that the results
presented here are essential for explaining the copy numbers
of TFs found in cells.

2. Results and analysis

The Hill curve, equation (1), is directly related to a formulation
of the problem using statistical mechanics and the knowledge
of the experimentally measured binding energy landscape of
the TF to the DNA. To illustrate this, we first focus on a
simple case where: (i) there is no cooperativity associated
with the structure of the TF or its binding properties, such that
one would naively expect \( n = 1 \). (ii) The probability of the
TF to be off the DNA or in a non-specific conformation on the
DNA is negligible. Note that by a non-specific conformation,
we mean one where the TF is on the DNA but does not
interact with the bases. This conformation, which typically
occurs due to electrostatic interactions, exists on any location
along the DNA, including the cognate site. The effects of
both simplifications are discussed in sections 3.1 and 3.2,
correspondingly.

Numbering the base pairs along the DNA, the location
of a TF on the DNA molecule, \( i \), is defined by the basepair
with the lowest index along the DNA among all the basepairs
bound by the protein. The steric repulsion is implemented by
restricting the occupation number of any location to 0 or 1\(^3\).

Under these assumptions, the problem can be formulated
as follows. For a cognate site with energy \( E_T \), the occupation
probability is given by

\[
P_T = \frac{1}{1 + e^{E_T - \mu}}, \tag{2}
\]

where throughout the paper we measure energy in units of \( k_B T \)
with \( k_B \) being the Boltzmann constant and \( T \) the temperature.
The chemical potential, \( \mu \), is set by the solution of the equation
\[\begin{align*}
e^{-F_{\mu} + \mu} + \sum_{i=0}^{N} \frac{1}{1 + e^{E_i - \mu}} &= m. \quad \tag{3}
\end{align*}\]

\(^3\) In fact, TFs are much larger than one basepair and, therefore, a TF bound
to a certain location on the DNA prevents binding of other proteins to several
locations (of the order of TF size in units of basepairs) with a higher index
along the DNA. This might make the effect of steric repulsion even more
important than illustrated here. Moreover, in our study we focus on the case of
high disorder in the energy landscape with several dominating energetic traps
along the DNA molecule. In this case it is highly non-probable for multiple
traps to be as close as the TF size, so that our definition the steric repulsion is
a good approximation.

Here, \( E_i \) is the binding energy of the TF at location \( i = 1, 2, \ldots, N \), with \( N \) being the number of accessible DNA
binding sites, on the DNA, presumably of the order of the genome length in units of basepairs. As stated above, different
copies of the TF exhibit steric repulsion, resulting in Fermi–
Dirac statistics \[19\] in equation (3). The first term in equation
(3) is the total copy number of TFs in the solution and in
nonspecific conformations on the DNA (with an associated
free energy \( F_{\mu} \)) \[20, 4, 5, 21\] and, for now, under our
assumption (ii) is negligible. In what follows it will be useful
to bear in mind that a non-negligible contribution can only reduce
the value of \( P_T \) and enhance \( m_{1/2} \). Its possible contribution is
discussed in section 3.2. We use the PRODORIC database for
PWMs of \( E. coli \) TFs \[22\], their cognate sequences and the
genomic sequence of \( E. coli \) of length \( N = 2 \times 4686077 \) \[23\].

To obtain the occupation probability of the cognate sites
of the analyzed TFs we solve numerically equation (3) for \( \mu \)
and find the occupation probability of the cognate site using
equation (2). In the following, we discuss the results of this
approach and show that it can lead to rather counterintuitive
Hill curves. We then give a simple theory which accounts for
the results analytically and gives precise conditions for these
effects to be important.

2.1. \( E. coli \) TFs Hill curves

First, we evaluate numerically the number of proteins needed
for the cognate site to be occupied with probability one-half,
\( m_{1/2} \), for all known cognate sites of three TFs. The results are
shown in figure 1 (for the moment focus on the data shown
as circles in panels (a), (b) and (c)). As can be seen (and
intuitively clear), the weaker the binding energy of the cognate
site, (larger \( E_T \)) the larger the \( m_{1/2} \). The values of \( m_{1/2} \) span
a large range for different TFs and for different cognate sites of
the same TF.

