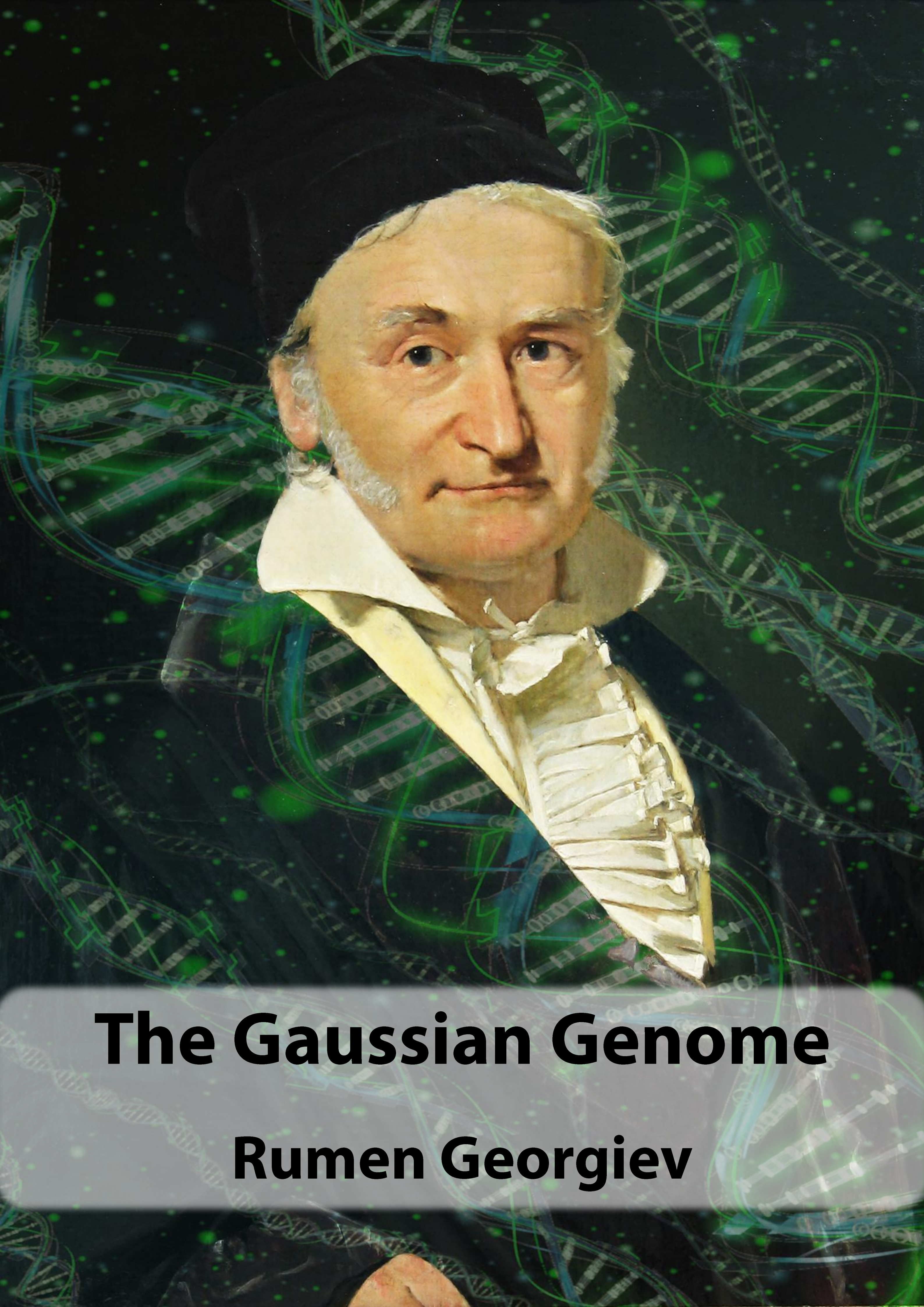


The Gaussian Genome

Rumen Georgiev

Cover design: Rumen Georgiev

Images used:

Jensen, Christian Albrecht.
Carl Friedrich Gauß. 1840.
Oil on canvas.
Archiv der Berlin-Brandenburgische
Akademie der Wissenschaften

Pringle, Jimmy.
"DNA." 2013.
L0kust's deviantART gallery.
deviantART. 2016. Web.
March 22, 2016

Universiteit Utrecht

Department of Chemistry

van't Hoff Laboratorium

The Gaussian Genome

Author:
Rumen Georgiev

Supervisors:
Jasper Landman
prof. Willem Kegel

May 16, 2016

Abstract

Recent studies in biophysics suggest transcription factor interactions with non-regulatory DNA are sequence-dependent and vary along the DNA strand. This makes numerical calculation of the grand canonical partition function cumbersome and renders predictions of genetic activity a seemingly insurmountable task. Using the cumulant-generating function of the normal distribution, we derive the partition function and define an effective binding energy, a single quantity which accounts for the contributions from the whole spectrum of binding energies. Applying our approach to LacI and RNAP, two prominent *lac* operon transcription factors, we obtain theoretical results which are in good accord with the actual biophysical picture, namely, the standard deviations of their binding energy distributions, on one hand, and binding mode differentiation on the other.

Table of Contents

Chapter 1

Introduction

Since its discovery in 1959 by Jacob and Monod [1], the *lac* operon, responsible for the the lactose metabolism in the cell, has been subjected to extensive studies due to the variety of possible regulatory scenarios which can take place within it. As an operon, it contains three genes controlled by a single promoter sequence P, *i.e.*, a single sequence of nucleobases, which signals RNA polymerase (RNAP) it should start synthesizing mRNA by reading out the genetic information encoded downstream from P. Control over the transcription process is carried out by two *transcription factors* – the *lac* repressor (LacI), which inhibits transcription by effectively blocking part of the promoter sequence, and the cAMP receptor protein (CRP, also known as catabolite activator protein, CAP), which activates transcription by forming contacts with RNAP. Each of these three proteins recognizes (at least) one specific sequence (*site*) in the *lac* operon and, upon binding to it, affects transcription. Due to the specific properties of these DNA-protein complexes, it is only natural to dub these functional sequences *specific sites.* RNAP has one specific site, the promoter, LacI binds specifically to the operator sequence O_1 and to the two auxiliary operators O_2 and O_3, while CRP's specific site is located close to the promoter sequence [2].

As we can see, even in this short part of DNA which is the *lac* operon, there are a total of 5 different specific sites for 3 regulatory proteins and binding to these sites is on average 5 to 15 $k_{\mathrm{B}}T$ stronger compared to any other part of the DNA strand [3, 4]. Since we called these tight-binding functional sites specific, patches of non-regulatory DNA, which do not play a crucial role in transcription control, shall be called *non-specific sites.* Here we should clarify that a protein does not discern between another protein's specific site and non-specific DNA. The repressor, for example, will bind with roughly equal affinity to a non-operator patch of DNA and part of the promoter sequence. The exact function of the non-specific sites has been debated for some time but the scientific community agrees that proteins use them to quickly find their respective specific site [5–8]. This is achieved by combining 3D-diffusion with 1D-diffusion along the DNA strand and jumps between spatially adjacent patches of DNA brought close together due to DNA supercoiling. Another question these sites raise is the distribution of their binding energies – are they uniformly distributed, *i.e.*, do they have constant energy, or do they have a broad distribution with a wide spectrum of binding energies. Recent work in the field [9–11] points to the latter and, since non-specific sites are considered as having a random sequence, a Gaussian distribution is to be expected. But calculating the partition function for this case is a seemingly insurmountable task because now, instead of having only one group of sites, we are faced with thousands even millions of groups all having different energies and numbers of sites.

The main goal of the current work is to derive a theory for the binding of proteins to non-specific sites under the grand canonical ensemble and check its validity in the context of the *lac* operon. We begin our study by providing the theoretical background in statistical mechanics, biophysics, statistics and tensor calculus needed in all our further considerations. Modelling the binding energy distribution under the grand canonical ensemble with the help of statistical generating functions leads us to the definition an effective energy for the whole distribution. In other words, the partition functions for a distribution of choice and a reservoir of non-specific sites with equal binding energy are identical as long as we take into account the shape of the distribution and scale the constant energy properly. We are interested not only in deriving a simple equation, which accurately predicts the behaviour of the system, but under what conditions our model breaks down, as well. To that end we derive a convergence criterion, which estimates the validity range of our model. Testing our model with *in vitro* experimental data yields interesting and reasonable results, which we try to link to biophysical quantities. Later on, however, when we expand our consideration to *in vivo* measurements, by obtaining the binding energy distribution of the whole *E. coli* genome via energy matrices [12], we are faced with a seemingly inexplicable inconsistency. To resolve the controversy, we conceive the idea of *binding modes*, that is, structural and chemical differences between specific and non-specific complexes. Finally, we conclude our work by summarising all results we have obtained and discussing what future research should focus on to fully resolve the issue at hand.

Chapter 2

Theoretical background

This chapter's aim is to provide the reader with an overview of all theoretical facts and knowledge needed further in our study. We begin our work by describing the system we are modelling, that is, the *lac* operon, by recalling its structure and function, listing all proteins involved in lactose metabolism control and briefly discussing the possible regulatory scenarios. Next, we plunge into statistical mechanics and discuss the two most prominent approaches in *lac* operon activity modelling – the canonical and grand canonical ensembles. We derive the occupation probability of a promoter site under both ensembles, compare the results and analyse them. Our next topic of interest is biophysics, in which we focus on the link between the binding energy of a protein to DNA and the salt conditions within a cell. To that end we review Manning's polyelectrolyte theory and derive $K^{\text{obs}}\left(\left[\text{M}^{+}\right]\right)$ by following Record *et al.*'s approach [13]. We conclude this theoretical background overview with a discussion of the mathematical and statistical tools we will be using in all further derivations.

2.1 The *lac* operon

In an ever-changing environment cells must adapt to the external conditions like temperature, nutrition source, salt concentration etc. in order to survive. This process of readjustment is carried out by proteins within the cell, which are synthesised *ad hoc* using messenger RNA or mRNA. The mRNA itself is created by a protein called RNA polymerase which, by sliding along the DNA of a cell, reads out (*transcribes*) the genetic information encoded in the form of nucleobases. This integrated process, from transcription to protein synthesis, is referred to as genetic expression. Being crucial to the cell's proper functioning, the expression of a given gene is heavily regulated by a variety of proteins capable of adsorbing onto DNA, thus hindering or enhancing transcription. This type of control is commonly known as *genetic regulation.*

In this work we deal with one such regulatory system, the *lac* operon, which is a patch of DNA found in many bacteria, *E. coli* among others, and is responsible for the transport and metabolism of lactose. As an operon, it has one promoter sequence followed by more than one (in this case three) genes. When RNA polymerase (RNAP) binds to the promoter it begins to read out the nucleobase sequence and synthesize mRNA. What sets operons apart from normal genes is the fact that several genes share a common promoter and their information is transcribed onto a single mRNA strand. For the *lac* operon these genes are:

1. LacZ, which carries the genetic information for β-galactosidase, an enzyme which splits the disaccharide lactose into glucose and galactose, which can then be metabolised.

2. LacY is the gene responsible for the production of lactose permease, a membrane protein facilitating lactose intake by the cell.

3. LacA, which encodes galactoside O-acetyltransferase, an enzyme which catalyses the deacetylation of acetyl-coenzyme A by β-D-galactoside and is involved in the detoxification of the cell.

As we already mentioned, production of these proteins is governed by RNAP, but the actual control over the transcription process is carried out by two other proteins, called transcription factors (TFs) – the *lac* repressor (LacI) and the cyclic adenosine monophosphate (cAMP) receptor protein (also known under various names like CRP and CAP). The repressor, as the name suggests, applies negative control over the transcription process by physically blocking RNAP from binding to the promoter. CRP, on the other hand, is an activator, which binds to DNA near the promoter and eases RNAP's binding to the promoter via direct protein-protein contacts with RNAP. Both TFs are thoroughly studied [14–16], so we will not go in much detail regarding their structures and just mention they are both dimeric proteins exhibiting two-fold symmetry, which ultimately affects the base sequence of their respective adsorption sites – CRP binds to a fully symmetric consensus site, while LacI's preferred binding site, the operator O_1, is pseudosymmetric. Both sequences are given below:

CRP consensus	A A A T G T G A T C T ‖ A G A T C A C A T T T
	T T T A C A C T A G A ‖ T C T A G T G T A A A
LacI O_1	A A T T G T G A G C **G** G A T A A C A A T T
	T T A A C A C T C G **C** C T A T T G T T A A

The two vertical lines in the CRP consensus site signify the axis of two-fold symmetry. Bases coloured in red for LacI's O_1 are the ones that break the symmetry of the site. The CRP consensus site is 22 base pairs (bp) in length, therefore the axis of symmetry goes in between the 11^{th} and 12^{th} bp. O_1, being 21 bp long, has a central base pair (given in bold), which lies on the axis of pseudosymmetry.

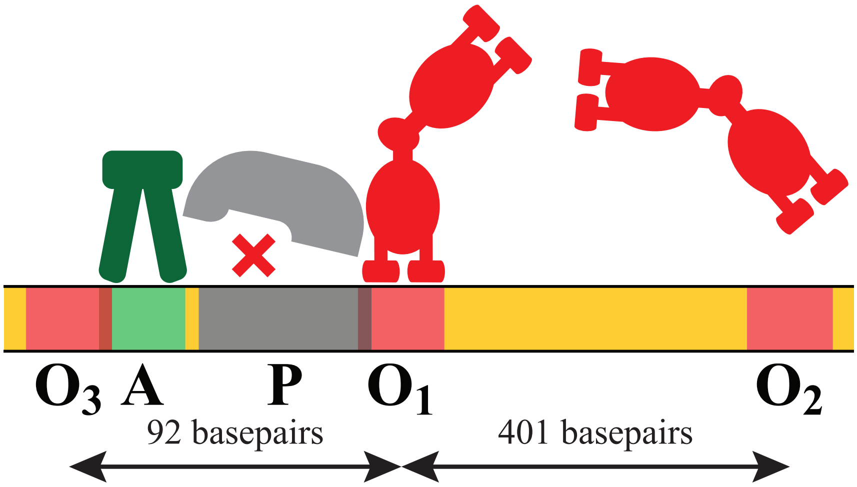


Figure 2.1: A sketch of the *lac* operon in the case of repression. LacI and the operator sites (O_1, O_2, O_3) are coloured in red, the promoter sequence P and RNAP are grey and the activator site A along with CRP are green.

Here we should point out that while CRP and RNAP have only one specific site within the *lac* operon, the repressor actually has three – we already mentioned the operator which is aided in repression by two auxiliary sites, O_2 and O_3, positioned 401 bases downstream and 92 bases upstream from the promoter, respectively. A sketch of the *lac* operon is presented in Fig 2.1. For clarity all three operator sequences are listed below and we have coloured in green the common base pairs (this time, to avoid confusion, we have dropped the complementary strands).

LacI O_1	A A T T G T G A G C G G A T A A C A A T T
LacI O_2	A A A T G T G A G C G A G T A A C A A C C
LacI O_3	G G C A G T G A G C G C A A C G C A A T T

The last molecule which affects transcription is the *lac* inducer – a sugar (allolactose) or a sugar derivative (IPTG) which binds to the repressor and changes its conformation, effectively leading to a massive reduction of its affinity for the operator sequences.

Now that we have introduced all species taking part in transcription regulation, we briefly discuss the genetic regulation mechanism which is influenced by cellular conditions.

We begin by assuming LacI occupies the operator site and hinders RNAP's binding to the promoter, which leads to very low levels of expression of the three Lac genes and lactose is hardly metabolised, leading to its accumulation. At a certain point its concentration is high enough and it can be partially transglycosylated by β-galactosidase, a process which yields the allosteric inducer of the *lac* operon, allolactose. It binds to the repressor leading to LacI's dissociation from the operator site and, due to the lack of hindering, RNAP can now begin transcribing the three Lac genes. The synthesised mRNA is used by ribosomes as a template for Lac protein synthesis. The proteins metabolise the accumulated lactose and when its concentration drops below a certain threshold allolactose desorbes from the repressor, which regains its high affinity for the operator and interferes with further transcription. CRP's function in the entire process is to increase the transcription level of LacZ when glucose levels are low.

As we now see, the *lac* operon is a complex system which renders prediction of its behaviour a complicated task. Furthermore, this is only one example of a genetic regulation architecture, and we should keep in mind there are thousands of other genes encoded on DNA. In that sense, the main goal of computational biology, *i.e.*, obtaining quantitative information for a biological system via generalised models applicable to many different systems, seems impossible. There are, however, methods with which genetic activity can be predicted. Although we shall discuss them only in the context of the *lac* operon, one should keep in mind they are applicable to many, if not all, genetic regulation scenarios.

2.2 Genetic statistical mechanics

After discussing how the *lac* operon functions, we now review how it is modelled in the context of both the canonical and grand canonical ensembles. After obtaining the partition functions under both frameworks, we draw parallels between them and see they yield identical results. Through this comparison we gain insight into the nature of the fugacity, a quantity which will prove crucial in all further considerations.

2.2.1 Canonical (NVT) ensemble

The statistical mechanical ensemble most commonly employed in modelling genetic activity is the canonical or NVT ensemble ([17–20] and references therein). Under this framework a constraint over the number of proteins in our system is applied, namely, there is a total of P protein molecules on the DNA strands, of which p are bound specifically and $P-p$ are adsorbed onto non-specific sites. After this clarification, we are now ready to derive the canonical partition function. For simplicity we shall consider the case of simple transcription, *i.e*, one RNAP molecule binds specifically to one promoter sequence – after all, this section serves only as an illustration of the statistical mechanical apparatus applied later on. For further generalized studies on the topic, one should refer to [21].

To begin our derivation, we recall the Boltzmann weights of the two possible binding situations: $\exp(-p\beta\epsilon_{\mathrm{s}})$ and $\exp(-(P-p)\beta\epsilon_{\mathrm{ns}})$, where $\beta = (k_{\mathrm{B}}T)^{-1}$, k_{B} is the Boltzmann constant, T is the absolute temperature in K, ϵ_i is the binding energy, the subscripts signify specific and non-specific binding, P is the RNAP copy number and p is the number of RNAP molecules occupying the specific site. Next, we need to realize that before promoter binding takes place, we have P molecules bound non-specifically and N_{total} sites, which can accommodate them. The number of possible combinations in which this can be achieved is given by the binomial coefficient:

$$\binom{N_{\mathrm{total}}}{P} = \frac{N_{\mathrm{total}}}{P!(N_{\mathrm{total}}-P)!} \tag{2.1}$$

The statistical weight of this purely non-specific binding case is given by the Boltzmann weight multiplied by the number of possible ways of realizing it:

$$Z_{\mathrm{ns}}(P) = \binom{N_{\mathrm{total}}}{P} \exp(-P\beta\epsilon_{\mathrm{ns}}) \tag{2.2}$$

In a similar fashion, we can define the statistical weight of specific binding, as well:

$$Z_{\mathrm{s}}(p) = \binom{1}{p} \exp(-p\beta\epsilon_{\mathrm{s}}) \Rightarrow Z_{\mathrm{s}}(p=0) = 1, \tag{2.3}$$

where p takes integer values of either 0 or 1. Since specific and non-specific binding are decoupled safe for the total number of RNAP molecules constraint, the statistical weight of this initial state is given by the product of the separate statistical weights:

$$Z_{\mathrm{state}}(p=0) = Z_{\mathrm{ns}}(P)Z_{\mathrm{s}}(p=0) = Z_{\mathrm{ns}}(P) \tag{2.4}$$

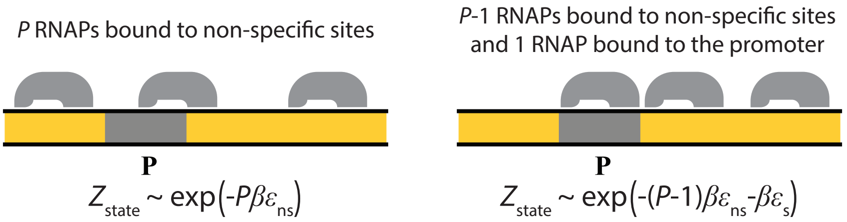

P RNAPs bound to non-specific sites
P-1 RNAPs bound to non-specific sites and 1 RNAP bound to the promoter
P
$Z_{\mathrm{state}} \sim \exp(-P\beta\varepsilon_{\mathrm{ns}})$
P
$Z_{\mathrm{state}} \sim \exp(-(P-1)\beta\varepsilon_{\mathrm{ns}}-\beta\varepsilon_{\mathrm{s}})$

Figure 2.2: Sketches of the two possible binding states of P number of RNAP molecules: (*left*) All RNAPs are bound non-specifically, therefore the statistical weight of the state scales exponentially with the number of molecules and their binding energy. (*right*) One RNAP occupies the promoter sequence and all other (P-1) are bound non-specifically.