Next, in figures 2 and 3 ((a) panels), we present the binding
energy distribution on the \( E. coli \) genome (calculated from the
PWMs and the genome sequence) of two representative TFs:
LexA and RpoN. One can see that the distribution can be
well fitted by a Gaussian with a standard deviation, \( \sigma \) \[4, 5\].
As we show in section 2.2, a disorder-induced HC appears
for large values of \( \sigma \). We note that in some cases the fit of
the energy distribution to a Gaussian is not perfect, especially
at the tails (for example, due to several cognate sites). As
will be clear this can only have a quantitative effect on our
results by modifying the exact value of the HC but it does
not have a qualitative effect. In fact, as shown below, for
example for the RpoN TF, such effects can lead to an even
larger HC than the one predicted by the analytical calculation
which assumes a perfect Gaussian distribution of the binding
energies.

In panels (b) of figures 2 and 3, the occupation probability
of typical cognate sites (one cognate site of LexA and two of
RpoN) for these two TFs is presented. In addition, we fit the
results to a Hill curve (see panels (c)). With the value of \( m_{1/2} \)
given, the only fit parameter is the HC, \( n \). Surprisingly, for
the three cases (and others not shown), we obtain \( n > 1 \) (in
figures 2 and 3 we obtain \( n = 1.7, 8.3, 2.03 \)), despite the
Figure 1. The value of $m_{1/2}$ is presented as a function of the energy of different cognate sequences, $E_T$ for (a) LexA, (b) RpoN and (c) Lrp. The dotted lines are based on a numerical calculation using the *E. coli* genome. The circles represent energies calculated from all known cognate sites of the TF. The black solid lines are based on the freezing regime approximation, equation (16), while the red dashed lines are based on the non-steric approximation, equation (11). Filled, gray horizontal areas show the typical range of the TF’s copy number in *E. coli*. The symbol $m_c$ in (c) marks the crossover value of the Lrp TF between the uncrowded and crowded regimes, predicted by equations (9) and (14). (d) The HC, obtained by a fit of the numerical data to equation (1), is shown as a function of the cognate site energy for the LexA TF. The solid line is the analytical prediction, equation (17), while the circles represent numerical data based on real DNA and cognate sites sequence. The filled circle represents the HC of a hypothetical cognate site with a perfect consensus sequence. The dashed line represents the result of the non-steric approximation, equation (8), which gives $n = 1$.

Figure 2. Binding energy distribution and Hill curves of LexA. (a) A histogram of the binding energies of the LexA TF to the *E. coli* genome. The red points represent numerical data while the line is based on a Gaussian fit with $\sigma = 5.76$. The big filled circle marks the binding energy to the recQ operon sequence with $E_T = -20.6$. (b) The occupation probability of the LexA protein to the recQ operon as a function of $m$. The circles represent numerical data based on the PMW of the protein and the *E. coli* DNA sequence. The dashed line is based on the non-steric approximation, equation (6), while the solid line on equation (15). The solid arrow shows the crossover value (nonphysical in this case) of the protein copy number, $m_c = 0.65$, predicted by equations (9) and (14). (c) The same data as shown in panel (b) with a Hill function, equation (1), fit (thin line) to the numerical data (circles) with $n = 1.7$ and $m_{1/2} = 3000$. The gray areas in (b) and (c) show the typical range of the LexA copy number in *E. coli* (200–4000) [24].

absence of cooperativity in the model between the TF copies apart from steric repulsion. In figure 1(d), one can see that for a given protein the HC can be significantly higher than 1 for weak cognate sites (high values of the binding energy of a target, $E_T$) and decreases to 1 for stronger targets (low values of $E_T$). As shown below this occurs for many TFs.

To understand this, we first perform the same procedure, but ignoring the steric repulsion of the TFs everywhere apart from the cognate site, one obtains qualitatively and quantitatively distinct results (see non-steric approximation in figures 1, 2 and 3). This suggests that steric repulsion between different copies of a TF plays an important role in transcription regulation despite the fact that in most cases the ratio of the copy number of TFs in the cell to the length of the DNA is very small. In fact, ignoring steric repulsion leads to a significant increasing in the value of $m_{1/2}$ (in some cases much above
the measured estimate of the copy number of TFs in the cell). Moreover, it is clear that without disorder in the binding energy of the TF to different sites along the DNA, one would obtain \( n = 1 \) in the regime \( m \ll N \).

In sum, to account properly for the Hill curve of TFs binding to DNA one must (a) account for the disordered binding energy of the TF to different sites along the DNA and (b) account for steric repulsion between different copies of the TFs. Without these, the calculated HC would be lower and (b) account for steric repulsion between different copies of the TFs have steric repulsion only on the cognate site. We refer to this as the non-steric approximation. As shown in appendix, in this case the occupation probability of the cognate site is given by

\[
P^{\text{non-steric}}_\text{T} = \frac{1}{1 + \frac{N}{m} e^{E_T + \frac{m}{2}}},
\]

where \( E_T \) is the binding energy of the DNA subsequence of interest. Thus, by comparing with equation (1), one has

\[
m_{1/2} = N e^{E_T + \frac{m}{2}},
\]

\[
n = 1.
\]

In particular, this implies that without steric repulsion there is no enhanced HC and one obtains the naive results.

From the numerical experiment presented before, clearly, the non-steric approximation can fail. In addition to giving a wrong value for \( n \), it gives, in some cases, unreasonable values of \( m_{1/2} \). Indeed, in figures 2 and 3 we show numerical results for LexA compared with equation (6). The values of \( m_{1/2} \) differ by more than two orders of magnitude and as stated above the value of \( n \) is larger than 1. Moreover, the number of LexA TFs in the cell is estimated to be between 200 and 4000, a value much lower than the \( m_{1/2} = 10^3 \) found in the non-steric approximation. A similar disagreement between the numerical results and the non-steric approximation occurs for RpoN as well as for many other TFs.

To analyze the problem more carefully, taking into account steric repulsion, we calculate the occupation probability of the cognate site, solving equation (4), in three limiting cases (see the appendix).

---

**Figure 3.** Binding energy distribution and Hill curves of RpoN. (a) A histogram of the binding energy of the RpoN TF to the *E. coli* genome. The red points represent the numerical data while the line is based on a Gaussian fit with \( \sigma = 7.8 \). The circle shows the binding energy to the argT operon with \( E_T = -41.8 \). The square shows the binding energy to the fixABCX operon with \( E_T = -28.9 \).

(b) The occupation probability of the RpoN protein for the operons argT (circles and 1 symbol) and fixABCX (squares and 2 symbol) as a function of the TFs. Without these, the calculated HC would be lower and (b) account for steric repulsion between different copies of the TFs. Without these, the calculated HC would be lower.

(c) The same numerical data as in panel (b) with a fit of a Hill function, equation (1), (thin, solid, black lines) to the numerical data with \( n = 8.3 \) and \( m_{1/2} = 3.3 \) for the argT operon and with \( n = 2.07 \) and \( m_{1/2} = 1300 \) for the fixABCX operon. Vertical areas in panels (b) and (c) show the typical number of the RpoN proteins in *E. coli* (~110) [25].
Uncrowded regime. In this regime, illustrated in figure 4(a), the copy number of TFs is smaller than a crossover value defined by the DNA length and the disorder width:

\[ m \ll m_c = Ne^{-\frac{\sigma^2}{2}}. \]  

The occupation probability in this regime is given by the non-steric approximation, equation (6), so that

\[ P^\text{uncrowded}_T = \frac{1}{1 + \frac{N}{m} e^{E_T + \frac{\sigma^2}{2}}} = P^\text{non-steric}_T. \]  

We note that the disorder width, \( \sigma \), has different values for different TFs with typical values in the range of 2–8. Thus, the crossover value of the protein’s copy number varies from a non-physical \( 10^{-7} \) (where the non-steric approximation surely fails) to \( 10^9 \) where the non-steric approximation is expected to hold unless the copy number of TFs is extremely high.

The above derivation implies that the non-steric approximation (valid in the uncrowded regime) will give a good estimation of the HC and the number of particles at half occupation,

\[ m_{1/2} \equiv Ne^{E_T + \frac{\sigma^2}{2}}, \]  

only when \( m_{1/2} \ll m_c \). Furthermore, to be in the uncrowded regime near \( m \simeq m_{1/2} \) the cognate site of interest, with energy \( E_T \), has to satisfy the condition

\[ E_T \ll -\sigma^2 \].

We refer to cognate sites which satisfy condition (13) as strong.