Upon specific binding, the number of RNAP molecules bound to non-specific DNA is reduced by one, which leads to a change in the binomial coefficient:

$$\begin{aligned}\binom{N_{\mathrm{total}}}{P-1} &= \frac{N_{\mathrm{total}}!}{(P-1)!(N_{\mathrm{total}}-P+1)!} = \frac{P}{(N_{\mathrm{total}}-P+1)}\frac{N_{\mathrm{total}}!}{P!(N_{\mathrm{total}}-P)!} \\ &\simeq \binom{N_{\mathrm{total}}}{P}\frac{P}{N_{\mathrm{total}}} \qquad \text{when } N_{\mathrm{total}} \gg P-1\end{aligned} \tag{2.5}$$

We obtain the statistical weight of this state in the same way as before while taking into account the reduced number of RNAP molecules bound to non-specific sites and the occupation of the promoter sequence:

$$\begin{aligned} Z_{\text{state}}(p=1) &= Z_{\text{ns}}(P-1)Z_{\text{s}}(p=1) \\ &= \binom{N_{\text{total}}}{P} \frac{P}{N_{\text{total}}} \exp\left(-\beta\left((P-1)\epsilon_{\text{ns}}+\epsilon_{\text{s}}\right)\right) \\ &= Z_{\text{state}}(p=0)\frac{P}{N_{\text{total}}} \exp\left(-\beta\Delta\epsilon\right), \end{aligned} \tag{2.6}$$

where in the last step we define the energy difference between specific and non-specific binding as $\epsilon_{\text{s}} - \epsilon_{\text{ns}} = \Delta\epsilon$ and we realize we can express the statistical weight of the occupied state as a function of $Z_{\text{state}}(p=0)$. Both states and their statistical weights are depicted in Fig 2.2.

If we now want to obtain the probability of having the promoter sequence occupied, we need to calculate the ratio of $Z_{\text{state}}(p=1)$ to the sum of all statistical weights $Z_{\text{system}} = Z_{\text{state}}(p=0) + Z_{\text{state}}(p=1)$:

$$\begin{aligned} p_{\text{occupied}} = \frac{Z_{\text{state}}(p=1)}{Z_{\text{system}}} &= \frac{Z_{\text{state}}(p=0)P \times N_{\text{total}}^{-1} \exp\left(-\beta\Delta\epsilon\right)}{Z_{\text{state}}(p=0)\left(1+P \times N_{\text{total}}^{-1}\exp\left(-\beta\Delta\epsilon\right)\right)} \\ &= \frac{P \times N_{\text{total}}^{-1} \exp\left(-\beta\Delta\epsilon\right)}{1+P \times N_{\text{total}}^{-1}\exp\left(-\beta\Delta\epsilon\right)} \end{aligned} \tag{2.7}$$

We now compare this result to the well-known Langmuir isotherm [22], describing localised adsorption of gas molecules on a solid lattice:

$$\theta_{\text{A}} = \frac{p_{\text{A}} \times K_{\text{ads}}}{1+p_{\text{A}} \times K_{\text{ads}}}, \tag{2.8}$$

and draw three interesting parallels:

1. The coverage fraction θ_{A} corresponds to the occupation probability of the promoter site
2. Both the pressure p_{A} and the RNAP copy number per non-specific site $P \times N_{\text{total}}^{-1}$ act as effective concentration of the adsorbing species
3. The adsorption constant K_{ads} can be linked to the adsorption free energy via the fundamental equation $\Delta G_{\text{ads}} = -k_{\text{B}}T\ln K_{\text{ads}}$, which makes K_{ads} identical with the exponential factors in Eq. 2.7.

Thus, we arrive at the conclusion that calculating the probability of finding the promoter site occupied by an RNAP molecule can be reduced to a Langmuir adsorption type of problem. This fact should not come as a surprise to us, since the system at hand represents a lattice of well-defined sites. The main difference is the dimensionality – while Langmuir derived his seminal equation for surfaces, we are looking at a one-dimensional string of adsorption sites, which in no way changes the physics behind Langmuir's isotherm.

2.2.2 Grand canonical (μVT) ensemble

After discussing the canonical ensemble, we turn to the grand canonical ensemble (also commonly referred to as μVT), which has recently been employed by Weinert *et al.* [23]

to substitute the NVT ensemble as a tool for genetic regulation modelling. Much like its canonical counterpart, the grand canonical ensemble's alternative name reflects the constraints which are applied to the system. In contrast to the canonical ensemble, under the grand canonical it is the chemical potential of the adsorbing species μ that is kept constant, rather than their total number. That being said, we can decouple the gene from the non-specific sites. Since we are interested only in the partition function of the gene, and not the whole genome, the problem is no longer complicated by the combinatorics we were forced to use under the canonical ensemble. Then the statistical weight of a state is simply given by its Boltzmann weight, in which the specific binding energy is offset by the non-specific energy. The Boltzmann weight itself is scaled by a factor, exponentially dependent on the chemical potential: $\mathrm{e}^{\beta p\mu}\mathrm{e}^{-\beta p(\epsilon_{\mathrm{s}}-\epsilon_{\mathrm{ns}})}$. Summing over all possible states, that is, over all possible occupation numbers, yields the grand canonical partition function Ξ:

$$\Xi = \sum_{p=0}^{1} \mathrm{e}^{\beta p\mu}\mathrm{e}^{-\beta p(\epsilon_{\mathrm{s}}-\epsilon_{\mathrm{ns}})} = \sum_{p=0}^{1} \lambda^{p}\mathrm{e}^{-\beta p\Delta\epsilon} = 1 + \lambda\mathrm{e}^{-\beta\Delta\epsilon} \tag{2.9}$$

Here, as in the previous section, we define the number of RNAP molecules occupying the promoter sequence as p and the specific-non-specific binding difference as $\Delta\epsilon$. The new quantity we encounter is the *fugacity* $\lambda = \exp(\beta\mu)$, which has a physical meaning of activity of the adsorbing protein. Obtaining the promoter occupation probability requires the definition of the grand potential, $\Phi = -k_{\mathrm{B}}T\ln\Xi$, taking its partial derivative with respect to the chemical potential and dividing by the total number of promoter sites. Since our system contains only one promoter, the derivative of the potential actually yields the occupation probability:

$$\begin{aligned} p_{\text{occupied}} &= -\frac{\partial\Phi}{\partial\mu} = \beta k_{\mathrm{B}}T\frac{\partial\ln\Xi}{\partial\ln\lambda} = \frac{\lambda}{\Xi}\frac{\partial\Xi}{\partial\lambda} \\ &= \frac{\lambda}{\Xi}\exp(-\beta\Delta\epsilon) \\ &= \frac{\lambda\exp(-\beta\Delta\epsilon)}{1+\lambda\exp(-\beta\Delta\epsilon)} \end{aligned} \tag{2.10}$$

where we used the definition of the fugacity.

Comparing the last lines of Eq. 2.7 and 2.10 points to an important fact, first conceived by Weinert *et al.*[23] – the expressions for the occupation probability, derived under both ensembles, have the Langmuir isotherm functional form, which means the fugacity of the adsorbing protein acts as an effective one-dimensional concentration and is approximately equal to the average number of proteins adsorbed per nucleobase.

Although yielding analogous results, the two approaches differ significantly in the ease with which one obtains the occupation probability. While the canonical ensemble requires the usage of tedious combinatorics and calculating the statistical weight for each state separately, under the grand canonical ensemble p_{occupied} follows directly from the grand canonical partition function function, given by the simple expression:

$$\Xi = \sum_{p=0}^{1} \lambda^{p}\exp(-\beta p\Delta\epsilon) \tag{2.11}$$

Having said that, now is the proper time to discuss one binding parameter, which is often overlooked, *i.e.*, the energy difference between specific and non-specific binding, $\Delta\epsilon$. Commonly in literature non-specific sites are deemed as a reservoir of adsorption sites with constant energy, which is set to zero. Recently, however, developments in the field of genetics have suggested otherwise – non-regulatory DNA has a binding energy distribution, which is close to a Gaussian [9, 10]. This poses a major problem to our considerations in both ensembles – there are thousands of non-specific reservoirs each having a different number of sites and a separate binding energy. Tackling this problem usually involves assuming only the sites around the peak of the distribution have a significant contribution to the partition function and centring it around zero. As we shall see in Chapter 3, this may often lead to contradictory results, especially if one is dealing with distributions with a standard deviation above $\frac{1}{2}k_\mathrm{B}T$. Having noted this, we can now clearly outline the goals of our study:

1. Derivation of the grand partition function of a one dimensional lattice of adsorption sites, obeying a statistical distribution

2. Calculation of distribution parameters from *in vivo* and *in vitro* binding experiments

3. Attempt to derive a theory for non-specific binding of regulatory proteins to DNA.

2.3 Biophysical conditions within the cell

The regulation of genes in *E. coli* relies on the interactions between the double-stranded co-polyelectrolyte DNA and a group of proteins, capable of adsorbing onto it. These interactions can be divided into two major groups: sequence specific which can vary depending on the precise order of nucleotides in the DNA strand, and sequence non-specific, which are constant throughout the entire strand. Hydrophobic interactions of apolar parts of the adsorbing protein with DNA and hydrogen bonding of nucleobases of DNA to side groups in the protein are the two most prominent specific interactions. On the other hand the non-specific interactions are mainly attributed to the electrostatic attraction between the negatively charged phosphate groups forming the DNA backbone and the cations in the active center of the adsorbing protein.

A main goal of quantitative biology is using the protein-DNA interactions as levers with which genetic regulations can be tailored. While sequence specific interactions are not easily adjustable (chemical modifications as methylation of nucleobases is needed, for example), one can fine-tune the electrostatics by screening the charges via the ionic strength of the solution used, thus altering the net DNA-protein interaction energy, which leads to a change in the observed equilibrium constant.

In this section we are mainly interested in the binding equilibrium constants K^obs (also known as affinity constant) obtained from *in vitro* measurements under conditions close to those observed *in vivo*. To that end, we present the theoretical background for the dependence of K^obs on ionic concentration by deriving $K^\mathrm{obs}\left(\left[\mathrm{M}^+\right]\right)$. We do that by essentially following the approximations made by Record *et al.* [13] while using a more straightforward and simple derivation (Appendix B).

To set the stage for it, we will first make an overview of the electrostatic interaction present on the DNA strand under Manning's polyelectrolyte theory [24]. In its framework DNA is seen as a string of negative charges distributed homogeneously along the length of the strand (Fig 2.3). Due to charge interactions between the phosphate groups along the strand, sodium ions condense from the solution onto the DNA strand. For this system we can define a characteristic dimensionless scale ξ_b given by the ratio of the average axial charge separation b to the Bjerrum length (λ_B, the distance at which the electrostatic interaction between two charges becomes equal to the thermal energy).

$$\xi_\mathrm{b} = \frac{b}{\lambda_\mathrm{B}} = b\frac{4\pi\varepsilon\varepsilon_0 k_\mathrm{B}T}{e^2}, \tag{2.12}$$

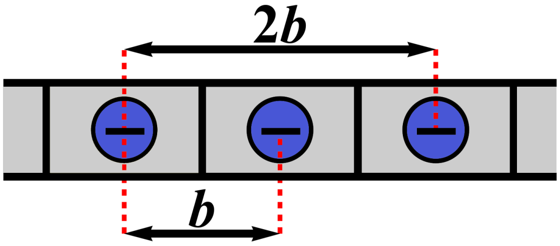


Figure 2.3: DNA strand modelled as 1D lattice of equally spaced negative charges with axial charge separation of $b = 1.7$ Å

where e is the elementary charge, ε_0 is the permittivity of vacuum and ε is the relative permittivity of the medium. Although fairly intuitive, this definition is not the one commonly used in biophysics literature and $\xi_{\rm el} = \xi_{\rm b}^{-1} = \lambda_{\rm B}/b$ is usually employed, instead. Plugging in the numbers for the system under consideration, *i.e.*, double-stranded DNA in water at 310 K, we obtain $\xi_{\rm el} = 4.27$. If we now multiply $\xi_{\rm el}$ by $k_{\rm B}T$ we obtain the energy of interaction of two adjacent phosphate groups on the DNA strand. According to the polyelectrolyte theory, charge condensation occurs when $\xi_{\rm el} > 1$ and the number of condensed counter-ions per charge on the polymer, *i.e.*, per phosphate group, is given by:

$$\theta = \frac{N_{\rm Na^+}}{N} = 1 - \xi_{\rm el}^{-1} = 0.77, \tag{2.13}$$

from which we can conclude that the effective charge of a phosphate group is:

$$q_{\rm eff} = (1-\theta)e = e/\xi_{\rm el} \tag{2.14}$$

We realise we should also include divalent ions like Mg^{2+} in our considerations, since they can be found within the cell and can affect the electrostatics in our system. Furthermore, they can adsorb to DNA and compete with proteins for DNA binding. We choose to neglect all these effects because most experimental studies reported in literature use monovalent ions as a main electrostatics-tuning ions and divalent ions are added only in minute quantities.

It is instructive to derive an expression for the electrostatic interactions between phosphate groups, which we will then use to obtain $K^{\rm obs}\left(\left[\mathrm{M}^+\right]\right)$. To that end, we will follow Manning's approach and assume a Debye-Hückel screened Coulomb interaction potential of the form $q^2 r^{-1}\exp(-\kappa r)$ and sum over all charge pairs, where κ is the Debye-Hückel screening parameter:

$$\kappa = \sqrt{8\pi N_{\rm A} I \lambda_{\rm B}} \tag{2.15}$$

Here, I is the ionic strength of the solution in moles per litre, $N_{\rm A}$ is Avogardo's number and $\lambda_{\rm B}$ is expressed in metres. Before we start deriving the pair potential, we recall that the phosphate groups are equally spaced on the DNA strand and we can express the distance r_{ij} between any two as $r_{ij} = |ib - jb|$, where ib and jb are the distances from an arbitrary point on the DNA strand and i and j are integers. Thus, we can write down the total pair interaction as:

$$U = \frac{1}{2}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{i=1}^{N}\frac{\exp(-\kappa|i-j|b)}{|i-j|b} = \frac{1}{2b}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{i=1}^{N}\frac{\exp(-\tilde{\kappa}|i-j|)}{|i-j|}, \tag{2.16}$$

where $\tilde{\kappa} = \kappa b$ and we divide by 2 to take into account double counting.

In order to tackle this sum, we make the substitution $|i-j| = k$. Therefore, we have two cases for k:

$$i - j = \begin{cases} -k & \Rightarrow \; j = i + k \\ k & \Rightarrow \; j = i - k, \end{cases} \tag{2.17}$$

which means the substitution splits the double sum into two double sums Σ1 and Σ2:

$$\Sigma_1 = \frac{1}{2b}\sum_{k=1}^{N-1}\sum_{i=k+1}^{N}\frac{\exp(-\tilde{\kappa}k)}{k} = \frac{1}{2b}\sum_{k=1}^{N-1}\frac{(N-k)\exp(-\tilde{\kappa}k)}{k} \qquad \text{when } j < i \tag{2.18}$$

and

$$\Sigma_2 = \frac{1}{2b}\sum_{k=1}^{N-1}\sum_{i=1}^{N-k}\frac{\exp(-\tilde{\kappa}k)}{k} = \frac{1}{2b}\sum_{k=1}^{N-1}\frac{(N-k)\exp(-\tilde{\kappa}k)}{k} \qquad \text{when } j > i \tag{2.19}$$

From these two equations it is evident that Σ_1 and Σ_2 are identical and we can express the pair potential by adding up Σ_1 and Σ_2 and splitting the resulting single sum into two contributions, S_1 and S_2:

$$U = \frac{1}{b}\sum_{k=1}^{N-1}\frac{(N-k)\exp(-\tilde{\kappa})^k}{k} = S_1 + S_2, \tag{2.20}$$

where

$$S_1 = \frac{N}{b}\sum_{k=1}^{N-1}\frac{\exp(-\tilde{\kappa})^k}{k} = -\frac{N}{b}\ln[1-\exp(-\tilde{\kappa})] \tag{2.21}$$

and

$$S_2 = -\frac{1}{b}\sum_{k=1}^{N-1}\exp(-\tilde{\kappa})^k = -\frac{1}{b}\frac{\exp(-\tilde{\kappa})}{1-\exp(-\tilde{\kappa})} = -\frac{1}{b}\frac{1}{\exp(\tilde{\kappa})-1} \tag{2.22}$$

Summing in S_1 and S_2 is only allowed if $\exp(-\tilde{\kappa}) < 1$ and $N \to \infty$. Since we are dealing with polymers ($N \gg 1$), we can make the approximation $N \to \infty$. On the other hand, for double-stranded DNA under physiological conditions, *i.e.*, 0.2 M NaCl solution, $\kappa = 1.48$ nm^{-1} and $b = 0.17$ nm, which leads to $\exp(-\tilde{\kappa}) = 0.78 < 1$. That being said, we realize we can drop the second sum since it is on the order of b^{-1}, while S_1 scales linearly with the length of the DNA – $S_1 \propto Nb^{-1}$. Therefore, we arrive at an approximate expression for the pair potential:

$$U = S_1 + S_2 \approx -\frac{N}{b}\ln[1-\exp(-\tilde{\kappa})] \approx -\frac{N}{b}\ln[1-1+\tilde{\kappa}+\mathcal{O}\left(\tilde{\kappa}^2\right)] \approx -\frac{N}{b}\ln(\tilde{\kappa}) \tag{2.23}$$

In the second to last step we expand the exponent within the logarithm and retain only the first two terms. We must keep in mind, though, this approximation overestimates the actual pair potential by 8%.

To obtain the interaction energy we multiply the pair potential U by the Coulomb factor and we keep in mind we are using effective charges (Eq. 2.14):

$$G^{\text{el}} = -\frac{N}{b}\ln(\tilde{\kappa})\frac{q_{\text{eff}}^2}{4\pi\epsilon\epsilon_0} = -\frac{N}{b}\ln(\tilde{\kappa})\frac{1}{4\pi\epsilon\epsilon_0}\underbrace{\frac{e^2}{\xi_{\text{el}}^2}}_{q_{\text{eff}}^2}\frac{k_{\text{B}}T}{k_{\text{B}}T} = -Nk_{\text{B}}T\frac{\ln(\tilde{\kappa})}{\xi_{\text{el}}} \tag{2.24}$$

Since $\tilde{\kappa} < 1$, $G^{\text{el}} > 0$, a fact which we expect since two negatively charged phosphate groups are repulsing each other. The attentive reader may argue the final result is questionable since G^{el} increases linearly with charge separation, while, according to Coulomb's law, it should go down linearly. This is due to the assumption made in Eq. 2.13 – the number of ions condensing on DNA decreases linearly with distance, which leads to a linear increase of the effective charge with b.

Now that we have derived an expression for the electrostatic interactions on the DNA strand, we turn our attention to the binding equilibrium of a regulatory protein to DNA (Fig 2.4):

$$\mathrm{D} + \mathrm{TF} \overset{K^{\text{obs}}}{\leftrightarrows} \mathrm{DTF}, \tag{2.25}$$

where D is a binding site (specific or non-specific) on DNA, TF is the transcription factor, and DTF is the complex formed. From this equilibrium we can express the observed affinity constant K^{obs}:

$$K^{\text{obs}} = \frac{[\mathrm{DTF}]}{[\mathrm{D}][\mathrm{TF}]} \tag{2.26}$$

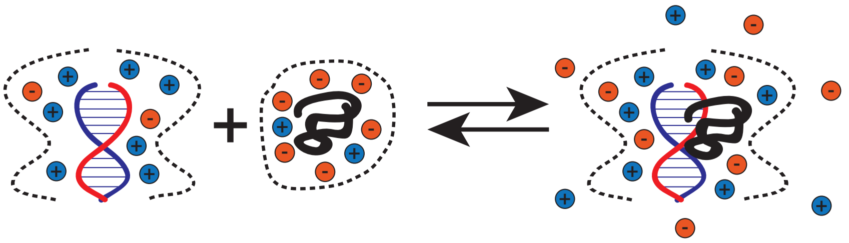


Figure 2.4: Sketch of DNA-protein binding. Both DNA and the adsorbing protein have counter-ions condensed onto them, a fraction of which are released into the bulk solution upon protein binding.

Although simple, this equation is not of much use in this form since it can not explain the strong dependence of K^{obs} on ionic strength – upon increasing the salt concentration, binding becomes weaker. From this well-established experimental fact and recalling Le Châtelier's principle we can deduce counter-ions are released from the DNA upon binding. Of course, anion release from the active center of TF and hydration effects might also play a major role, but for small and simple anions like Cl^- and dilute solutions (concentrations used are on the order of nM), experiments have shown that these are irrelevant [25, 26]. Furthermore, in this form the equation gives little to no information regarding the relationship between the energy of binding and the observed affinity constant. Our analysis of this relationship is even further complicated by the fact that the constant is not dimensionless which prevents us from applying $\Delta G = -k_{\mathrm{B}}T \ln K$ directly.

To resolve the latter issue, we first define the total number of binding sites as D_0. Then $[\mathrm{DTF}] = \langle \mathrm{TF} \rangle / V$ and $[\mathrm{D}] = (\mathrm{D}_0 - \langle \mathrm{TF} \rangle)/V$, where V is the volume of the system and $\langle \mathrm{TF} \rangle$ signifies the average number of transcription factors bound to the DNA strand (further in the text we will often use this notation for averaging). We obtain the average number of adsorbed transcription factors from the grand partition function:

$$\langle \mathrm{TF} \rangle = \lambda \frac{\partial \ln \Xi}{\partial \lambda} = \mathrm{D}_0 \frac{\lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \mathrm{e}^{-\beta\epsilon}}, \tag{2.27}$$

where, as usual, $\lambda = \mathrm{e}^{\beta\mu_{\mathrm{TF}}}$ is the fugacity of the transcription factor and ϵ is the binding energy of TF to D. Substituting Eq. 2.27 in the definitions of [DTF] and [D], we can write down their ratio as:

$$\frac{[\mathrm{DTF}]}{[\mathrm{D}]} = \frac{\lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \mathrm{e}^{-\beta\epsilon}} \left(1 - \frac{\lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \mathrm{e}^{-\beta\epsilon}} \right)^{-1} = \lambda \mathrm{e}^{-\beta\epsilon} \tag{2.28}$$

On the other hand, for sufficiently dilute solutions ($n_{\mathrm{TF}} \ll n_{\mathrm{w}}$), the molar concentration is proportional to the molar fraction (here v_{w} denotes the molar volume of water in L/mol):

$$[\mathrm{TF}] = \frac{n_{\mathrm{TF}}}{V} = \frac{1}{v_{\mathrm{w}}}\frac{n_{\mathrm{TF}}}{n_{\mathrm{w}}} \approx \frac{x_{\mathrm{TF}}}{v_{\mathrm{w}}} \tag{2.29}$$

The molar fraction of the transcription factor is linked to its fugacity via the relationship:

$$\mu = \mu^{\ominus} + k_{\mathrm{B}}T \ln x \tag{2.30}$$

Re-writing Eq. 2.30, we obtain $x_{\mathrm{TF}} = \lambda \exp\left(-\beta\mu^{\ominus}_{\mathrm{TF}}\right)$, which, combined with Eq. 2.28 and Eq. 2.29, yields:

$$K^{\mathrm{obs}} = v_{\mathrm{w}} \exp\left(\beta\mu^{\ominus}_{\mathrm{TF}} - \beta\epsilon\right) \tag{2.31}$$

Since we are free to decide what the standard state for TF is, we choose it in such a way that $\mu^{\ominus}_{\mathrm{TF}} = 0$:

$$K^{\mathrm{obs}} = v_{\mathrm{w}} \exp\left(-\beta\epsilon\right) \tag{2.32}$$

Physically, this means we assume the transcription factor is bound to the DNA and additional energy is needed to make it desorb from the strand. This, is in fact, true: due to their low solubility transcription factors are always bound to DNA. Thus, we obtain a simple expression for the observed equilibrium constant, which is proportional to the molar volume of water and scales exponentially with the energy of binding.

Let us now ponder what kind of interactions take part in the binding energy. As we already said, the affinity constants are highly sensitive to ionic strength conditions, which means a major part of the binding energy is due to electrostatics and entropy gain from counter-ion release. On the other hand since both the adsorbing protein and DNA contain apolar groups, there should be a hydrophobic contribution, as well. Last, but not least, we should consider the formation of hydrogen bonds upon binding.

The dependence of K^{obs} on the binding energy, though insightful, does not answer our main question, that is, how the observed equilibrium constant depends on salt conditions. Obtaining this relation requires us to assume there is another participant in the reaction in Eq. 2.25, namely, the sodium cations, released from the DNA strand:

$$\mathrm{D} + \mathrm{TF} \overset{K^{\ominus}_{\mathrm{T}}}{\leftrightarrows} \mathrm{DTF} + \nu\mathrm{M}^{+}, \tag{2.33}$$

where ν is the number of cations released upon protein binding. That being said, we can write down the intrinsic equilibrium constant of the process:

$$K^{\ominus}_{\mathrm{T}} = \frac{a_{\mathrm{DTF}}\ a^{\nu}_{\mathrm{M}^{+}}}{a_{\mathrm{D}}\ a_{\mathrm{TF}}}, \tag{2.34}$$

To provide a link between $K^{\ominus}_{\mathrm{T}}$ and the observed constant, we remember the relationship between the activity and the concentration for dilute solutions:

$$a_{\mathrm{A}} = \gamma_{\mathrm{A}} x_{\mathrm{A}} = \gamma_{\mathrm{A}}\frac{n_{\mathrm{A}}}{n_{\mathrm{total}}} \simeq \gamma_{\mathrm{A}}\frac{n_{\mathrm{A}}}{n_{\mathrm{w}}} = \gamma_{\mathrm{A}}\frac{n_{\mathrm{A}}}{c_{\mathrm{w}}V} = \gamma_{\mathrm{A}}[\mathrm{A}]v_{\mathrm{w}}, \tag{2.35}$$

where c_{w} is the molar concentration of water. Expressing concentrations in Eq. 2.26 in terms of activities and multiplying both the numerator and the denominator by $a^{\nu}_{\mathrm{M}^{+}}$ yields:

$$K^{\mathrm{obs}} = \frac{a_{\mathrm{DTF}}v^{2}_{\mathrm{w}}}{a_{\mathrm{D}}a_{\mathrm{TF}}v_{\mathrm{w}}}\frac{a^{\nu}_{\mathrm{M}^{+}}}{a^{\nu}_{\mathrm{M}^{+}}}\frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} = v_{\mathrm{w}}K^{\ominus}_{\mathrm{T}}a^{-\nu}_{\mathrm{M}^{+}}\frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} \tag{2.36}$$

Rearranging Eq. 2.36 and taking the natural logarithm on both sides of the equation yields:

$$\ln(v_{\mathrm{w}}^{-1}K^{\mathrm{obs}}) = \ln K_{\mathrm{T}}^{\ominus} + \ln a_{\mathrm{M}^+}^{-\nu} + \ln \frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} \tag{2.37}$$

If we now compare Eq. 2.32 and 2.37, it becomes evident that each of the three terms on the right-hand side of Eq. 2.37 represents a contribution to the "observed" change of free energy of the system:

1. $\ln K_{\mathrm{T}}^{\ominus}$ is the binding energy under standard conditions
2. $\ln a_{\mathrm{M}^+}^{-\nu}$ is a concentration term, which depends on the ionic environment and reflects the change in solution conditions upon binding
3. $\ln \frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}}$ accounts for the non-ideality of the interacting macromolecules and how this departure from ideal behaviour changes in the process of binding.

This reasoning follows from the fundamental equation $\Delta G = k_{\mathrm{B}}T \ln K$ and the definitions of the activity a and the activity coefficient γ.

In order to obtain an explicit relationship between K^{obs} and $[\mathrm{M}^+]$ we follow the approach of Record *et al.*[13] and make 2 assumptions:

1. Binding of TF to DNA does not give rise to a net change of charge on the strand – a Z-charged macrocation (TF) binds tightly, effectively neutralizing Z charges from the DNA strand. It may be argued that this leads to a net charge flux of Z/ξ positive charges condensing onto the DNA strand but we must keep in mind that TF actually also has counter-ions (X^-) condensed at the active site, which are released upon binding and do not condense on the formed complex.
2. Charge density in both the native DNA and the complex is uniform and equal. Although binding of certain proteins like CRP to DNA introduce a bend in the strand and there are transcription factors like the *lac* repressor which induce DNA-looping ([14, 27, 28] and references therein), these effects are local and span no more than 500 base pairs in 5×10^6 base pair long prokaryotic DNA.

With these two assumption we derive the linear dependence of $\ln K^{\mathrm{obs}}$ on the logarithm of the salt concentration. The full derivation can be found in Appendix B while here we only present the final result:

$$\ln(v_{\mathrm{w}}^{-1}K^{\mathrm{obs}}) = \ln K_{\mathrm{T}}^{0} + Z\frac{\ln\left(\delta\gamma_{\mathrm{M}^+}\right)}{\xi} - Z\ln\overline{[\mathrm{M}^+]}\left(1 - \frac{3}{2\xi}\right), \tag{2.38}$$

where $\delta = b\sqrt{8\pi N_{\mathrm{A}}\lambda_{\mathrm{B}}c^{\ominus}}$ and $\overline{[\mathrm{M}^+]} = [\mathrm{M}^+]/c^{\ominus}$ is the molar concentration of cations divided by the standard molar concentration $c^{\ominus}$ of 1 M.

The expression we arrive at is very similar to the one derived by Record *et al.* but with one major difference – the concentration (third) term in Eq. 2.38 is proportional to $1 - 3 \times (2\xi)^{-1}$, contradicting the result of the aforementioned researchers who derive a dependence of $1 - 1 \times (2\xi)^{-1}$. For double-stranded DNA under physiological conditions the difference between the two factors may not seem striking (according to Record *et al.* the value should be 0.88 while we obtain 0.65) but we need to keep in mind one important feature of Eq. 2.38 – the concentration term scales linearly with the number of counter-ions released upon binding, *i.e.*, using Record's result we overestimate Z by roughly one third. Despite this discrepancy, we shall adopt Record *et al.*'s result, because it is widely used in the literature and is supported by experimental measurements.

Calculating the functional dependence of K^{obs} on cation concentration requires we take the total derivative of $\ln K^{\mathrm{obs}}$ with respect to $\ln \overline{[\mathrm{M}^+]}$:

$$\frac{\mathrm{d}\ln K^{\mathrm{obs}}}{\mathrm{d}\ln \overline{[\mathrm{M}^+]}} = -Z\left(1-\frac{1}{2\xi}\right) + \frac{Z}{\xi}\frac{\mathrm{d}\ln \gamma_{\mathrm{M}^+}}{\mathrm{d}\ln \overline{[\mathrm{M}^+]}} \approx -0.88Z \tag{2.39}$$

For sufficiently low salt concentrations the derivative term is negligible and can be dropped out. Thus, we conclude that $\ln K^{\mathrm{obs}}$ depends linearly on $\ln \overline{[\mathrm{M}^+]}$ provided there is no significant dependence of the activity of ions on their concentration (*i.e.* this is only valid for dilute solutions).

2.4 Genetic Mathematics

In this section we give a brief overview of the mathematical background needed to obtain the main theoretical results, presented in Chapter 3. We start by introducing the idea of the energy matrix and showing how one can calculate the binding energy of a site from its nucleobase sequence using vector and tensor products. Next, we argue what the distribution of the binding energies of a DNA strand should be and arrive at the conclusion a normal distribution is to be expected. Finally, we recall what the moment-generating function (MGF) and a cumulant-generating function (CGF) are and derive the MGF for the normal and Laplace distributions as examples.

2.4.1 Energy matrices and binding energy of a site

The need to predict the genetic activity *in vivo* has lead to the idea of the *energy matrix* **E** – a mathematical concept which assigns an additive energy contribution to each nucleobase within a specific binding site and which makes theoretical calculation of its binding energy possible. A similar idea was initially introduced by G. D. Stormo *et al.* in 1982 [29] but the position weight matrices (PWMs) these researchers proposed evaluated the information carried by a nucleobase at a given position for the recognition of the site as specific. Recently, combining modern high-throughput sequencing and cell sorting techniques, Kinney *et al.* [30] have managed to further develop the concept of the PWM and evolve it into the energy matrix (Fig. 2.5).

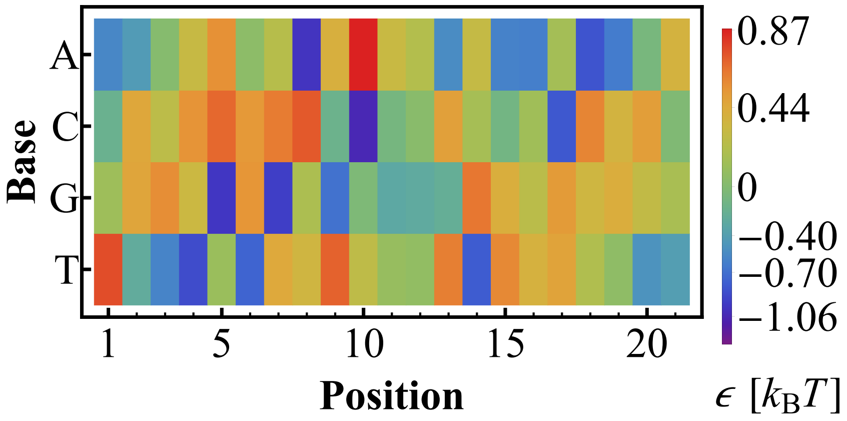


Figure 2.5: Energy matrix **E** of the *lac* repressor as obtained by Razo-Mejia *et al.* [12]. The matrix gives the energy contribution of each possible base at each position of the binding site. The binding energy is obtained by summing all contributions.

To illustrate how one can theoretically predict the binding energy of a site, we make use of the *lac* repressor energy matrix found by Razo-Mejia *et al.* [12]:

$$\mathbf{E} = \begin{pmatrix} \epsilon_{1,\mathrm{A}} & \epsilon_{1,\mathrm{C}} & \epsilon_{1,\mathrm{G}} & \epsilon_{1,\mathrm{T}} \\ \epsilon_{2,\mathrm{A}} & \cdot & \cdot & \epsilon_{2,\mathrm{T}} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \epsilon_{21,\mathrm{A}} & \epsilon_{21,\mathrm{C}} & \epsilon_{21,\mathrm{G}} & \epsilon_{21,\mathrm{T}} \end{pmatrix}_{21\times 4}, \tag{2.40}$$

and define a generic binding site represented by the 21-dimensional sequence vector **s**:

$$\mathbf{s} = (\mathrm{GAT...CA})_{21} \tag{2.41}$$

In both **E** and **s** A, C, G, and T signify the 4 nucleobases comprising DNA. The sequence vector **s** contains one of the four nucleobases at each position, according to the DNA sequence of the site of interest. Each element of the matrix, on the other hand, tells us what is the energy contribution at a given position provided we have a specific base there. For instance, the element on row 2, column 4 $\epsilon_{2,\mathrm{T}}$ informs us what the energy gain (or loss) is when the *lac* repressor binds to DNA and needs to interact with thymine at position 2 of the binding site.

Our next step is to define a base vector **b** – a vector of length 4, which contains all four different nucleobases, one for each position of the vector. The structure of the energy matrix defines the base order of **b**, that is, since the first column of **E** represents A, the second – C, etc. the base vector must be:

$$\mathbf{b} = (\mathrm{ACGT}), \tag{2.42}$$

Calculating the binding energy of the site requires the introduction of the sequence matrix **S**, a logical matrix which has one element per row equal to unity and all other elements in the row are 0. The position of the unity element tells us what nucleobase we have at the given position. **S** is defined as the tensor product of the sequence vector **s** and the base vector **b**. In order to obtain a logical matrix from this tensor product we define the linear operator $\hat{\delta}$ as (here I and J both signify one of the 4 nucleobases):

$$\hat{\delta}(\mathrm{IJ}) = \begin{cases} 1 & \text{if } \mathrm{I} = \mathrm{J} \\ 0 & \text{if } \mathrm{I} \neq \mathrm{J} \end{cases} \tag{2.43}$$

and apply it to each entry of the obtained tensor product. Therefore,

$$\mathbf{S} = \hat{\delta}\ \mathbf{b} \otimes \mathbf{s} = \begin{pmatrix} \hat{\delta}(\mathrm{GA}) & \hat{\delta}(\mathrm{GC}) & \hat{\delta}(\mathrm{GG}) & \hat{\delta}(\mathrm{GT}) \\ \hat{\delta}(\mathrm{AA}) & \hat{\delta}(\mathrm{AC}) & \hat{\delta}(\mathrm{AG}) & \hat{\delta}(\mathrm{AT}) \\ \hat{\delta}(\mathrm{TA}) & \hat{\delta}(\mathrm{TC}) & \hat{\delta}(\mathrm{TG}) & \hat{\delta}(\mathrm{TT}) \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \hat{\delta}(\mathrm{CA}) & \hat{\delta}(\mathrm{CC}) & \hat{\delta}(\mathrm{CG}) & \hat{\delta}(\mathrm{CT}) \\ \hat{\delta}(\mathrm{AA}) & \hat{\delta}(\mathrm{AC}) & \hat{\delta}(\mathrm{AG}) & \hat{\delta}(\mathrm{AT}) \end{pmatrix}_{21\times 4} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}_{21\times 4} \tag{2.44}$$

If we now take the double inner product of **E** and **S**, we obtain the total interaction energy:

$$\mathbf{E} : \mathbf{S} = \epsilon_{1,\mathrm{G}} + \epsilon_{2,\mathrm{A}} + \epsilon_{3,\mathrm{T}} + ... + \epsilon_{20,\mathrm{C}} + \epsilon_{21,\mathrm{A}} = \epsilon_{\mathrm{site}} \tag{2.45}$$

2.4.2 Distribution of binding energies

As we already saw in the previous section, the binding energy of a site (be it specific or non-specific) can be modelled as a sum of independent contributions, which are assigned to each nucleobase in the sequence of the site. It is now of interest to see what should the energy distribution of several thousand such sites be.

One formulation of the central limit theorem states that the sum of a number of independent and identically distributed random variables with finite variances tends to a normal distribution as the number of variables grows. For a generic non-specific site the probability for each nucleobase to occur at a given position is 25%, that is all possible outcomes of our random process are identically distributed. On the other hand, since the presence of A at position 3, for example, does not influence the nucleobases present at any other position, we can say that the variables are independent, as well. That being said, the application of the central limit theorem to our system seems straightforward.