Crowded regime. In this regime, illustrated in figure 4(b),

\[ m \gg m_c = Ne^{-\frac{\sigma^2}{2}} \]

and the occupation probability is given by

\[ P^\text{crowded}_T = \left( 1 + \frac{e^{E_T + \frac{\sigma^2}{2}} \sqrt{\frac{W}{\pi}} \left( \frac{\sigma}{\sqrt{2}} \right)^4}{W^4 N} \right)^{-1} \]  

\[ \gg P^\text{non-steric}_T, \]  

where \( W \) is the Lambert W-function [26]. If close to saturation of the cognate site, the system is in the crowded regime, \( m_{1/2} \gg m_c \). Then comparing equations (1) and (15) one can see that

\[ m_{1/2} \equiv Ne^{w_1 - E_T - \sigma^2} \]  

and

\[ n = \frac{2 + \sqrt{2\sigma^2 W \left( \frac{2}{\pi \sigma^2} \right)^4} - W \left[ \frac{2}{\pi \sigma^2} \right]^4}{\left[ 1 - W \left( \frac{2}{\pi \sigma^2} \right)^4 \right]^4} \]  

\[ \left( 1 + W \left( \frac{2}{\pi \sigma^2} \right)^4 \right)^4 \].

Note that to be in the crowded regime close to saturation of the cognate site, the condition

\[ -1 \gg E_T \gg -\sigma^2 \]  

has to be satisfied. We refer to cognate sites which satisfy condition (18) as weak.

Full-DNA regime. For the sake of completeness, we analyze here the irrelevant, full-DNA regime where the DNA is almost completely covered with TFs and for which

\[ m \gg m_c = Ne^{-\frac{1}{2\sigma^2}} \]  

and the occupation probability is given by

\[ P^\text{full-DNA}_T = \left( 1 + \frac{e^{E_T + \frac{1}{2\sigma^2}} \sqrt{\frac{2\ln \pi}{\sigma}}}{\sqrt{2\ln \pi}} \right)^{-1}. \]
When the typical number of the TFs in the cell is larger than the crowded regime, as demonstrated by equation (15).

Figure 5. A classification of cognate sites of different TFs. For each TF (y-axis) $E_T = -\langle\sigma^2\rangle$ is plotted for all its known cognate sites (taken from [22]). Empty rectangles represent a hypothetical cognate site’s energy that is half-occupied if the number of TFs is as specified in the database [27] (see the legend). For TFs with $m < m_c$, filled red areas represent the same hypothetical energy, as predicted by equations (11). When the typical number of the TFs in the cell is larger than $m_c$, the hypothetical cognate sites’ energies calculated from the non-steric approximation, equation (7), are presented by the red filled area while the results of applying equation (16) are presented by the blue filled areas.

If close to saturation of the cognate site, then the system is in the crowded regime, $m_{1/2} \gg m_c$. Then comparing equations (20) and (1), one can see that

$$m_{1/2} = Ne^{-\frac{\sigma^2}{2}}$$

$$n = \frac{1 + \sigma \sqrt{2 \ln \frac{N}{m} - 2 \ln \frac{N}{m}}}{2 \ln \frac{N}{m} \left(1 - \frac{1}{2} \frac{\sqrt{2 \ln \frac{N}{m}}}{m}\right)}.$$  \hspace{1cm} (22)

Note that to be in the full-DNA regime close to saturation of the cognate site the condition

$$E_T \gg -1$$

has to be satisfied. Such weak cognate sites do not seem to occur. In addition, the distribution of the binding energy for such high energies might deviate from the Gaussian one. Thus it is doubtful that the full-DNA regime results can be relevant even for, say, in vitro situation or for evaluating occupation probability of a non-cognate site.

2.3. Comparison of the analytical and numerical results

To check our analytical results, we plot both the non-steric approximation and the results from the crowded regime in figures 2 and 3. The non-steric approximation, equation (6), clearly fails. In contrast, equation (15) agrees well with the numerical data (strictly the uncrowded regime result gives a good approximation below $m_c = Ne^{-\frac{\sigma^2}{2}}$). In particular, the HC of the numerical data in the analyzed cases is clearly above 1 and depends on the disorder width and the cognate site’s energy, as demonstrated by equation (15).