We must, however, ponder the question how large should the number of variables be. It is commonly accepted that a sum of 31 and more random variables can be deemed normally-distributed [31]. If we are considering RNAP which occupies 41 bases, then we can assume its binding energies to NS DNA obey a normal distribution without hesitation. On the other hand, for transcription factors with shorter binding sites, like the *lac* repressor (21 bases) and CRP (22 bases), this assumption may not hold and one may need to resort to Student's t distribution to estimate the mean and standard deviation of the distribution. If we, however, compare t distributions with 20 and 30 degrees of freedom (the degrees of freedom are defined as $L-1$, where L is the number of bases in the site) to a standardised normal distribution, we hardly see any differences – most notably the tails of the t distributions are slightly heavier and the maxima are lower compared to the normal distribution (Fig. 2.6). Therefore, we shall approximate the energy distributions of these proteins to a Gaussian.

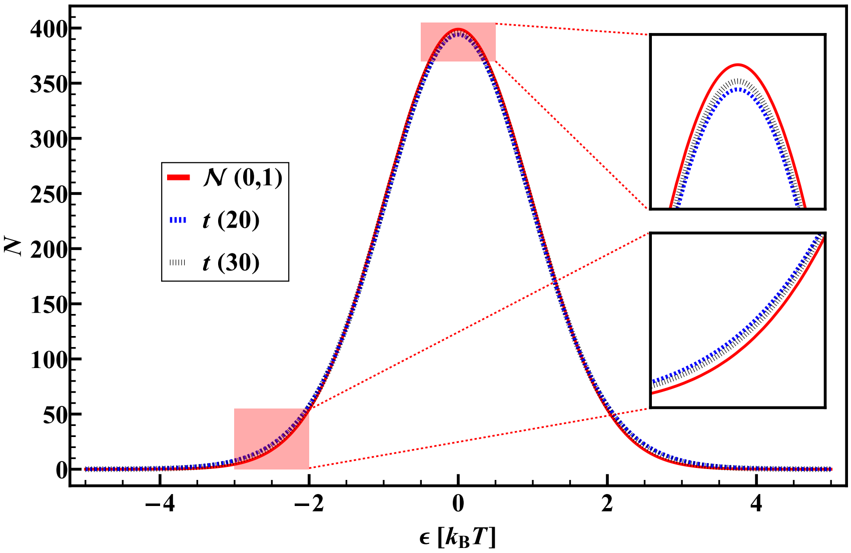


Figure 2.6: Comparison between three distributions: Student's t with 20 and 30 degrees of freedom (t (20) and t (30), respectively) and a standard Gaussian ($\mathcal{N}$(0,1)). There are hardly any differences among the three – the peak of the Gaussian is slightly higher than those of the Student's ts and its tail is slightly less heavy compared to theirs (top and bottom inset, respectively)

2.4.3 Moment- and cumulant-generating functions

Establishing the normal distribution of the binding sites on a DNA strand has set the path to one of the main goals of our work – calculating the grand partition function of a one dimensional lattice of adsorption site, whose energies obey a statistical distribution. While tackling this task, we shall encounter an interesting problem – we will have to average over the Boltzmann weights $\exp(-\beta\epsilon_i)$ of all sites, present on the DNA strand. In other words, we will have to calculate $\langle\exp(-\beta\epsilon)\rangle$. When we consider energies, obeying a Gaussian, this task can be reduced to finding the average of a log-normal distribution, since the exponent of a normally distributed variable with mean $-\beta\langle\epsilon\rangle$ and a standard deviation of $\beta\sigma$ is actually a log-normally distributed variable. Or in mathematical terms this reads: for $-\beta\epsilon \sim \mathcal{N}(-\beta\langle\epsilon\rangle, \beta^2\sigma^2)$

$$\langle\exp(-\beta\epsilon)\rangle = \exp\left(-\beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2}\right) \tag{2.46}$$

Although this result is pivotal to our work, it is instructive to forget for a moment we are dealing with normally-distributed variables and calculate the average exponent of energies, obeying a distribution of choice. This generalisation aims at demonstrating our considerations are not only applicable to analytical distributions but also ones that can not be expressed via a function.

We begin our derivation of the average exponent in a rigorously mathematical manner, which requires we first expand the exponent into a Taylor series:

$$\langle\exp(-\beta\epsilon)\rangle = \left\langle \sum_{n=0}^{\infty} \frac{(-\beta\epsilon)^n}{n!} \right\rangle = \sum_{n=0}^{\infty} (-\beta)^n \frac{\langle\epsilon^n\rangle}{n!} = \sum_{n=0}^{\infty} \mu_n \frac{(-\beta)^n}{n!} = M_\epsilon(-\beta). \tag{2.47}$$

This series is knows as the *moment-generating function* or MGF $M_\epsilon(-\beta)$ and is used to calculate the raw moments $\mu_n = \langle\epsilon^n\rangle$ of a distribution, hence its name. Straightforward as this approach may be, it is not particularly useful, because it fails to yield a closed form for the MGF. To illustrate this, we list the first few moments of the normal distribution under consideration:

$$\begin{aligned} \mu_1 &= \langle\epsilon\rangle \\ \mu_2 &= \langle\epsilon\rangle^2 + \sigma^2 \\ \mu_3 &= \langle\epsilon\rangle(\langle\epsilon\rangle^2 + 3\sigma^2) \\ \mu_4 &= \langle\epsilon\rangle^4 + 6\langle\epsilon\rangle^2\sigma^2 + 3\sigma^4 \end{aligned} \tag{2.48}$$

If we choose a different distribution, the moments also change, *i.e.* this approach fails to yield a universal, distribution-independent result. Obtaining a relatively simple expression for the average exponent requires the use of another statistical function, closely related to the MGF – the *cumulant-generating function* (CGF) $K(-\beta)$, which is defined as the natural logarithm of the MGF:

$$K(-\beta) = \ln M_\epsilon(-\beta) \Rightarrow \langle\exp(-\beta\epsilon)\rangle = M_\epsilon(-\beta) = \mathrm{e}^{K(-\beta)} \tag{2.49}$$

Much like the MGF, CGF is a series expansion, but the coefficients in the sum are the *cumulants* κ_n of the distribution (in contrast to the MGF, where the coefficients are moments):

$$K(-\beta) = \sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^n}{n!}, \tag{2.50}$$

and, again mirroring the MGF, it is commonly used to calculate κ_n. The cumulants are also known as Ursell functions [32] in the fields of statistical mechanics.

We now take a step back and clarify what the cumulants are and why we prefer to use them instead of the moments. First and foremost, when dealing with a set of data, we can easily calculate them without making any assumptions regarding the nature of the distribution since the first cumulant is the average, the second is the variance, and the third and fourth are the skewness γ_1 and the excess kurtosis γ_2, scaled by the variance:

$$\begin{aligned}
\kappa_1 &= \langle\epsilon\rangle \\
\kappa_2 = \sigma^2 &= \left\langle(\epsilon - \langle\epsilon\rangle)^2\right\rangle \\
\kappa_3 = \gamma_1\kappa_2^{3/2} &= \left\langle(\epsilon - \langle\epsilon\rangle)^3\right\rangle \\
\kappa_4 = \gamma_2\kappa_2^2 &= \left\langle(\epsilon - \langle\epsilon\rangle)^4\right\rangle - 3\kappa_2^2
\end{aligned} \tag{2.51}$$

Here, we should clarify what kind of information γ_1 and γ_2 give us about the data. The third cumulant reflects any asymmetry around the mean the data might exhibit. Actually, all cumulants of odd order (except the first), describe asymmetry – the higher the order of the cumulant, the more descriptive it is for the tails of the distribution. In other words, for symmetrical distributions all κ_n, where n is odd and greater than 1, should be zero, due to the lack of any skew. On the other hand, even order cumulants (with the exception of κ_2), are a way to estimate the pointiness or flatness of the distribution. One such is κ_4, which is a function of the excess kurtosis γ_2. γ_2 is a means to compare the distribution of interest to the normal distribution in terms of how far the tails of the distribution stretch and how heavy they are [33]. To grasp the physical meaning of κ_4, in Fig. 2.7 we compare three symmetric distributions centred around zero with variances of 1 ($\kappa_1 = \kappa_3 = 0$ and $\kappa_2 = 1$): the Laplace distribution, with its long tails, has an excess kurtosis of $\gamma_2 = 3$, the uniform distribution, which lacks any form of tails, has $\gamma_2 = -1.2$ and a Gaussian, which is an intermediate case, with $\gamma_2 = 0$. It is now clear that distributions like the Laplace one which are "pointier" and have tails stretching further away from the mean compared to a Gaussian of the same variance have a positive excess kurtosis, while "boxy" distributions, with short tails have a negative fourth cumulant.

As we see from Eq. 2.51 the first few cumulants can readily be calculated from any data set. Higher order cumulants cannot be expressed as simply, but that should not worry us, since, for most practical uses, the first four cumulants are sufficient to obtain a good estimate of the CGF.

That being said, we can outline three main properties of $K(-\beta)$ that make it particularly convenient for calculating $\langle\exp(-\beta\epsilon)\rangle$:

1. It conserves the functional form of the Boltzmann weight – $\langle\exp(-\beta\epsilon)\rangle$ is still an exponential function, but its argument is a sum of contributions, rather than one single energy.

2. Cumulants can be extracted rather easily from a set of data without having to fit it to a distribution, while moments require assuming a specific distribution and only then are we able to calculate them. This is actually a major advantage of the CGF over the MGF since we can apply the cumulant-generating function to any set of binding energies even if it does not obey an analytical distribution.

3. Most commonly used distributions have a finite number of non-zero cumulants or an infinite number of cumulants, which form a series and one can easily calculate their sum.

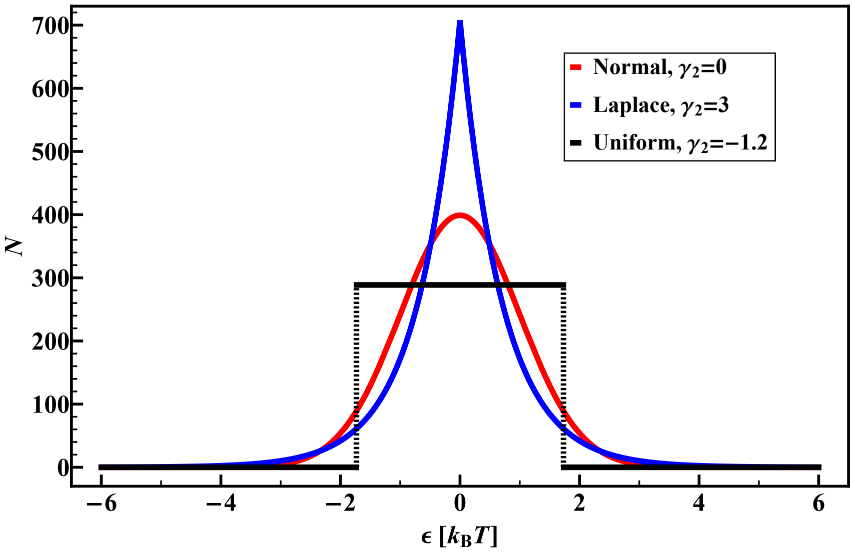


Figure 2.7: Comparison between three symmetric and zero-centred distributions with equal variances, but varying excess kurtoses. The shape, heaviness and outstretch of the tails defines the value of the fourth cumulant.

While the first property of the CGF is self-explanatory and obvious (we shall expand on it in Chapter 3, where we put the CGF to use), and we already explained the second, the third requires some additional reasoning and examples. To that end we calculate $\langle \exp(-\beta\epsilon) \rangle$ for the Gaussian and Laplace distributions.

MGF of Gaussian distribution In this example we will calculate $\langle \exp(-\beta\epsilon) \rangle$ for a normal distribution by using the definition of the average value of a variable obeying a continuous distribution. We draw inspiration from the 'Technical Notes on Statistics' by G. Lebanon [34].

The probability density function PDF of a set of normally distributed binding energies $\beta\epsilon$ with average $-\beta\langle\epsilon\rangle$ and standard deviation $\beta\sigma$ is given by:

$$f_{\mathrm{ND}} = \frac{1}{\beta\sigma\sqrt{2\pi}} \exp\left(- \frac{(\epsilon - \langle\epsilon\rangle)^2}{2\sigma^2} \right) \tag{2.52}$$

To average a variable, one multiplies it by the PDF and integrates the product over the entire support of the distribution (in our case $\beta\epsilon \in (-\infty;\infty)$):

$$M_\epsilon(-\beta) = \langle \exp(-\beta\epsilon) \rangle = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(\epsilon - \langle\epsilon\rangle)^2}{2\sigma^2} \right) \exp(-\beta\epsilon) \mathrm{d}\epsilon \tag{2.53}$$

We change the integration variable to $r = (\epsilon - \langle\epsilon\rangle)/\sqrt{2}\sigma$, which require us to multiply the expression by $\sqrt{2}\sigma$ and transforms $\exp(-\beta\epsilon)$ into $\exp(-\sqrt{2}\beta\sigma r - \beta\langle\epsilon\rangle)$ (all terms independent of r are brought outside the integral):

$$M_\epsilon(-\beta) = \frac{\sqrt{2}\sigma \exp(-\beta\langle\epsilon\rangle)}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\big(- (r^2 + \sqrt{2}\beta\sigma r) \big) \mathrm{d}r \tag{2.54}$$

We recognize the incomplete square of $r + \beta\sigma/\sqrt{2}$ in the argument of the integrand. In order to obtain the complete square, we multiply and divide by $\exp(\beta^2\sigma^2/2)$:

$$M_\epsilon(-\beta) = \frac{\exp(-\beta\langle\epsilon\rangle + \beta^2\sigma^2/2)}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\big(-(r+\beta\sigma/\sqrt{2})^2\big)\mathrm{d}r \tag{2.55}$$

We make one more change of variables, namely $z = r + \beta\sigma/\sqrt{2}$, which results in a Gaussian integral equalling $\sqrt{\pi}$:

$$M_\epsilon(-\beta) = \frac{\exp(-\beta\langle\epsilon\rangle + \beta^2\sigma^2/2)}{\sqrt{\pi}} \underbrace{\int_{-\infty}^{\infty} \mathrm{e}^{-z^2}\mathrm{d}z}_{\sqrt{\pi}} = \exp\left(-\beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2}\right) \tag{2.56}$$

Thus, we obtain the exact same result as the one following from our realisation that $\exp(-\beta\epsilon)$ is log-normally distributed, which was to be expected. From Eq. 2.56 we draw the conclusion that it is not only the position (the mean energy) that matters for the value of the average Boltzmann weight, but also the width of the distribution reflected by the variance term. We shall see what the physical reasoning for this result is when we consider it in the context of the grand canonical partition function in Chapter 3.

MGF of Laplace distribution In contrast to the previous example where our approach was purely mathematical, in this derivation we will make use of the symmetry of the distribution and the cumulant-generating function.

One thing to know before we start our derivation is whether the even order cumulants follow any trend. To that end we consult with Abramowitz and Stegun's *Handbook of Mathematical Functions* [35] and notice two tendencies:

1. All odd order cumulants, except the first, are equal to zero, as one would expect for a symmetric distribution
2. All even order cumulants can be expressed with the simple formula:

$$\kappa_n = 2(n-1)!s^n, \tag{2.57}$$

where s is the shape parameter of the distribution.

With this knowledge in hand, we are now ready to derive the MGF for a Laplace distribution:

$$M_\epsilon(-\beta) = \exp\big(K(-\beta)\big) = \exp\left(\sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^n}{n!}\right) \tag{2.58}$$

For now we focus only on the CGF, separate the first term from the rest of the sum and simplify the remainder:

$$K(-\beta) = -\beta\langle\epsilon\rangle + \sum_{n=2}^{\infty} 2(n-1)!\frac{(-\beta s)^n}{n!} = -\beta\langle\epsilon\rangle + \sum_{n=2}^{\infty} \frac{2}{n}(-\beta s)^n \tag{2.59}$$

Since only even summands matter, we can change the summation variable $n = 2j$ and then make the substitution $z = \beta^2 s^2$ (for brevity):

$$K(-\beta) = -\beta\langle\epsilon\rangle + \sum_{j=1}^{\infty} \frac{2}{2j}(-\beta s)^{2j} = -\beta\langle\epsilon\rangle + \sum_{j=1}^{\infty} \frac{z^j}{j} = -\beta\langle\epsilon\rangle - \ln(1-z) \tag{2.60}$$

Plugging Eq. 2.60 into Eq. 2.58 then yields:

$$M_\epsilon(-\beta) = \exp\big(-\beta\langle\epsilon\rangle - \ln(1-z)\big) = \frac{\mathrm{e}^{-\beta\langle\epsilon\rangle}}{1-z} = \frac{\mathrm{e}^{-\beta\langle\epsilon\rangle}}{1-\beta^2 s^2}, \tag{2.61}$$

To compare this result with results known from literature, we must compute the characteristic function $\varphi(-\beta)$ of the Laplace distribution, since this is the most often cited generating function. To that end, we carry out the same procedure as the one described above, but now instead of calculating the MGF for $-\beta$, we use $-\mathrm{i}\beta$, where i is the imaginary unit. Carrying all the calculation in this case yields:

$$\varphi(-\beta) = M_\epsilon(-\mathrm{i}\beta) = \frac{\mathrm{e}^{-\mathrm{i}\beta\langle\epsilon\rangle}}{1+\beta^2\sigma^2}, \tag{2.62}$$

which is identical with the characteristic function according to Abramowitz and Stegen, which brings us to the conclusion that our approach is correct. The aim of these two examples is to prove that $\langle\exp(-\beta\epsilon)\rangle$ can be expressed as a Boltzmann weight, in which instead of one single energy, we plug in a series of energy contributions according to the distribution, which we are studying. This result serves as basis for out main findings, presented in Chapter 3.

Chapter 3

Grand canonical partition function of a 1D lattice of non-specific sites obeying a distribution

In this chapter we derive the grand partition function for a transcription factor binding to a strand of non-specific DNA. As we pointed out in Section 2.2 this system does not differ conceptually from a one dimensional lattice of localized adsorption sites with binding energies obeying a statistical distribution. What sets non-specific DNA apart from a simple 1D lattice of sites, where one adsorbent molecule occupies only one site, is the idea of site overlap which will be discussed briefly in Section 3.4. Our first task will be to derive the grand canonical partition function for a Gaussian with a low variance and see that the partition function is dominated entirely by the contribution of the peak. Then we will generalise our considerations and, with the help of the cumulant-generating function, we will find a general form of the partition function of a unimodal (having one peak) distribution. We will further expand our model to mixture distributions and demonstrate its robustness. Our last task in this chapter will be to derive a criterion for the fugacity at which our model breaks down.

3.1 Normal distribution with low variance

In this section we focus on the simplified case of a transcription factor adsorbing on a DNA strand comprising sites with energies obeying a normal distribution and we assume this Gaussian has a low variance, *i.e.*, most sites have a binding energy very similar to the average. We will construct the grand canonical partition function Ξ for this system stepwise by first considering a single binding site with energy ϵ_i. Then we will look at the case when there are several copies of this site, forming a reservoir. Finally, we will write down the partition function for a set of N reservoirs having different numbers of sites N_i each. Since we are considering a normal distribution, we will make use of its symmetry.

We start off with one binding site with binding energy ϵ_i and occupation number p, for which the grand canonical partition function is given by:

$$\Xi_{1,i} = \sum_{p=0}^{1} \lambda^p \mathrm{e}^{-\beta p \epsilon_i}, \tag{3.1}$$

where $\beta = (k_\mathrm{B}T)^{-1}$ is the reciprocal of the thermal energy, $\lambda = \mathrm{e}^{\beta\mu}$ is the fugacity of the transcription factor adsorbing and μ is its chemical potential. There are two possible states of this system:

1. No protein adsorbs therefore $p = 0$, corresponding to unity
2. One protein molecule adsorbs at the site with energy ϵ_i therefore $p = 1$ and $\epsilon_i \neq 0$, corresponding to $\lambda \mathrm{e}^{-\beta\epsilon_i}$

This yields a partition function $\Xi_{1,i}$ for one site with energy ϵ_i:

$$\Xi_{1,i} = 1 + \lambda \mathrm{e}^{-\beta\epsilon_i} \tag{3.2}$$

On the other hand, for N_i such sites, independent of each other, the function is:

$$\Xi_{N_i} = \Xi_{1,i}^{N_i} = (1 + \lambda \mathrm{e}^{-\beta\epsilon_i})^{N_i} \tag{3.3}$$

Finally, when we take into consideration all possible sites with different energies (ranging from $\epsilon_1 = \epsilon_{\text{min}}$ to $\epsilon_N = \epsilon_{\text{max}}$), we derive the partition function for the entire set of non-specific sites:

$$\Xi = \prod_{i=1}^{N} \Xi_{N_i} = \prod_{i=1}^{N} \left(\sum_{p=0}^{1} \lambda^p \mathrm{e}^{-\beta p \epsilon_i} \right)^{N_i} = \prod_{i=1}^{N} (1 + \lambda \mathrm{e}^{-\beta\epsilon_i})^{N_i} \tag{3.4}$$

The boundaries of the product (N_1 and N_N) correspond to the sites with the lowest and the highest adsorption energy, respectively. If we add up all numbers of sites with energies ranging from ϵ_1 to ϵ_N, we will get the total number of sites N_{total}:

$$\sum_{i=1}^{N} N_i = N_{\text{total}} \tag{3.5}$$

For all derivations in this chapter it is more convenient to use a sum rather than a product. Therefore, we will take the natural logarithm of the partition function:

$$\ln \Xi = \sum_{i=1}^{N} N_i \ln(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}) \tag{3.6}$$

Then, we express all energies ϵ_i as functions of $\Delta\epsilon_i$:

$$\ln \Xi = \sum_{i=1}^{N} N_i \ln \left(1 + \lambda \mathrm{e}^{-\beta\langle\epsilon\rangle} \mathrm{e}^{-\beta\Delta\epsilon_i}\right), \tag{3.7}$$

where $\Delta\epsilon_i$ is defined as the energy deviation with respect to the average ($\Delta\epsilon_i = \epsilon_i - \langle\epsilon\rangle$). Since the distribution is symmetrical around $\langle\epsilon\rangle$, deviations from the average energy are also distributed symmetrically, *i.e.*, $-\Delta\epsilon_i = \Delta\epsilon_{N+1-i}$ and the number of sites having these energies is equal ($N_i = N_{N+1-i}$). Then, we make the substitution $x = \exp(-\beta\langle\epsilon\rangle)$ and present Eq. 3.7 as follows :

$$\begin{aligned} \ln \Xi &= N_1 \ln \left(1 + \lambda x \mathrm{e}^{-\beta\Delta\epsilon_1}\right) + N_{\mathrm{N}} \ln \left(1 + \lambda x \mathrm{e}^{-\beta\Delta\epsilon_{\mathrm{N}}}\right) + ... = \\ &= N_1 \ln \left[(1 + \lambda x \mathrm{e}^{-\beta\Delta\epsilon_1})(1 + \lambda x \mathrm{e}^{\beta\Delta\epsilon_1})\right] + ... = \\ &= \sum_{i=1}^{N/2} N_i \ln \left[(1 + \lambda x \mathrm{e}^{-\beta\Delta\epsilon_i})(1 + \lambda x \mathrm{e}^{\beta\Delta\epsilon_i})\right] = \\ &= \sum_{i=1}^{N/2} N_i \ln \left[1 + 2\lambda x \cosh(\beta\Delta\epsilon_i) + \lambda^2 x^2\right] \end{aligned} \tag{3.8}$$

Here, the upper boundary of the sum changes from N to $N/2$ because each product like the one on the second line of Eq. 3.8 uses 2 terms from the sum in Eq. 3.6: one term that corresponds to an adsorption energy lower than $\langle\epsilon\rangle$ and one corresponding to an adsorption energy higher than $\langle\epsilon\rangle$. Next, we focus on the argument of the hyperbolic cosine and perceive that for $\Delta\epsilon_{\max} < 0.5\ k_{\mathrm{B}}T$ the value of $\cosh\Delta\epsilon_i$ is approximately 1. To get a physical sense of a distribution complying with this requirement, we recall the '3σ rule', which states that roughly 99.7% of a Gaussian lie within three standard deviations around the mean. We then realise our approximation is applicable only to normal distributions with $\sigma < 0.17\ k_{\mathrm{B}}T$. Having said that, we further simplify the last line of Eq. 3.8:

$$\begin{aligned}\sum_{i=1}^{N/2} N_i \ln\left[1 + 2\lambda x\cosh(\beta\Delta\epsilon_i) + \lambda^2x^2\right] &\approx \sum_{i=1}^{N/2} N_i \ln(1+\lambda x)^2 = 2\ln(1+\lambda x)\sum_{i=1}^{N/2} N_i \\ &= N_{\mathrm{total}}\ln(1+\lambda x)\end{aligned} \tag{3.9}$$

In the last step we keep in mind that summing over all sites with energies ranging from ϵ_1 to $\langle\epsilon\rangle$ (since the distribution is symmetrical $\epsilon_{N/2}$ is the average energy) will yield half of the total number of sites:

$$\sum_{i=1}^{N/2} N_i = \frac{N_{\mathrm{total}}}{2} \tag{3.10}$$

Finally, we can derive the expression for the grand canonical partition function:

$$\Xi = \left(1 + \lambda \mathrm{e}^{-\beta\langle\epsilon\rangle}\right)^{N_{\mathrm{total}}} \tag{3.11}$$

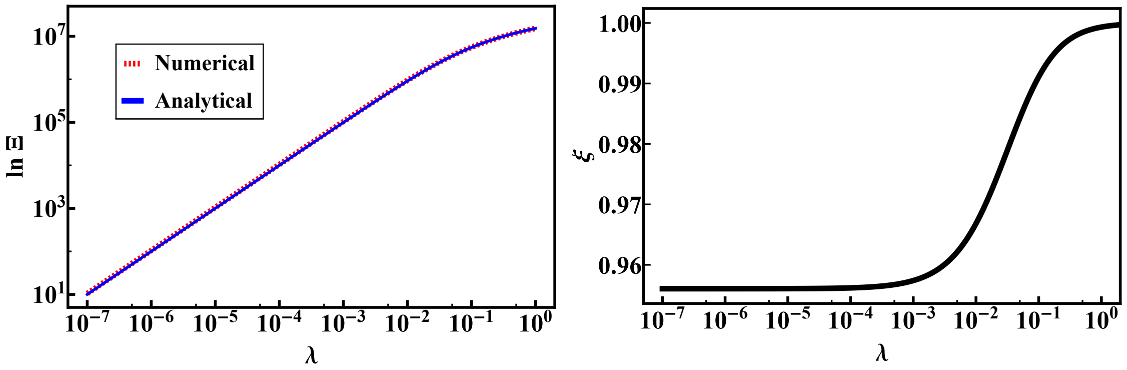
This result implies the partition function for a 1D lattice of normally distributed in energy binding sites with low variance is entirely dominated by the peak, *i.e.*, the shear number of sites with energy close to the average overpowers the contribution from the low-energy tail. The sites on the left of the mean may have a lower energy (thus, a greater Boltzmann weight) but they are too few in numbers to significantly change the value of the partition function.

To test this derivation, we compare the logarithm of the partition function Ξ_{Th} following from Eq. 3.11 to $\ln\Xi_{\mathrm{Num}}$ calculated numerically according to Eq. 3.6 (here, the two subscripts, "Th" and "Num" stand for theoretical and numerical approach). To do that, we construct a normal distribution of 5×10^6 sites with a mean energy $\langle\epsilon\rangle = -3\ k_{\mathrm{B}}T$ and standard deviation $\sigma = 0.3\ k_{\mathrm{B}}T$. Using the obtained distribution, we calculate Ξ numerically via Eq. 3.6. On the other hand, by plugging the parameters of the distribution in Eq. 3.11, we arrive at Ξ_{Th}. Comparison between the theoretical $\ln\Xi_{\mathrm{Th}}$ and numerical $\ln\Xi_{\mathrm{Num}}$ result is given in Fig. 3.1a.

From the figure it is evident that even for this value of σ, which does not comply with our assumption ($\sigma < 0.17\ k_{\mathrm{B}}T$), there is hardly any difference between the numerical and theoretical result. This result, however, seems suspicious and to get a real sense of the deviation from the theoretical formula, we present the ratio of $\ln\Xi_{\mathrm{Th}}$ to $\ln\Xi_{\mathrm{Num}}$, ξ, as a function of the fugacity in Fig. 3.1b

The ratio of the logarithms of the two partition functions is at a low plateau at $\xi \sim 0.95$ in the low fugacity range and quickly rises to 1 as $\lambda\exp(-\beta\langle\epsilon\rangle)$ approaches unity. The reasoning behind this is the fact that we have neglected the contribution from the low-energy sites in the left tail of the distribution, thus underestimating the value of the partition function.

We notice that even for this low standard deviation we have a difference of about 5% between the two partition functions. A table of $\xi = \ln\Xi_{\mathrm{Th}}/\ln\Xi_{\mathrm{Num}}$ for different standard



(a) Gaussian, $\langle\epsilon\rangle = -3\ k_\mathrm{B}T$ and $\sigma = 0.3\ k_\mathrm{B}T$

(b) $\xi(\lambda) = \ln \Xi_\mathrm{Th}(\lambda)/\ln \Xi_\mathrm{Num}(\lambda)$

Figure 3.1: (a) Comparison between the logarithms of the theoretical grand canonical partition function $\ln \Xi_\mathrm{Th}$ and its numerical counterpart $\ln \Xi_\mathrm{Num}$ calculated for a set of 5×10^6 sites obeying a normal distribution with an average $\langle\epsilon\rangle = -3\ k_\mathrm{B}T$ and standard deviation $\sigma = 0.3\ k_\mathrm{B}T$. The two partition functions are calculated according to Eq. 3.11 and 3.6, respectively. (b) The ratio of $\ln \Xi_\mathrm{Th}$ to $\ln \Xi_\mathrm{Num}$, ξ, as a function of the fugacity.

deviations is given below (Table 3.1). The presented values for ξ make it obvious now that Eq. 3.11, which only takes into account the contribution of the peak to Ξ, is far from accurate, especially for distributions with $\sigma > 0.5\ k_\mathrm{B}T$.

Table 3.1: $\xi(\lambda = 10^{-7})$ calculated for different standard deviations

σ, $[k_\mathrm{B}T]$	0.30	0.50	0.75	1.00	1.50	2.00	3.00
ξ	0.956	0.883	0.755	0.607	0.326	0.139	0.014

It is also interesting to see what happens when we plug in extreme values (+50 and $-50\ k_\mathrm{B}T$, for example) for the average adsorption energies, while retaining a standard deviation of $0.5\ k_\mathrm{B}T$. We should keep in mind that the non-specific binding energy and the chemical potential are coupled, *i.e.*, shifting one variable causes a shift in the other. This is due to the fact that in the case of non-specific binding the reservoir of constant chemical potential is the cell itself and the fugacity is determined by the number of adsorbing proteins. In other words, shifting the average energy without correcting the fugacity effectively leads to a change of the total number of adsorbents available in the cell. Having said that, we realize we should always scale the fugacity by the adsorption energy term. Therefore, it is only natural to plot $\ln \Xi$ vs. λx, where $x = \exp(-\beta\langle\epsilon\rangle)$.

To illustrate this, we investigate the two cases (with extremely high and extremely low binding energy) within a fixed fugacity range: $\lambda \in [10^{-7}, 10^2]$. For an average binding energy of zero $\langle\epsilon\rangle = 0$, the lower limit corresponds to only one molecule bound to the entire strand, while the upper limit corresponds to a 100-fold excess of adsorbents with respect to N_total. For the two cases we are interested in common sense dictates that we should observe two distinct cases:

1. When all sites have a strongly positive adsorption energy (in the range of 48.5 to 51.5 $k_\mathrm{B}T$), adsorption is highly unfavourable and no sites will be occupied. From this we conclude that the partition function will be 1. This assumption is confirmed

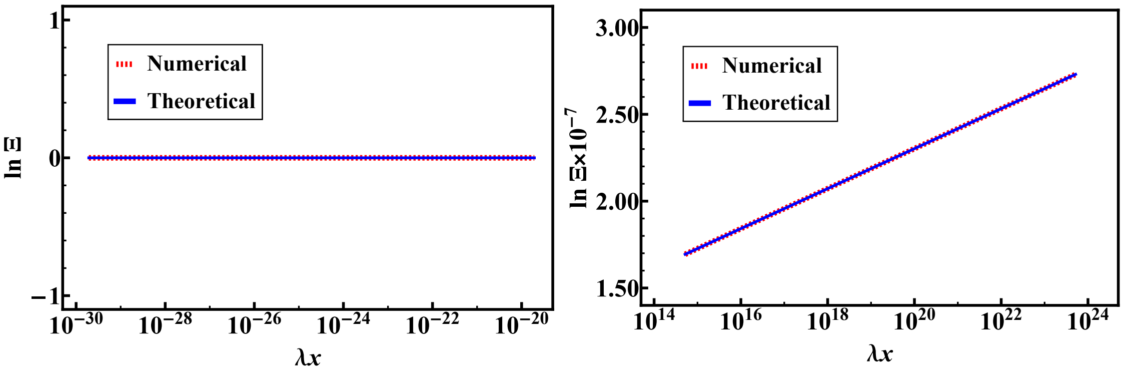
by Eq. 3.11, as well. For strongly positive values of $\langle\epsilon\rangle$ the second term in brackets is virtually zero. Therefore, we can expand the final line of Eq. 3.9 into a Maclaurin series:

$$\frac{\ln\Xi}{N_{\text{total}}} = \ln(1+\lambda x) \xrightarrow{\lambda x\to 0} \sum_{n=1}^{\infty}(-1)^{(n+1)}\frac{(\lambda x)^n}{n} \approx \lambda x \tag{3.12}$$

As we already pointed out in Section 2.2.2, the fugacity of a species A is approximately equal to the ratio of the number of molecules A adsorbed on the DNA strand to the length of the strand expressed in number of sites, *i.e.*, $\lambda \approx A/N_{\text{total}}$. This,however, is only valid in the case of $\langle\epsilon\rangle = 0$ – when the energy distribution is not centred around zero, one should scale the number of adsorbed proteins by x^{-1}. Therefore, we can simplify Eq. 3.12 even further:

$$\ln\Xi \approx N_{\text{total}}\lambda x \approx N_{\text{total}}\frac{Ax^{-1}}{N_{\text{total}}}x = A = 0, \tag{3.13}$$

a result, which we anticipated, because of the high adsorption energy.



(a) $\ln\Xi(\lambda x)$ for $\langle\epsilon\rangle = 50\ k_{\text{B}}T$

(b) $\ln\Xi(\lambda x)$ for $\langle\epsilon\rangle = -50\ k_{\text{B}}T$

Figure 3.2: (a) $\ln\Xi_{\text{Th}}$ and $\ln\Xi_{\text{Num}}$ as functions of λx for a Gaussian with $\langle\epsilon\rangle = 50\ k_{\text{B}}T$ and $\sigma = 0.5\ k_{\text{B}}T$. Both partition functions are equal to unity for the entire studied fugacity range $\lambda \in [10^{-7}; 10^{2}]$, as expected from Eq. 3.13. (b) $\ln\Xi_{\text{Th}}$ and $\ln\Xi_{\text{Num}}$ as functions of λx for a Gaussian with $\langle\epsilon\rangle = -50\ k_{\text{B}}T$ and $\sigma = 0.5\ k_{\text{B}}T$. Both partition functions are linear functions with zero intercepts and slope equal to the total number of sites, as expected from Eq. 3.14.

2. In the other extreme, a strongly negative average energy means adsorption on any of the non-specific sites will be highly favourable and molecules adsorb on all available sites. The second line of Eq. 3.9 supports this assumption – for strongly negative $\langle\epsilon\rangle$ the second term in brackets is much larger than unity and $\ln\Xi$ is proportional to $-\beta\langle\epsilon\rangle N_{\text{total}}$:

$$\ln\Xi = N_{\text{total}}\ln(1+\lambda x) \xrightarrow{\lambda x\gg 1} N_{\text{total}}\ln(\lambda x) = N_{\text{total}}(-\beta\langle\epsilon\rangle + \beta\mu) \tag{3.14}$$

It turns out that, in this extreme case, the natural logarithm of the partition function is linear with respect to $\ln\lambda x$, has a zero intercept and its slope is equal to the total number of sites. A plot of $\ln\Xi_{\text{Th}}$ and $\ln\Xi_{\text{Num}}$ vs. λx depicts this result (Fig 3.2b).

Finally, let us plot $\ln\Xi$ vs. λx for the entire λx-range studied up to now, *i.e.*, from 10^{-30} to 10^{20} for all cases (theoretical, numerical at positive and negative energies). The

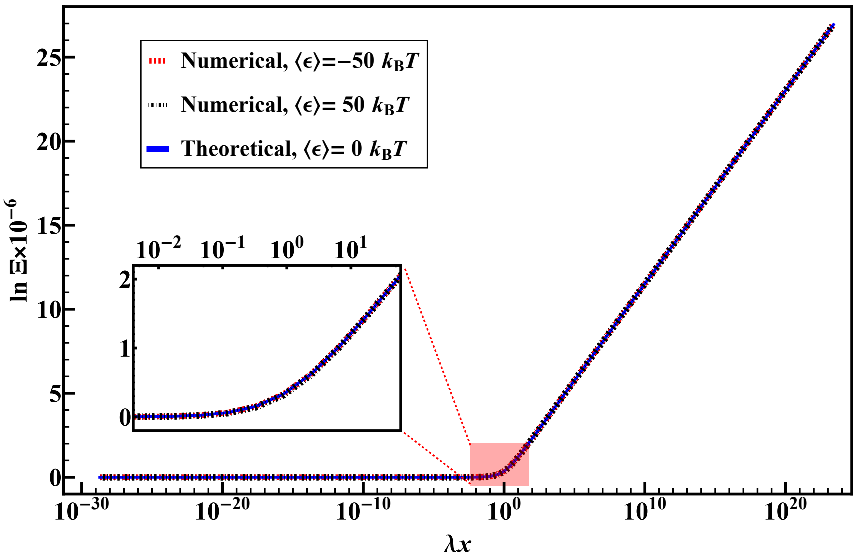


Figure 3.3: Comparison of two numerically calculated $\ln \Xi(\lambda x)$ at different average adsorption energies with the theoretical expression. The three partition functions overlap over the λx range $\lambda x \in [10^{-30}; 10^{20}]$

results are presented in Fig. 3.3. It is now clear that we can present every moderately broad distribution (up to $\sigma = 0.5\ k_{\mathrm{B}}T$) with a function $\ln \Xi = N_{\mathrm{total}} \ln(1 + \lambda x)$. The function has two distinct branches, reflecting each of the cases discussed above:

1. When $\lambda x \ll 1$ the partition function equals unity, corresponding to the physical picture of no binding.

2. In the case of $\lambda x \gg 1$ $\ln \Xi$ is linear with respect to $\ln(\lambda x)$, has a zero intercept and the slope of the line equals N_{total}, reflecting the case of binding to any site present on the strand.