It is interesting to note that close to saturation the Hill curve (determined by the value of $m_{1/2}$ and the steepness of $P_T$) is well described by equations (11), (12), (16) and (17), as presented in figure 1. However, in some cases there is a disagreement between the theory and the numerical results for small values of the protein copy number (see the small $m$ values in figures 3(b) and (c) and the low $E_T$ values in figures 1(b) and (d)). This disagreement is due to the deviation of the binding energy probability density from a normal distribution at the low energy tail (see figures 2(a) and 3(a)). This effect, which is easy to calculate numerically, increases both the value of $m_{1/2}$ and $n$ relative to the analytical predictions. As shown in figures 3(a) and (b) in some cases, this effect may be significant and lead to a very large HC ($n = 8.3$ for the argT operon of RpoN).

The analytical formulas, being based on the assumption of the Gaussian distribution of the binding energy, gives a different quantitative value for the HC in this case. In all the presented cases, the non-steric approximation clearly fails by several orders of magnitude.

2.4. Relevance to other cognate sites and other TFs

To examine the validity and relevance of these results to other TFs, we use a database of 17 known PWMs of TFs of *E. coli*, chosen randomly, and their known binding sites’ sequences [22]. Many of the TFs are present in large copy numbers, larger than their $m_c$ value (see the legend in figure 5). Therefore, the non-steric approximation of the occupation probability, (6), does not approximate well the occupation probability of many cognate sequences in the biologically relevant regime. In addition, as one can see in figure 5, the value of $E_T + \sigma^2$ for many cognate sites of many proteins is positive. As suggested by equation (18), to describe the occupation probability of such cognate sites, equations (10) and (15) have to be used, while the non-steric approximation, equation (6), fails.

It is instructive to calculate the chemical potential or the value of a hypothetical cognate site energy, denoted by $\mu$,
which is half-occupied when the number of the TF, \( m \), is as it is measured in experiments. The results are shown in figure 5. One can clearly see that the value of \( \mu \) is smaller (larger) than \( -\sigma^2 \) if \( m \) is smaller (larger) than \( m_c \), as suggested by conditions (13) and (18). Moreover, the non-steric approximation predicts well the value of \( \mu \) for \( m < m_c \), but fails for \( m > m_c \), predicting much lower values and, therefore predicting very small occupation probability for all the cognate sites with an energy above \( \mu \). In contrast, the expressions for the chemical potential calculated in appendix—equations (A.9) and (A.13)—predict well the location of the hypothetical, half-occupied energy for both weak and strong cognate sequences.

3. Inherent cooperativity and nonspecific states

The analysis in the paper relied on (i) no inherent cooperativity in the TF, and (ii) a negligible probability to be in non-specific states either on or off the DNA. We now turn to discuss the influence relieving these assumptions.

3.1. Effects of cooperativity

In our study, so far, we ignored the possibility of cooperative effects between distinct molecules of the TF. However, many TFs are active only in their \( \bar{n} \)-meric form. In this case, the HC is usually expected to be close to \( \bar{n} \) since the number of active molecules of the TF is proportional to the \( \bar{n} \)th power of the concentration of monomers [10]. One can reformulate the above derivation with \( n \) acting as the number of active, \( \bar{n} \)-meric copies of the TF. In this case, the HC according to the arguments given above is \( \bar{n} \cdot n \). Thus, a combination of cooperativity and quenched disorder can naturally lead to a high values of the HC.

3.2. The effect of nonspecific states

In our study, so far, we ignored nonspecific states of the protein. These states exist and correspond either to the TF being in the solution or in a nonspecific conformation bound to the DNA [20, 4, 21]. Namely, the nonspecific free energy is given by

\[
F_{ns} = \ln(e^{-E_{3D}} + N e^{-E_2}),
\]

where \( E_{3D} \) is the free energy of an unbound TF (modeled say by the free-energy of a solution of TFs) and \( E_2 \) is the energy of the bound protein in the nonspecific conformation which for simplicity we assume to have the same energy for all sites along the DNA (the results are unchanged, being proportional to \( N \), in the presence of small disorder in this binding energy with a slight reinterpretation of \( E_2 \)). Clearly, these nonspecific states can only reduce the occupation probabilities of the cognate sites calculated in equations (10) and (15). However, as suggested previously the existence of these nonspecific states can be an important component for the dynamics of the search and recognition process carried out by the TF [20, 4, 5, 28].