3.2 Generalized derivation of the grand canonical partition function

In the previous section we derived the grand canonical partition function of a lattice with sites having normally distributed binding energy, but we made an assumption which might not always be true – we considered only narrow Gaussians with standard deviation of no more than 0.5 $k_{\mathrm{B}}T$. Hence, now we wish to expand our considerations to wider distributions. Furthermore, we pursue to derive the partition function as generally as possible, *i.e.*, we wish to include other distributions, as well, since we believe our result is applicable not only to DNA, but to any Langmuir lattice.

We again start off with the expression for the grand canonical partition function for a 1D lattice of statistically-distributed independent adsorption sites (Eq. 3.4), and, as we pointed out in the previous section, it is more convenient to derive $\ln \Xi$:

$$\ln \Xi = \sum_{i=1}^{N} N_i \ln(1 + \lambda \mathrm{e}^{-\beta \epsilon_i}) \tag{3.15}$$

where N_i, as usual, is the number of sites having adsorption energy ϵ_i and $\lambda = \mathrm{e}^{\beta\mu}$ is the fugacity of the adsorbing molecule. Next, we present the number of sites with energy ϵ_i in terms of the total number of sites N_{total} and the probability for a site to have this energy:

$$N_i = f(\epsilon_i)N_{\text{total}}, \tag{3.16}$$

where $f(\epsilon)$ is the probability density function for the given distribution. Our next step is to substitute Eq. 3.16 in Eq. 3.15, which yields:

$$\ln \Xi = N_{\text{total}} \sum_{i=1}^{N} f(\epsilon_i) \ln(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}) \tag{3.17}$$

We are allowed to isolate the total number of sites from the sum, because it does not depend on the summation variable. By definition, summing over all values of a random variable, multiplied by their probabilities, yields the average value of the variable:

$$\sum_{\{i\}} x_i f(x_i) = \langle x \rangle \tag{3.18}$$

That being said, we transform the right part of Eq. 3.17 into an average logarithm:

$$\ln \Xi = N_{\text{total}} \Big\langle \ln(1 + \lambda \mathrm{e}^{-\beta\epsilon}) \Big\rangle \tag{3.19}$$

One might be tempted to approximate the average logarithm to a logarithm of an average and this will, in fact, be a fairly good approximation, but only under certain conditions. To gain insight into what these conditions are, we will seemingly make our derivation more complicated than needed but in this way we will manage to express $\ln \Xi$ as a sum of a leading term, a logarithm of an average, and a term, which vanishes for low enough fugacities.

Having said that, we introduce a shorthand for the logarithm of the average, which we shall call a:

$$a = \ln\Big\langle 1 + \lambda \mathrm{e}^{-\beta\epsilon} \Big\rangle = \ln\Big(1 + \lambda \Big\langle \mathrm{e}^{-\beta\epsilon} \Big\rangle\Big) \tag{3.20}$$

We are allowed to make the last transformation since λ is independent of the adsorption energy. Next, we add and subtract $N_{\text{total}}a$ from Eq. 3.19:

$$\begin{aligned} \ln \Xi &= N_{\text{total}} \left(\left\langle \ln\left(1 + \lambda \mathrm{e}^{-\beta\epsilon}\right) \right\rangle + a - a \right) \\ &= N_{\text{total}} \left(a + \left\langle \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon}}{\mathrm{e}^{a}} \right\rangle \right) \\ &= N_{\text{total}} \left(a + \left\langle \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle} \right\rangle \right) \end{aligned} \tag{3.21}$$

Bringing $-a$ inside the averaging angle brackets is possible since, for a given fugacity, a is a function only of the average exponent, which as we already mentioned in Section 2.4.3 depends on the cumulants of the distribution and is constant for a given set of cumulants. Our next step is to substitute the second term in the last line of Eq.3.21 with b thus obtaining an expression for the partition function in a compact form:

$$\ln \Xi = N_{\text{total}} \left(a + \left\langle \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle} \right\rangle \right) = N_{\text{total}} \left(a + b\right) \tag{3.22}$$

In order to explain why a is the leading term in the expression, we must discuss how b behaves upon decreasing the fugacity:

$$b = \left\langle \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle} \right\rangle \tag{3.23}$$

For sufficiently low fugacity the second terms in both the numerator and the denominator become negligibly small compared to unity, the fraction itself tends to 1, which means $b \approx 0$ and it can be neglected in Eq. 3.22. It can be argued, however, that the same reasoning is applicable to the first term, as well. Although intuitively plausible, this is not the case – since roughly half of the values for $1 + \lambda \mathrm{e}^{-\beta\epsilon_i}$ are greater than $1 + \lambda \left\langle \mathrm{e}^{-\beta\epsilon} \right\rangle$ and the other half are smaller than it, half of the summands in $N_{\text{total}} b$ are negative and the other half positive:

$$N_{\text{total}} b = \sum_{i=1}^{N} \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon_i}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle} \approx \underbrace{\sum_{i=1}^{N/2} \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon_i}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle}}_{>0} + \underbrace{\sum_{i=N/2}^{N} \ln \frac{1 + \lambda \mathrm{e}^{-\beta\epsilon_i}}{1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle}}_{<0} \approx 0 \tag{3.24}$$

That being said the average of these summands, *i.e.* b, will converge to zero much more rapidly than a. To illustrate this, we plot the ratio of $-b$ to a as a function of λ (Fig. 3.4). The calculation of $-b/a$ is made for a normal distribution with mean $\langle \epsilon \rangle = -1.3\ k_{\mathrm{B}}T$ and standard deviation $\sigma = 2.3\ k_{\mathrm{B}}T$. For biological systems the typical values for the fugacity range roughly from 1×10^{-7} to 1×10^{-5}. From the graph we see that for this range the ratio monotonously increases from approximately 2×10^{-4} to 2×10^{-2}.

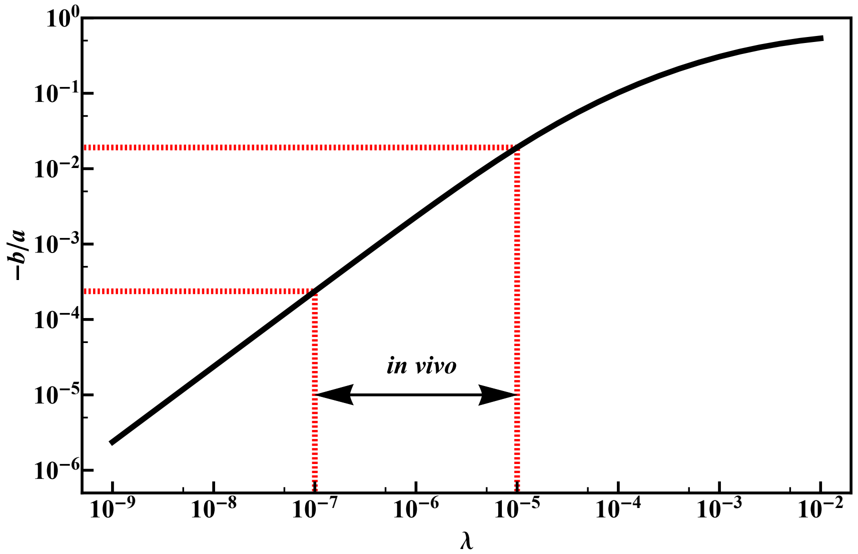


Figure 3.4: Ratio of $-b$ to a (see text for definitions) as a function of the fugacity λ for a normal distribution with $\langle \epsilon \rangle = -1.3\ k_{\mathrm{B}}T$ and $\sigma = 2.3\ k_{\mathrm{B}}T$. Under *in vivo* conditions the second term in Eq. 3.22 is negligible.

We can now safely drop the second term in Eq. 3.22 and write the first one explicitly:

$$\ln \Xi \simeq N_{\text{total}} \ln \left(1 + \lambda \left\langle \mathrm{e}^{-\beta\epsilon} \right\rangle \right) \tag{3.25}$$

As we already discussed in Section 2.4, the average exponent of a variable obeying a distribution of choice is the moment-generating function of the distribution. We also argued why the MGF is not convenient for our goals and decided to use its close relative, the cumulant-generating function:

$$\left\langle \mathrm{e}^{-\beta\epsilon} \right\rangle = M_\epsilon(-\beta) = \exp\left(K(-\beta)\right) = \exp\left(\sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^n}{n!}\right), \tag{3.26}$$

where κ_n is the n^{th} cumulant of the distribution. After this short recap, we finally re-write Eq. 3.25 in terms of a series of cumulants:

$$\ln \Xi \simeq N_{\mathrm{total}} \ln \left[1 + \lambda \exp\left(\sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^n}{n!}\right)\right] \tag{3.27}$$

If we now recall Eq. 2.51, we can write out the cumulant-generating function explicitly:

$$K(-\beta) = -\beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2} - \frac{\beta^3\gamma_1\sigma^3}{6} + \frac{\beta^4\gamma_2\sigma^4}{24} + ... \tag{3.28}$$

Having pointed this out, it is now clear that one can easily calculate $\ln \Xi$ for any distribution with a convergent moment-generating function via Eq. 3.27.

We now arrive at the conclusion that, for a distribution of choice, the expression for $\ln \Xi$ is much more complicated than the simple expression in Eq. 3.11. We can, however, define an effective energy ϵ_{eff}:

$$\epsilon_{\mathrm{eff}} \equiv \langle\epsilon\rangle - \frac{\beta\sigma^2}{2} + \frac{\beta^2\gamma_1\sigma^3}{6} - \frac{\beta^3\gamma_2\sigma^4}{24} + ..., \tag{3.29}$$

thus obtaining Eq. 3.27 in a simple, yet accurate form:

$$\ln \Xi \simeq N_{\mathrm{total}} \ln\left(1 + \lambda \exp\left(-\beta\epsilon_{\mathrm{eff}}\right)\right) \tag{3.30}$$

Although qualitatively identical, Eq. 3.11 and Eq. 3.30 differ significantly in quantitative terms – while Eq. 3.11 takes into account only the location ($\langle\epsilon\rangle$) of the distribution, Eq. 3.30 considers both the location and the shape of the distribution, which, as we will see, may lead to significant departures from Eq. 3.11. To illustrate this, we again present the $\xi = \ln \Xi_{\mathrm{Th}} \times \ln \Xi_{\mathrm{Num}}^{-1}$ for several standard deviations (Table 3.2), but this time we use Eq. 3.30 to calculate the partition function analytically. From the values listed it becomes evident Eq. 3.30 is exact for standard deviations up to $\sigma = 2k_{\mathrm{B}}T$, but its accuracy deteriorates rapidly for wider distribution.

Table 3.2: $\xi(\lambda = 10^{-7})$ calculated for different standard deviations

$\sigma, [k_{\mathrm{B}}T]$	0.30	0.50	0.75	1.00	1.50	2.00	3.00
ξ	1.000	1.000	1.000	1.001	1.005	1.025	1.255

Let us now take a step back and contemplate the physical reasoning behind Eq. 3.30. To that end, we will take Eq. 3.29 up to the first term and gradually add more and more terms. We will start by considering a set of sites with equal energy, which can be deemed a normal distribution with zero variance, that is, $K(-\beta) = -\beta\langle\epsilon\rangle$. This is actually how non-specific sites are usually modelled – as a set of sites with uniform binding energy. Thus, $\ln \Xi$ depends only on one parameter, the (mean) energy of the sites:

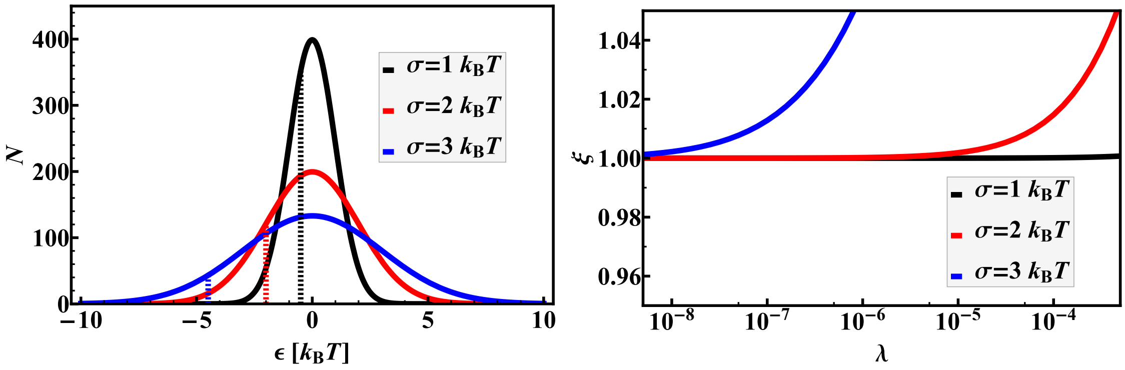
$$\ln \Xi = N_{\text{total}} \ln \Big[1 + \lambda \exp\Big(- \beta\langle\epsilon\rangle\Big)\Big] \quad \text{(for constant } \epsilon\text{)} \tag{3.31}$$

We now start to broaden the distribution ($\sigma > 0$), sites begin to differ in energy, two tails are formed – left, with energy lower than $\langle\epsilon\rangle$, and right, with higher energy. The molecules will preferentially bind to the sites with lower energy – the higher the variance, the further away from the average the tails stretch, therefore the lower the energy of these preferred sites. To account for this "bias", we add the second term in the cumulant-generating function:

$$\ln \Xi \simeq N_{\text{total}} \ln \left[1 + \lambda \exp\left(- \beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2}\right)\right] \quad \text{(for Gaussian)} \tag{3.32}$$

From this result it becomes clear why our derivation in Section 3.1 gave a fairly accurate estimate of $\ln \Xi$ up to $\sigma \approx 0.5\ k_{\text{B}}T$ – for a standard deviation of $\frac{1}{2}k_{\text{B}}T$, the second term in Eq. 3.32 is no more than 0.125, which (if we split the exponent), introduces a factor of 1.13 to the fugacity term. Recalling the numbers, presented in Table 3.1, we see that our theoretical partition function overestimates the actual one by exactly 1.13.

To demonstrate the role of the shape of the distribution, *i.e.*, its standard deviation, we present several Gaussians with $\langle\epsilon\rangle = 0$ and varying widths (Fig. 3.5a). We clearly see that the wider the distribution, the more ϵ_{eff} is shifted to the left. This means that one of the arguments we made, while deriving Eq. 3.30, might not be accurate, namely neglecting b in Eq. 3.22. We dropped this term under the pretext that it converges to zero much more quickly than a, because of symmetry around the effective energy. We now see that the effective energy slices the distribution more non-symmetrically the greater the standard deviation is. This explains why ξ for broader Gaussians diverges from unity at lower fugacities compared to narrow distributions (Fig. 3.5b and data presented in Table 3.2).

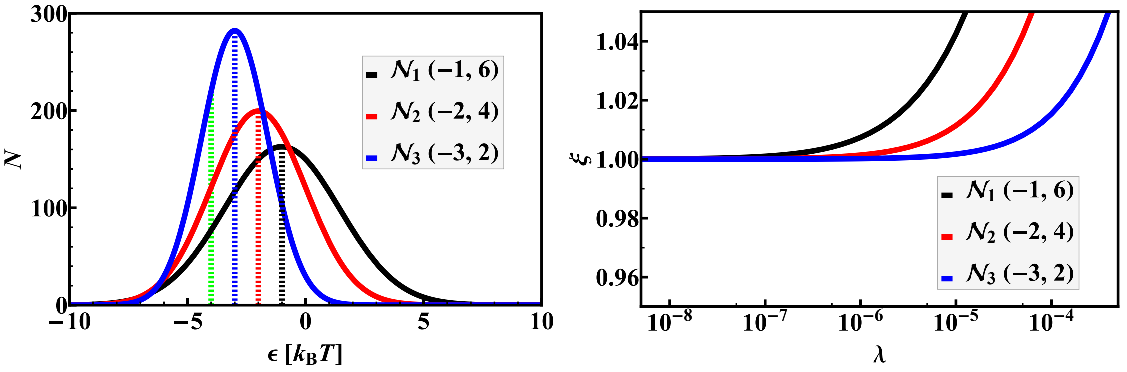


(a) ϵ_{eff} (dashed vertical lines)

(b) $\xi(\lambda)$ for different effective energies

Figure 3.5: (a) Comparison of the effective energies (dashed vertical lines) of normal distributions with $\langle\epsilon\rangle = 0\ k_{\text{B}}T$. For standard deviations up to $\sigma \leq 1\ k_{\text{B}}T$, ϵ_{eff} is hardly shifted with respect to the mean, in good agreement with our findings in Section 3.1. (b) $\xi(\lambda)$ for Gaussians with $\langle\epsilon\rangle = 0$ and varying standard deviation. The wider the distribution, the lower the fugacity at which ξ diverges from 1.

It is also interesting to see how ξ behaves when we fix the effective energy at, say $-4\ k_{\text{B}}T$, and use different combinations of $\langle\epsilon\rangle$ and σ. To that end, we plot three normal distributions and their corresponding $\xi(\lambda)$ functions (Fig 3.6). Using the mathematical notation for a normal distribution with an average $-\beta\langle\epsilon\rangle$ and variance $\beta^2\sigma^2$, $\mathcal{N}\left(-\beta\langle\epsilon\rangle,\ \beta^2\sigma^2\right)$, we choose the following Gaussians: $\mathcal{N}_1(-1,\ 6)$, $\mathcal{N}_2(-2,\ 4)$, and $\mathcal{N}_3(-3,\ 2)$. From Fig. 3.6b it is evident it is not only the effective energy, which determines the convergence of b (Eq. 3.23) to zero – the width of the distribution plays a separate role in b, as well.



(a) Constant ϵ_{eff} and varying $\langle\epsilon\rangle$ and σ

(b) $\xi(\lambda)$ for Gaussians with constant ϵ_{eff}

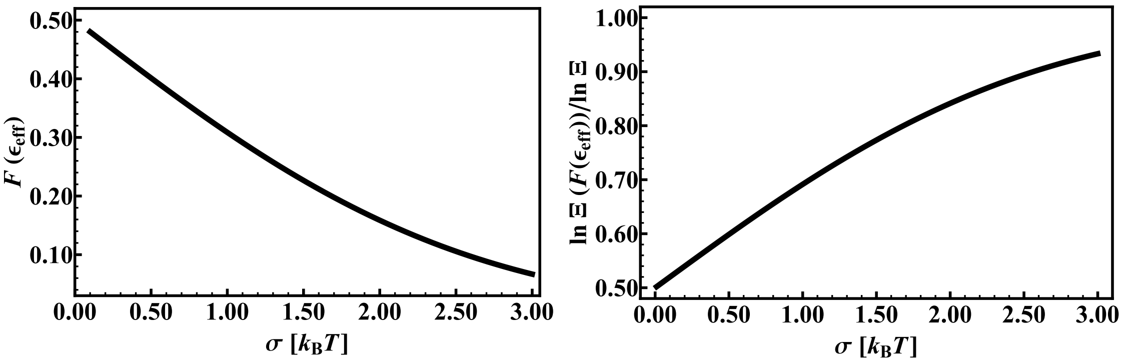
Figure 3.6: (a) Normal distributions with a fixed effective energy $\epsilon_{\text{eff}} = -4\ k_{\text{B}}T$ (green vertical dashed line) and varying means (the black, red, and blue vertical dashed lines). Shifting the distribution to the left leads to its narrowing, due to the constrained effective energy. (b) ξ for the three distributions on the left. It is clear that the variance of the distribution plays a separate role in the convergence of b (Eq. 3.23) to zero.