From equation (3), one can see that the nonspecific energy is negligible when

\[
F_{ns} \gg \mu + \ln m.
\]

Otherwise a significant fraction of the TFs are in the nonspecific state so that the chemical potential is well approximated by \( F_{ns} \). Thus, the occupation probability of the cognate site may be approximated by

\[
P_{T} \simeq \min \left( \frac{1}{e^{m_{n} \mu} + 1}, \frac{1}{e^{m_{n} \mu} + 1} \right),
\]

where \( \mu \) is given by equations (A.9) and (A.13). In figure 6 we show that indeed these results agree well with the numerical calculations for different values of the nonspecific energy. Thus, condition (25) implies that a half occupation of the cognate site occurs in the freezing regime (so that \( n > 1 \)) if

\[
F_{ns} \gg E_{T} + \ln m_{1/2}.
\]
eliminate the disorder-induced HC for all values of $m$ only if the condition $F_{ns} \ll \mu + \ln m_c$ is satisfied.

4. Discussion

The considerations discussed in this paper suggest the existence of a disorder-enhanced HC. They provide an estimate of the TF’s copy number needed to significantly occupy its cognate sites. This estimate is shown in many cases to be much smaller and more consistent with the existing data than a naive estimate, based on the non-steric approximation. Steric interactions between two competing TFs were already pointed out to lead to effective cooperativity previously in the case of an unusually large number of cognate binding sites [17]. In this study, we demonstrate that HC can be enhanced due to existence of high-affinity non-cognate binding sites along the DNA molecule. The results presented here are significantly different as they deal in a limit where naively one would expect different TFs to be independent of each other. The dependence arises due to the strong disorder of the binding energy landscape.

By analyzing the disordered statistical mechanics problem analytically, we show that three regimes are possible. In the uncrowded regime the number of TFs, $m$, is much smaller than the crossover value $m_c$, and steric repulsion can be ignored. In contrast, in the crowded regime, $m \gg m_c$, the steric repulsion play an important role and change dramatically the saturation curve, both quantitatively and qualitatively. The above results are summarized in figure 5 where we plot for several TFs and all their known cognate sites the value of $E_T + \sigma^2$. The regime $E_T + \sigma^2$ smaller/greater than zero corresponds to a regime without/with disorder-enhanced HC. In addition, we calculate using data on TF copy numbers the value of $E_T + \sigma^2$ which would yield $P_T = 1/2$ with the typical estimated copy number of TFs in E. coli. As can be seen, the numbers indicate that a disorder-induced HC is likely for a significant fraction of the TFs and their cognate sites. This happens since, as shown in the legend of figure 5, many TFs are present in numbers much larger than $m_c$ and, therefore, are located in the crowded regime. This study was performed using the measured PWMs of several TFs and can easily be extended to others.

Acknowledgments

We acknowledge useful comments from A Horovitz, R Voituriez, O Benichou, L Mirny, E Braun, A Sharma and M Depken. MS thanks FOM/NWO for financial support.

Appendix

In this appendix, we calculate the occupation probability of the target by a transcription factor in the different regimes discussed in the text. Assuming a pseudorandom DNA sequence implies that equation (3) is well approximated, for $N \gg 1$, by

$$\left\langle \frac{1}{1 + e^{E - \mu}} \right\rangle = \int_{-\infty}^{\infty} \frac{P_T(E)}{1 + e^{E - \mu}} dE = \frac{m}{N},$$  \hspace{1cm} (A.1)

where the binding energy probability density is given by equation (5) and the angular brackets are defined as an average over $E$ weighted by $P_T(E)$.

First we discuss the non-steric approximation. In this case, the Boltzmann statistics takes place of the Fermi–Dirac one everywhere except from the cognate site. Since occupation of

Figure 7. Nonspecific energy changes the Hill curve parameters. The parameters of the fitted Hill curve, equation (1), are presented for the occupation probability of the LexA TF to 1 of its cognate sequences, the recQ operon as a function of the nonspecific free energy. The circles represent $m_{c/2}$ while the squares represent $n$. The filled gray horizontal area shows the typical range of the LexA copy number in E. coli (200–4000) [24].

\[\text{Phys. Biol. 9 (2012) 056006} \hspace{1cm} \text{M Sheinman and Y Kafri}\]
the target site is much smaller than \( m \), it can be neglected and the constraint on the chemical potential is given by
\[
\langle e^{-E + \mu} \rangle = e^{\frac{\mu}{m}} = \frac{m}{N}. \tag{A.2}
\]
Substituting the solution for \( \mu \) in equation (2), one gets the non-steric approximation
\[
\rho_{T}^{\text{non-steric}} = \frac{1}{1 + \frac{N}{m} e^{E_{s} + \frac{\mu}{m}}}. \tag{A.3}
\]

As is shown in the main text the non-steric approximation clearly fails. We now turn to evaluate the integral in equation (4) using a saddle-point approximation. First, we evaluate the integral in equation (A.1) assuming that the saddle point of the integrand is known and then find the equation for the saddle point. In this way instead of one integral equation (A.1) for the chemical potential, one gets two coupled algebraic equations for the chemical potential and the saddle point of the integrand.

Performing this procedure, the resulting set of equations are
\[
\mu = E_{s} + \ln \left( -\frac{\sigma^{2}}{E_{s}} - 1 \right) \tag{A.4}
\]
and
\[
\frac{e^{-\sigma^{2}/E_{s}}}{\sigma \sqrt{\sigma^{2} - E_{s}} - E_{s} \sigma^{2}} = \frac{m}{N}, \tag{A.5}
\]
where \( E_{s} \) is the value of \( E \) at the saddle point. Equation (A.5) may be solved self-consistently in three limiting cases:

**Uncrowded regime.** In this regime the saddle point occurs at \( E_{s} \approx -\sigma^{2} \), so that equation (A.5) implies
\[
E_{s} = -\sigma^{2} \left( 1 - \frac{m}{N} e^{\sigma^{2}/2} \right). \tag{A.6}
\]
Self-consistency in this regime requires
\[
m \ll m_{c}, \tag{A.7}
\]
where
\[
m_{c} = N e^{-\frac{\sigma^{2}}{2}}. \tag{A.8}
\]
The chemical potential (A.4) in this limit is given by
\[
\mu = -\frac{\sigma^{2}}{2} - \ln \left( \frac{N}{m} \right) \tag{A.9}
\]
and the occupation probability is given by equation (10).

**Crowded regime.** In this regime the saddle point satisfies \(-\sigma^{2} \ll E_{s} \ll -1 \) (recall that in our units \( k_{B}T = 1 \)) and equation (A.5) implies
\[
E_{s} = -\frac{\sigma}{\sqrt{2}} \sqrt{W \left[ \frac{2}{\sigma^{2}} \left( \frac{N}{m} \right)^{4} \right]}, \tag{A.10}
\]
where \( W \) is the Lambert \( W \)-function [26]. Self-consistency requires
\[
m_{s} \gg m \gg m_{c}, \tag{A.11}
\]
where
\[
m_{c} = N e^{-\frac{1}{\sqrt{2} \sigma^{2}}}. \tag{A.12}
\]
The chemical potential in this limit is then given by
\[
\mu = -\frac{\sigma}{\sqrt{2}} \sqrt{W \left[ \frac{2}{\sigma^{2}} \left( \frac{N}{m} \right)^{4} \right]} + \ln \left\{ \frac{\sqrt{2} \sigma}{\sqrt{W \left[ \frac{2}{\sigma^{2}} \left( \frac{N}{m} \right)^{4} \right]}} - 1 \right\}. \tag{A.13}
\]
and the occupation probability by equation (15).

**Full-DNA regime.** In this regime, the saddle point satisfies \( E_{s} \gg -1 \) and equation (A.5) implies
\[
E_{s} = -\sigma \sqrt{2 \ln \frac{N}{m}}. \tag{A.14}
\]
Self-consistency requires
\[
m \gg m_{s}. \tag{A.15}
\]
The chemical potential in this limit, is then given by
\[
\mu = -\sigma \sqrt{2 \ln \frac{N}{m} + \ln \left( \frac{\sigma}{\sqrt{2 \ln \frac{N}{m}}} - 1 \right)}. \tag{A.16}
\]
This regime is irrelevant in our study, since the copy number of any TF in the cell is much smaller than the DNA length.