The reader may ask the reasonable question 'But why doesn't the transcription factor bind to one of the low energy sites in the tail?', e.g. sites with energy $\epsilon < \langle\epsilon\rangle - 2.5\sigma$. The answer to that question lies in the low abundance of these sites, *i.e.*, although having a great Boltzmann weight, thus being highly favourable for binding, there are too few sites like that on the DNA strand. What is more, the number of sites with energy lower than $\epsilon = \langle\epsilon\rangle - \beta\sigma^2/2$ progressively diminishes as the distribution broadens. To illustrate this, we plot the cumulative distribution function CDF $F(\beta\epsilon_{\text{eff}})$ (not to be mistaken with the cumulant-generating function) at $\epsilon_{\text{eff}} = \langle\epsilon\rangle - \beta\sigma^2/2$ vs. σ (Fig. 3.7a). In this case, again for simplicity, we have set the average to zero. Upon calculating the CDF for a given energy, we obtain what fraction of the sites have an energy no higher than the chosen one:

$$F(\beta\epsilon_{\text{eff}}) = \int_{-\infty}^{\beta\epsilon_{\text{eff}}} f(\epsilon)\mathrm{d}\epsilon = \frac{1}{2}\text{erfc}\left(\frac{\beta\sigma}{2\sqrt{2}}\right) \tag{3.33}$$

From the figure it is evident that the CDF is roughly 0.5 when $\sigma = 0.1\ k_{\text{B}}T$, which is to be expected since the effective energy is virtually the same as the average. Increasing the standard deviation leads to a monotonous decrease in the CDF, and at $\sigma = 3\ k_{\text{B}}T$ less then 10% of all sites have energy no higher than the effective.

Although very few, these sites absolutely dominate the partition function due to their great Boltzmann weight. To demonstrate this we need to calculate the partition function for the sites with $\epsilon \leq \epsilon_{\text{eff}}$. To that end we use Eq. 3.17, but now, instead of summing over the entire spectrum of binding sites, we only take the ones on the left hand side of the



(a) $F(\beta\sigma)$ for a Gaussian with $\langle\epsilon\rangle = 0\ k_{\text{B}}T$ (b) Relative weight of sites with energy $\epsilon \leq \epsilon_{\text{eff}}$

Figure 3.7: (a) Cumulative distribution function F at the effective energy as function of the standard deviation. The fraction of sites having energy no higher than the effective ($\epsilon_i \leq \epsilon_{\text{eff}}$) decreases for wider distributions (b) $\ln \Xi$ of the sites having energy $\epsilon \leq \epsilon_{\text{eff}}$ compared to the total $\ln \Xi$ as function of the standard deviation.

effective energy:

$$\begin{aligned}\ln \Xi \left(F\left(\beta\epsilon_{\text{eff}}\right)\right) &= N_{\text{total}} \sum_{i=1}^{\text{eff}} f\left(\epsilon_i\right) \ln(1+\lambda \mathrm{e}^{-\beta\epsilon_i}) \simeq \\ &\simeq N_{\text{total}} \ln\left[1+\frac{1}{2}\lambda \mathrm{e}^{-\beta\epsilon_{\text{eff}}}\left(1+\operatorname{erf}\left(\frac{\beta\sigma}{2\sqrt{2}}\right)\right)\right] = \\ &= N_{\text{total}} \ln\left[1+\lambda \mathrm{e}^{-\beta\epsilon_{\text{eff}}}\left(1-F\left(\beta\epsilon_{\text{eff}}\right)\right)\right]\end{aligned} \tag{3.34}$$

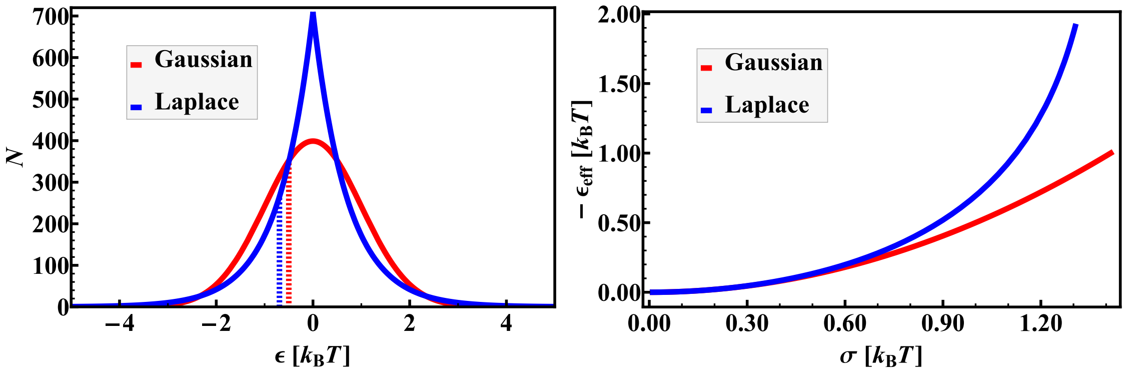
where we use the equation linking the error function and its complementary: $\operatorname{erf}(x) = 1 - \operatorname{erfc}(x)$. If we now divide this result by the logarithm of the total partition function, we obtain the relative statistical weight of the sites with energy $\epsilon_i \leq \epsilon_{\text{eff}}$ and see they contribute the most to the total partition function (Fig. 3.7b). Although comprising only a small fraction of the total number of sites, these sites' Boltzmann weights make them favourable for binding and they manage to outplay the rest of the genome. This finding in the context of non-specific binding can actually be parallel to specific binding, as well. While very few in numbers, the specific sites on DNA manage to outcompete the entire genome by having a binding energy several $k_{\text{B}}T$ lower than the effective one for the non-specific reservoir.

After this extensive analysis of the normal distribution, we turn our attention to curved, but symmetric distributions (that is, excessively long tails or the lack of such). Let us take for example the Laplace distribution with an average of $\langle\epsilon\rangle$ and a shape parameter s (note that this is not the standard deviation of the distribution). Here, $\gamma_2 > 0$ and all additional even-order terms in $K(-\beta)$ will be positive, which will lead to larger values of $\ln \Xi$. The reason for that is the existence of a small number of sites having energy considerably lower than $\langle\epsilon\rangle$.

$$\begin{aligned}\ln \Xi &\simeq N_{\text{total}} \ln\left[1+\lambda \exp\left(-\beta\langle\epsilon\rangle + \frac{2\beta^2 s^2}{2} + \frac{4\beta^4\gamma_2 s^4}{24} + \sum_{i=3}^{\infty}\frac{(\beta s)^{2i}}{i}\right)\right] \\ &= N_{\text{total}} \ln\left(1+\lambda\frac{\mathrm{e}^{-\beta\langle\epsilon\rangle}}{1-\beta^2 s^2}\right) \quad \text{(for Laplace distribution)}\end{aligned} \tag{3.35}$$

In order to compare the normal and the Laplace distributions, we need to set their common cumulants, *i.e.*, the average and the variance, equal. To that end, we remember the variance formula for the Laplace distribution – $\kappa_2 = 2s^2$, that is, the shape parameter s should be $\sqrt{2}$ times smaller than σ. Also we must take into account another constraint over s, namely, $\beta s < 1 \Rightarrow s < 1\ k_{\mathrm{B}}T$, otherwise the moment-generating function diverges, which physically means the tails of the distribution stretch so far away from the mean that the effective energy tends to infinity.

With this knowledge in mind, we choose to center both distributions at 0 and have $\sigma = 1\ k_{\mathrm{B}}T$ and $s = \sqrt{2}^{-1}\ k_{\mathrm{B}}T$. With these parameters, we can immediately calculate the effective energies of the two distributions: $-0.50\ k_{\mathrm{B}}T$ for the Gaussian and $-0.69\ k_{\mathrm{B}}T$ for the Laplace distribution (Fig. 3.8a). According to Eq. 3.30, we can collapse the entire information carried by the distribution to a single energy value. In that sense, comparing the effective energies as functions of σ (since we have fixed the ratio of σ to s, we have only one free parameter) for both distributions gives us a good estimate of how their partition functions relate (Fig. 3.8b). We see that the two effective energies are (almost) identical up to $\sigma = 0.5\ k_{\mathrm{B}}T$, but for greater standard deviations the Laplace effective energy begins to diverge to infinity as σ approaches $\sqrt{2}$.



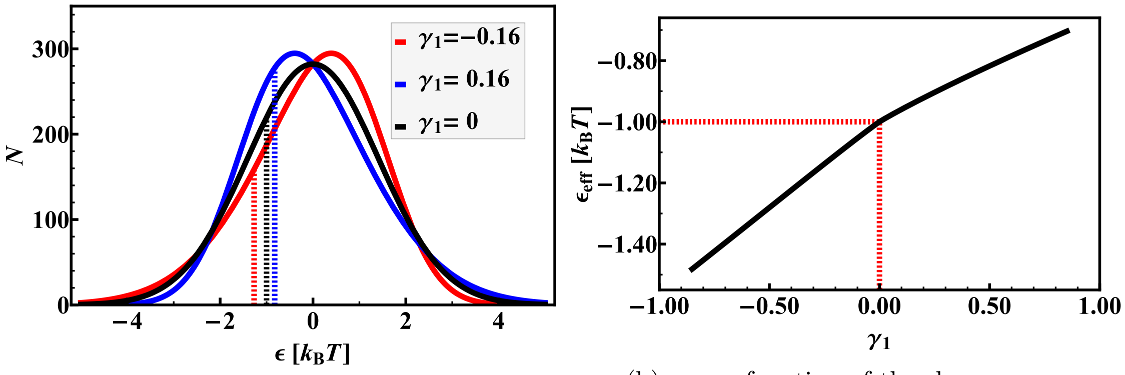
(a) ϵ_{eff} for normal and Laplace distributions

(b) $\epsilon_{\mathrm{eff}}(\sigma),\ s = \sigma/\sqrt{2}$

Figure 3.8: (a) Comparison between the Laplace distribution and a Gaussian and their effective energies. Both distributions are centred around zero and have a fixed variance of $1\ (k_{\mathrm{B}}T)^2$ (b) Effective energies for a Gaussian and a Laplace distribution as functions of the standard deviation of the Gaussian. Both distributions are centred around zero and the shape parameter of the Laplace distribution is constrained to $s = \sigma/\sqrt{2}$

Finally, if we introduce a skew to our distribution, we need to add the third term in Eq. 3.29, as well. Since a skewed distribution means having a long tail, we also need to take into account all other terms in the expansion of the CGF. To realize the physical reasoning behind the third term, we must consider a skew normal distribution, where one of the tails is heavier (not longer; just having a greater statistical weight).

If the left tail (the one corresponding to energies lower than $\langle\epsilon\rangle$) is heavier, then the distribution has a negative skewness ($\gamma_1 < 0$) which makes the third term in Eq. 3.28 positive, which, on the other hand, leads to a larger value of $\ln \Xi$. The reason for that is the greater number of sites with energy lower than $\langle\epsilon\rangle$ "competing" in molecule binding, compared to the number of sites with energy higher than $\langle\epsilon\rangle$. The same reasoning applies if $\gamma_1 > 0$, because the right tail of the distribution is heavier and sites having higher energy are more likely to be encountered by the molecule. In this case we, again, have competition between favourable binding energy and abundance – on one hand, the energy of these sites



(a) ϵ_{eff} for different skewnormal distributions

(b) ϵ_{eff} as function of the skewness γ_1

Figure 3.9: (a) Comparison of skewnormal distributions and their effective energies. In all three cases $\langle\epsilon\rangle = 0\ k_{\text{B}}T$ and $\sigma^2 = 2\ k_{\text{B}}T$ (b) $\epsilon_{\text{eff}}\left(\gamma_1\right)$ for a skewnormal distribution.

makes molecule binding to them highly unfavourable, on the other hand, their abundance makes up for that, which effectively leads to smaller values of $\ln \Xi$. Fig. 3.9b is a good illustration of our reasoning – moving the skew from left to right leads to an increase in the effective energy.

It is interesting to note there are two distinct branches of $\epsilon_{\text{eff}}\left(\gamma_1\right)$ - for negatively-skewed distributions the slope of $\epsilon_{\text{eff}}\left(\gamma_1\right)$ is greater than the slope for positive skew. This is due to dependence of the excess kurtosis on the skewness:

$$\text{slope} = \frac{\mathrm{d}\epsilon_{\text{eff}}}{\mathrm{d}\gamma_1} = \frac{\beta^2\sigma^3}{6} - \frac{\beta^3\sigma^4}{24}\frac{\mathrm{d}\gamma_2}{\mathrm{d}\gamma_1} \tag{3.36}$$

Increasing γ_1 from a negative value to 0 leads to a decrease in the excess kurtosis from a positive value to 0 (as $\gamma_1 \to 0^-$, $\gamma_2 \to 0^+$, because the distribution becomes more and more normal). Thus, the derivative $\mathrm{d}\gamma_2/\mathrm{d}\gamma_1$ in Eq. 3.36 is negative, leading to an increased slope. The same reasoning can be applied to the second branch of $\epsilon_{\text{eff}}\left(\gamma_1\right)$, as well – due to departures from normality upon increasing γ_1, the derivative term is positive which decreases the slope.

It is now clear that, within reasonable approximation, the effective energy is able to summarise the entire information carried by a distribution, as long as we are not dealing with pathologically wide distributions for which the assumptions made while deriving Eq. 3.30 are inaccurate.

After discussing the applicability of our model to distributions with a single peak, we wish to push it to its limits and check how well it performs with mixture distributions, as well. A distribution like that can be presented as a linear combination of several other distributions, multiplied by the weight they have in the total distribution. To illustrate this, we will consider a mixture distribution, which is a sum of two Gaussians:

$$f_{\text{total}}(\epsilon_i) = p_1 f_1(\epsilon_i) + p_2 f_2(\epsilon_i), \tag{3.37}$$

where p_i is the weight of the distribution i:

$$p_i = \int_{-\infty}^{\infty} f_i(\epsilon)\mathrm{d}\epsilon \Big/ \sum_{j=1}^{2}\int_{-\infty}^{\infty} f_j(\epsilon)\mathrm{d}\epsilon \tag{3.38}$$

When written in terms of the total probability density function $f_{\text{total}}(\epsilon_i)$, Eq. 3.17 reads:

$$\ln \Xi = N_{\text{total}} \sum_{i=\text{min}}^{\text{max}} f_{\text{total}}(\epsilon_i) \ln \left(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}\right), \tag{3.39}$$

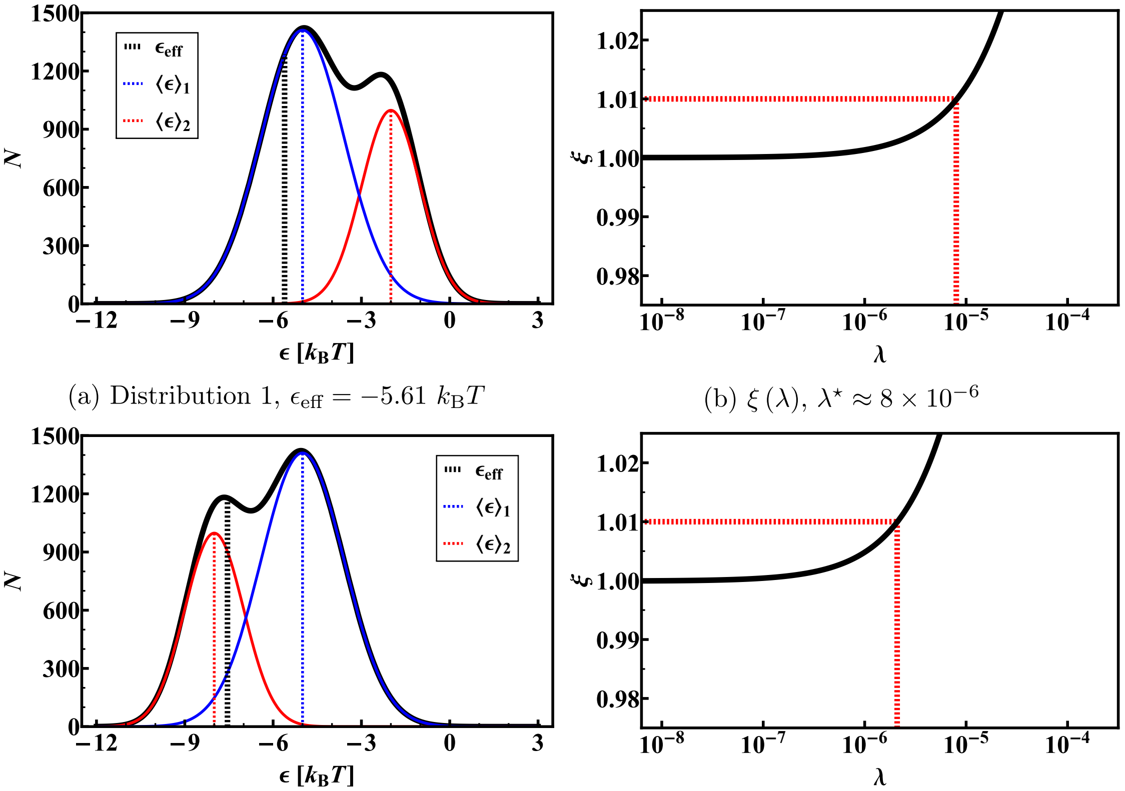
We can now split the sum in Eq. 3.39 into two sums, according to Eq. 3.37:

$$\ln \Xi = p_1 N_{\text{total}} \sum_{i=\text{min}}^{\text{max}} f_1(\epsilon_i) \ln \left(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}\right) + p_2 N_{\text{total}} \sum_{i=\text{min}}^{\text{max}} f_2(\epsilon_i) \ln \left(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}\right), \tag{3.40}$$

which we readily transform into:

$$\ln \Xi \simeq N_{\text{total}} \ln \left[1 + \lambda p_1 \exp\left(\sum_{n=1}^{\infty} \kappa_n^{(1)} \frac{(-\beta)^n}{n!}\right) + \lambda p_2 \exp\left(\sum_{n=1}^{\infty} \kappa_n^{(2)} \frac{(-\beta)^n}{n!}\right)\right], \tag{3.41}$$

following the procedure described above. The superscripts (1) and (2) signify that these are the cumulants of $f_1(\epsilon_i)$ and $f_2(\epsilon_i)$, respectively.



(a) Distribution 1, $\epsilon_{\text{eff}} = -5.61\ k_{\text{B}}T$

(b) $\xi\left(\lambda\right)$, $\lambda^{\star} \approx 8 \times 10^{-6}$

(c) Distribution 2, $\epsilon_{\text{eff}} = -7.55\ k_{\text{B}}T$

(d) $\xi\left(\lambda\right)$, $\lambda^{\star} \approx 2.1 \times 10^{-6}$

Figure 3.10: (a) and (c) Mixture distributions composed as sums of two Gaussians each. In both cases the ratios of weights is $p_1/p_2 = 2$. Both mixture distributions are composed of a normal distribution with $\beta\sigma_1 = \sqrt{2}$ (blue line) and another normal with $\beta\sigma_2 = 1$ (red line) (b) and (d) $\xi(\lambda)$ for the mixture distributions on the left. The divergence points $\lambda^{\star}$ are given in red dash

Eq. 3.41 can be even further generalized to a mixture distribution in which the single distributions are not necessarily normal:

$$\ln \Xi \simeq N_{\text{total}} \ln \left[1 + \lambda \sum_{i=1}^{\infty} p_i \exp\left(\sum_{n=1}^{\infty} \kappa_n^{(i)} \frac{(-\beta)^n}{n!}\right)\right] \tag{3.42}$$

Furthermore, we can again define ϵ_{eff} even for this general case:

$$\beta\epsilon_{\text{eff}} = -\ln \sum_{i=1}^{\infty} \exp\left(\ln p_i + \sum_{n=1}^{\infty} \kappa_n^{(i)} \frac{(-\beta)^n}{n!}\right) \tag{3.43}$$

but we will not go into that much detail and just present the results for two mixture distributions consisting of two Gaussians each (Fig. 3.10).