**References**

[1] Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson J D 1994 The Molecular Biology of the Cell 4th edn (New York: Garland)
[2] Perelson A S and Weisbuch G 1997 Immunology for physicists Rev. Mod. Phys. 69 1219
[3] Zhang J, Maslov S and Shakhnovich E I 2008 Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size Mol. Syst. Biol. 4 1
[4] Gerland U, Moroz J D and Hwa T 2002 Physical constraints and functional characteristics of transcription factor—DNA interaction Proc. Natl Acad. Sci. USA 99 12015
[5] Slutsky M and Mirny L A 2004 Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential Biophys. J. 87 4021
[6] Zakrzewska K and Lavery R 2012 Towards a molecular view of transcriptional control Curr. Opin. Struct. Biol. 22 160
[7] Wang J, Lu J, Gu G and Liu Y 2011 In vitro DNA-binding profile of transcription factors: methods and new insights J. Endocrinol. 210 15
[8] Berg O G and von Hippel P H 1987 Selection of DNA binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters J. Mol. Biol. 193 723
[9] Stormo G D and Fields D S 1998 Specificity, free energy and information content in protein–DNA interactions Trends Biochem. Sci. 23 109
[10] Phillips R, Konev J and Therio J 2009 Physical Biology of the Cell (New York: Taylor and Francis)
[11] Buchler N E, Gerland U and Hwa T 2005 Nonlinear protein degradation and the function of genetic circuits Proc. Natl Acad. Sci. USA 102 9559–64
[12] Kuhlman T, Zhang Z, Saier M H and Hwa T 2007 Combinatorial transcriptional control of the lactose operon of Escherichia coli Proc. Natl Acad. Sci. USA 104 6043
[13] Kim H D, Shai T, O’Shea E K and Regev A 2009 Transcriptional regulatory circuits: predicting numbers from alphabets Science 325 429
[14] Garcia H G, Sanchez A, Kuhlman T, Kondev J and Phillips R 2010 Transcription by the numbers redux: experiments and calculations that surprise *Trends Cell. Biol.* 20 7233

[15] Weiss J N 1997 The Hill equation revisited: uses and misuses *FASEB J.* 11 835

[16] Holt J M and Ackers G K 2009 The Hill coefficient: Inadequate resolution of cooperativity in human hemoglobin *Biothermodynamics, Part A (Methods in Enzymology* vol 455) ed M L Johnson, J M Holt and G K Ackers (New York: Academic) chapter 7 pp 193

[17] Werner M, Zhu L and Aurell E 2007 Cooperative action in eukaryotic gene regulation: physical properties of a viral example *Phys. Rev. E* 76 061909

[18] von Hippel P H and Berg O G 1986 On the specificity of protein interactions *Proc. Natl Acad. Sci. USA* 83 1608

[19] Landau L D and Lifshitz E M 1980 *Statistical Physics* (Oxford: Pergamon)

[20] Berg O G, Winter R B and von Hippel P H 1981 Diffusion-driven mechanisms of protein translocation on nucleic acids: I. Models and theory *Biochemistry* 20 6929

[21] Kalodimos C G, Biris N, Bonvin A M J J, Levandoski M M, Guennuegues M, Boelens R and Kaptén R 2004 Structure and flexibility adaptation in nonspecific and specific protein–DNA complexes *Science* 305 586

[22] Munch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E and Jahn D 2003 PRōDORIC: prokaryotic database of gene regulation *Nucleic Acids Res.* 31 266

[23] Blattner F R et al 1997 The complete genome sequence of *Escherichia coli* K-12 *Science* 277 1453

[24] Dri A-M and Moreau P L 1994 Control of the LexA regulon by pH: evidence for a reversible inactivation of the LexA repressor during the growth cycle of *Escherichia coli* *Mol. Microbiol.* 12 621

[25] Jishage M, Iwata A, Ueda S and Ishihama A 1996 Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions *J. Bacteriol.* 178 5447

[26] Abramowitz M and Stegun I A 1970 *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (New York: Dover)

[27] Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl F U, Kerner M and Frishman D 2008 Protein abundance profiling of the *Escherichia coli* cytosol *BMC Genomics* 9 102

[28] Sheinman M, Bénichou O, Kafri Y and Voituriez R 2011 Classes of fast and specific search mechanisms for proteins on DNA *Rep. Prog. Phys.* 75 026601

[29] Derrida B 1980 Random-energy model: limit of a family of disordered models *Phys. Rev. Lett.* 45 79

[30] Li G-W, Berg O G and Elf J 2009 Effects of macromolecular crowding and DNA looping on gene regulation kinetics *Nature Phys.* 5 294