In both cases the energy difference between the individual peaks is 3 $k_\text{B}T$, but in Distribution 1 the less prominent peak is at energy higher than -5 $k_\text{B}T$, while in Distribution 2 it is at lower energy. From the right panels in Fig. 3.10 we conclude that centring the less prominent peak at lower energy simply shifts the effective energy ϵ_{eff} and the maximum fugacity $\lambda^\star$ for which the model works to lower values but does not affect the accuracy of the model in the low fugacity range. Here, we defined $\lambda^\star$ as the fugacity at which ξ becomes 1.01 and diverges even further from unity. In the next section we shall refer to $\lambda^\star$ as the *divergence point.* $\lambda^\star$ is shifted from roughly 10^{-5} in Distribution 1 to 10^{-6} in Distribution 2 and the effective energy is lowered by ≈ 2 $k_\text{B}T$. This is solely due to the greater number of sites having energy lower than -5 $k_\text{B}T$.

From all said thus far it is clear that the proposed model can be used to quickly and accurately calculate the grand canonical partition function, if three conditions are met:

1. The moment-generating function of the distribution of choice (be it analytical or not) needs to be convergent or dominated by the first order terms in the expansion

2. The product of the fugacity and effective Boltzmann weight is much smaller than unity ($\lambda \exp(-\beta\epsilon_{\text{eff}}) \ll 1$)

3. The standard deviation is not excessively high

Furthermore, from Eq. 3.29 and Eq. 3.30, we can conclude that instead of using a constant energy for the non-specific sites we are supposed to use an effective energy:

$$\epsilon_{\text{eff}} = \sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^{n-1}}{n!} \tag{3.44}$$

Finally, we should keep in mind that the true power of our model lies in its accurate prediction of the effective energy of any set of binding energies, even if the set cannot be fitted to an analytical distribution.

3.3 Convergence criterion

After defining the effective energy of the distribution of non-specific sites as the cumulant-generating function in the last chapter, we compared three Gaussians with constant ϵ_{eff} and noticed their divergence points $\lambda^\star$ differed, more specifically $\lambda^\star$ decreased as variance went up. We are now interested in investigating what is the role of the variance apart from being a contribution in the effective energy. Furthermore, we wish to establish a quantitative link between the parameters of a distribution and the fugacity range in which our model holds.

To that end we focus on the only approximation we made while deriving Eq. 3.30, *i.e.*, neglecting the term b in Eq. 3.22:

$$b = \left\langle \ln \frac{1 + \lambda e^{-\beta\epsilon}}{1 + \lambda\langle e^{-\beta\epsilon}\rangle} \right\rangle, \tag{3.45}$$

and check under what conditions b indeed converges to zero. First, we take into account that for broad distributions ($\sigma \geq 2\ k_\mathrm{B}T$) the effective energy is considerably shifted to the left from the average energy, which means that the greater part of the summands in Eq. 3.45 would be negative (the majority of the $\lambda\mathrm{e}^{-\beta\epsilon}$ terms will be smaller than $\lambda\left\langle\mathrm{e}^{-\beta\epsilon}\right\rangle$). Using this reasoning we can explain why we observe a sudden rise in the ratio of the theoretically to the numerically calculated partition functions – from a certain fugacity onwards b starts to become more and more negative. By neglecting this fact, we overestimate the partition function, which leads to a ratio greater than unity.

To define a convergence to zero criterion, we first convert the logarithm of a ratio to a logarithm difference and then expand the logarithms around zero (for brevity we use the substitution $q = \lambda\mathrm{e}^{-\beta\epsilon}$):

$$b = \langle \ln(1+q) - \ln(1+\langle q\rangle)\rangle = \left\langle \sum_{i=1}^{\infty}(-1)^{i+1}\frac{q^i}{i} - \sum_{i=1}^{\infty}(-1)^{i+1}\frac{\langle q\rangle^i}{i}\right\rangle \tag{3.46}$$

We now remember averaging is in its essence integration and an integral of a sum is a sum of integrals. Thus, we arrive at:

$$b = \sum_{i=1}^{\infty}(-1)^{i+1}\frac{\langle q^i\rangle - \langle q\rangle^i}{i} \tag{3.47}$$

We already established that q is log-normally distributed, since ϵ has a normal distribution. Therefore, the infinite sum of its moments is divergent and we must use a finite number of summands but more than one, otherwise we obtain zero, which is to be expected, since, in first approximation, b vanishes. If we use 2 as an upper limit of the sum, we obtain an expression which strongly resembles the definition of a variance:

$$b = \sum_{i=1}^{2}(-1)^{i+1}\frac{\langle q^i\rangle - \langle q\rangle^i}{i} = \langle q\rangle - \langle q\rangle - \frac{\left\langle q^2\right\rangle - \langle q\rangle^2}{2} = -\frac{1}{2}\left(\left\langle q^2\right\rangle - \langle q\rangle^2\right) \tag{3.48}$$

Therefore, b calculated up to the second term is actually proportional to the variance of a log-normal distribution:

$$\begin{aligned} b = -\frac{1}{2}\mathrm{Var}\left(q \sim \ln\mathcal{N}\left(-\beta\langle\epsilon\rangle, \beta^2\sigma^2\right)\right) &= -\frac{1}{2}\lambda^2\mathrm{e}^{-2\beta\langle\epsilon\rangle+\beta^2\sigma^2}\left(\mathrm{e}^{\beta^2\sigma^2} - 1\right) \\ &= -\frac{1}{2}\lambda^2\mathrm{e}^{-2\beta\epsilon_\mathrm{eff}}\left(\mathrm{e}^{\beta^2\sigma^2} - 1\right) \end{aligned} \tag{3.49}$$

From Eq. 3.49 explicitly follows that $b < 0$, since the term in parentheses is always greater than unity for non-δ distributions. Also, it is now obvious why distributions with equal ϵ_eff may have different divergence points – b, which causes the divergence of ξ from unity, contains a factor solely dependent on the variance of the Gaussian.

We now focus on our main task in this section, namely, finding an upper limit for the fugacity at which our approximation works well. To that end, we set an the arbitrary upper limit of ξ, which we shall name the convergence criterion $\xi^\star$:

$$\xi = \frac{a}{a+b} = (1+b/a)^{-1} \leq \xi^\star, \tag{3.50}$$

where $a = \ln\left(1+\lambda\left\langle\mathrm{e}^{-\beta\epsilon}\right\rangle\right)$ as in the previous section. Rearranging Eq. 3.50 and keeping in mind that $\xi^\star > 1$, we obtain:

$$-\frac{b}{a} \leq 1 - \frac{1}{\xi^\star} \tag{3.51}$$

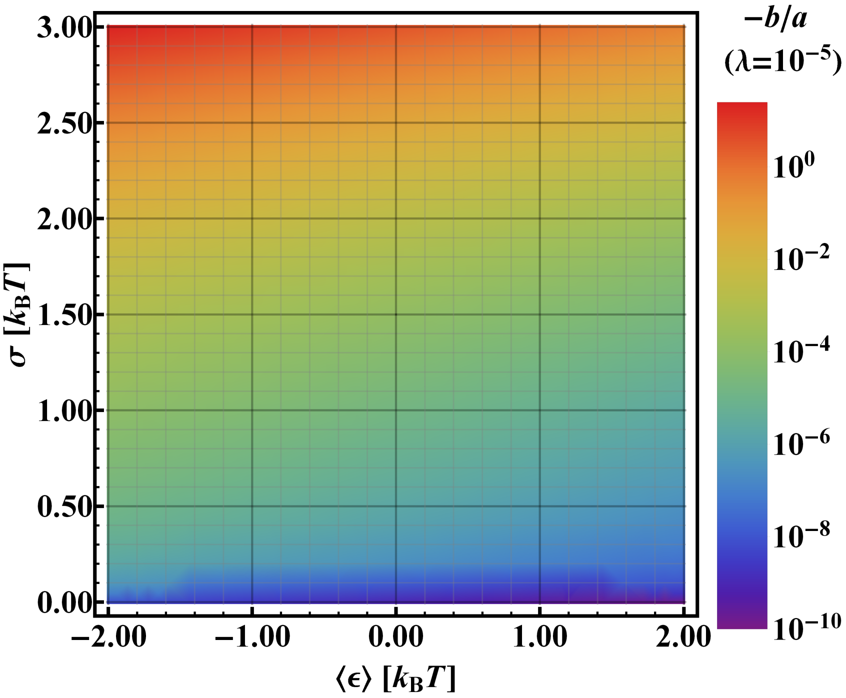


Figure 3.11: Density plot of $-b/a$ for $\lambda = 10^{-5}$ as function of the standard deviation and the average energy.

Before we continue with our derivation, we present the dependence of $-b/a$ on $\langle\epsilon\rangle$ and σ in Fig. 3.11 – as we can see for most parameter combinations this ratio is negligibly small compared to unity. When we, however, work with pathologically wide distributions ($\sigma > 2.5\ k_\mathrm{B}T$) $-b/a$ becomes greater than unity, a clearly non-physical result, which points to $\xi < 0$.

Obtaining an explicit relationship between the fugacity and the upper limit requires calculating b/a. We expand the denominator, a, in a Taylor series around zero up to the second term to keep consistency with b. Thus:

$$\frac{b}{a} = -\frac{1}{2}\frac{\lambda^2 \mathrm{e}^{-2\beta\epsilon_\mathrm{eff}}(\mathrm{e}^{\beta^2\sigma^2}-1)}{\lambda \mathrm{e}^{-\beta\epsilon_\mathrm{eff}}(1-\frac{\lambda}{2}\mathrm{e}^{-\beta\epsilon_\mathrm{eff}})} = -\frac{\lambda \mathrm{e}^{-\beta\epsilon_\mathrm{eff}}}{2-\lambda \mathrm{e}^{-\beta\epsilon_\mathrm{eff}}}(\mathrm{e}^{\beta^2\sigma^2}-1) \approx -\frac{1}{2}\lambda \mathrm{e}^{-\beta\epsilon_\mathrm{eff}}(\mathrm{e}^{\beta^2\sigma^2}-1) \quad (3.52)$$

We are allowed to make the last approximation since we expect $\lambda \mathrm{e}^{-\beta\epsilon_\mathrm{eff}}$ to still be far smaller than 2, when the ratio becomes greater than $\xi^\star$. Plugging Eq. 3.52 into Eq. 3.51 and isolating λ yields:

$$\lambda_\mathrm{max} = 2\Big(1-\frac{1}{\xi^\star}\Big)\frac{\mathrm{e}^{\beta\epsilon_\mathrm{eff}}}{\mathrm{e}^{\beta^2\sigma^2}-1} \quad (3.53)$$

Here, we define the analytical counterpart of the numerical divergence point $\lambda^\star$ introduced in the last section as λ_max

To illustrate the applicability of our result, we compare λ_max to $\lambda^\star$ when a threshold of 1.001 is applied and we are interested in a normal distribution with mean $\langle\epsilon\rangle = -1.3\ k_\mathrm{B}T$ and a standard deviation of $\sigma = 2.3\ k_\mathrm{B}T$. Both values are in good agreement pointing to an upper limit of $\lambda = 1.95 \times 10^{-7}$. If we are interested in a higher limit, say 1.01, the criterion becomes less forgiving to the deviations and underestimates the limiting fugacity – the numerical results point to $\lambda^\star = 2.8 \times 10^{-6}$, while our criterion suggests $\lambda_{\max} = 2.0 \times 10^{-6}$. Nonetheless, our criterion should be more accurate the stricter the upper limit is.

Table 3.3: Comparison of numerical ($\lambda^\star$) and analytical ($\lambda_{\max}$) values for the fugacity, at which ξ is no greater than an arbitrary threshold $\xi^\star$

Distribution	$\xi^\star = 1.001$			$\xi^\star = 1.01$		
	$\lambda^\star$	$\lambda_{\max}$	$\lambda^\star/\lambda_{\max}$	$\lambda^\star$	$\lambda_{\max}$	$\lambda^\star/\lambda_{\max}$
$\mathcal{N}(-1, 6)$	1.01×10^{-7}	9.09×10^{-8}	1.11	1.49×10^{-6}	9.01×10^{-7}	1.65
$\mathcal{N}(-2, 4)$	7.21×10^{-7}	6.77×10^{-7}	1.06	8.76×10^{-6}	6.83×10^{-6}	1.28
$\mathcal{N}(-3, 2)$	5.82×10^{-6}	5.73×10^{-6}	1.02	6.35×10^{-5}	5.67×10^{-5}	1.12

Let us now check if Eq. 3.53 accurately predicts the lowering of $\lambda^\star$ when Gaussians with a constant ϵ_{eff} are compared (Fig. 3.6b). The results are presented in Table 3.3, from which we conclude that our method of estimating $\lambda^\star$ is fairly accurate, that is to say, its accuracy increases upon lowering ξ's threshold and upon narrowing the distribution. Although it may seem $\lambda_{\max}$ fails to predict the divergence point, especially for wide distribution under a more relaxed convergence criterion ($\lambda^\star/\lambda_{\max} = 1.65$ for $\mathcal{N}(-1, 6)$ and $\xi^\star = 1.01$), we should keep in mind that both $\lambda^\star$ and $\lambda_{\max}$ scale exponentially with the chemical potential of the adsorbed species – $\lambda = \exp(\beta\mu)$. Therefore, our criterion underestimates the true chemical potential at the divergence point by no more $\frac{1}{2}k_{\text{B}}T$, while $\mu \approx -13.5k_{\text{B}}T$. Furthermore, given the fact that most research in the field of computational biology aims for order of magnitude estimates, we realise the criterion can be deemed accurate enough to be applicable.

3.4 Transcription factors copy number and site occupation

Up until this moment we have considered only a lattice and one adsorbing protein, without taking into account the copy number of transcription factors within the cell. Since many regulatory proteins are almost insoluble in water, they constantly reside on the DNA strand, thus blocking some of the sites and making them inaccessible for other proteins. This might turn out to be a major flaw in all our considerations up to now, since inaccessible sites cannot be included in Eq. 3.15, thus altering the site distribution and, consequently, the partition function. Delving into this issue requires us to see what the copy number of each protein is (for simplicity we limit ourselves to the *lac* operon), how many sites are blocked upon its adsorption, and whether this is crucial to the accuracy of Eq. 3.27.

We begin our argument with the simplistic case, in which a single protein molecule blocks only one site on the strand. In this case site inaccessibility will not be much of a problem because of the small number of transcription factor molecules (on the order of 10^2) compared to the enormous number of available sites (10^7). To make our explanations clearer, we consider the regulatory proteins involved in lactose control, namely, LacI, RNAP and CRP. Their respective copy numbers within the cell are roughly 20, 1000, and 50 [3]. Even for the most abundant protein, RNAP, site inaccessibility should not be a problem, because the blocked sites form only a minute part of the whole strand, *i.e.* $\sim 10^{-4}$. Here we should point out that 10^7 is actually the length in nucleobases of both DNA strands and not the number of non-specific sites. For highly specific TFs (like LacI) with a low number of specific sites (3), the two quantities are practically identical. When we, however, discuss promiscuous proteins like RNAP, which has roughly 2600 specific sites (promoter sequences)[36, 37], non-specific sites begin to differ from the total number of sites. Still, we do not expect a major discrepancy. From all said thus far, we conclude that if proteins were to block only the site they occupy, site inaccessibility is not a major issue and our model should provide accurate results for the partition function.

This treatment of the problem is, however, rather oversimplified and does not take into account two important microscopic characteristics of protein binding:

1. Sites comprise tens of adjacent bases, *i.e.*, when a protein adsorbs on the DNA it covers a patch of a certain length, which makes it inaccessible for other proteins

2. Adjacent sites, due to base sharing, overlap, that is, moving even only one base upstream or downstream from the initial base of a site yields a different site.

Understanding how many sites a protein blocks upon binding involves realising that if the protein covers 3 bases, then the two bases located immediately upstream from the beginning of the binding site cannot be the initial base of another site because of steric repulsion. The same reasoning applies to the bases comprising the binding site itself. For longer binding sites, these 'forbidden' bases will be more – 7 for four base-long site, 9 for five etc. or $2L - 1$ for a binding site with length L. To see how this affects our derivation, we look into the binding site length of the TFs involved in the *lac* operon [12]:

1. All three specific binding sites of the *lac* repressor are 21 bases long. Therefore, upon adsorption, it blocks a total of 41 potential binding sites.

2. CRP's consensus site is 22 bases long, which means the protein renders 43 sites inaccessible for other CRP molecules

3. Promoter sequences, to which RNAP binds specifically, are usually 41 bases in length, resulting in a total of 81 blocked sites.

Combining this knowledge with our previous reasoning about protein copy number, we arrive at the conclusion that the *lac* repressor molecules block a total of roughly 800 sites, CRPs make another 2200 inaccessible, and RNAP binding takes up 80 000 sites. While the first two proteins block a negligible fraction of the total number of sites ($\approx 10^{-4}$), RNAP renders almost 1% of all sites inaccessible. While this may seem an obvious flaw in all considerations up to now, we believe these blocked DNA sites should follow a similar distribution as the one for the whole genome, that is, site accessibility does not alter the cumulants of the distribution, only the total number of sites, forming it. The reasoning behind this assumption is the same as the one made in Section 2.4.2 – for big enough sets of sites the central limit theorem points to a Gaussian. Proving this assumption, however, requires extensive simulation work, which reaches beyond the scope of this work.

There is one more factor we have not discussed up to now, namely, the presence of proteins not related to lactose metabolism on the DNA strand. There is a wide variety of regulatory proteins within the cell, some of which insoluble in water, which will constantly reside on the strands and consequently block patches of it. Will this be crucial to our model, then? We are inclined to answer 'Important, definitely, but not enough to render our model inapplicable.', because, after all, we are dealing with a dynamic system with proteins constantly moving along the DNA strand. In other words, a site which is blocked at one point in time, will be free in the next and can be occupied by the protein we are modelling.

All that being said, we draw the final conclusion that site accessibility is hardly relevant for TFs with low copy number and short site lengths and it may become a serious problem when dealing with proteins, present in large numbers and with long binding sites. But even then, we expect a statistical behaviour of these inaccessible sites, which, while important, should not crucially hinder our further considerations.

Chapter 4

Comparison with experimental results

After deriving the grand partition function for a one dimensional lattice of sites with binding energy, obeying a statistical distribution, we wish to test the applicability of our model to real-life systems. To that end, we use experimental data, both *in vivo* and *in vitro*, obtained over the course of almost four decades and see how well our model performs. We start with an extensive overview of binding affinity measurements of the *lac* repressor for all interesting binding scenarios, that is, adsorbing onto the three operator sites, binding to the synthetic symmetric operator and non-specific binding. After obtaining binding energies for each specific case, we compare the specific-non-specific binding energy difference to *in vivo* values and see a remarkable (given the timespan and technique diversity) agreement. We also list a few studies dealing with non-specific RNAP-DNA binding. Using the non-specific binding energies of RNAP and LacI, and with the aid of Eq. 2.38, we obtain the standard deviations for these two proteins' binding energy distributions. When we, however, compare these standard deviations to their *in vivo* counterparts, we see rather disturbing discrepancies. To explain them, we turn our attention to protein configuration and the possibility of different modes of binding, an assumption, which is confirmed by NMR structure analyses.

4.1 *In vitro* binding data

Before testing our model, we need to obtain experimental data through a thorough overview of *in vitro* LacI affinity studies, in which we list most findings researchers have made spanning several decades. Our main focus is the binding sites present in the wild type *E. coli* bacterium, *i.e.*, non-specific, O_1, O_2, and O_3. Nevertheless we also mention studies dealing with the perfectly symmetric O_{id}, which, although synthetic, has been commonly used *in vivo* to explain the base sequence-binding affinity relation. Due to scarcity of experimental data we will only touch upon the binding affinity of the repressor for the auxiliary sites. We will also mention the non-specific binding of RNAP to DNA, but we will limit ourselves only to this case, because ϵ_{NS} is the energy via which we can calculate the standard deviation of the binding energy distribution for this protein.

In order to obtain binding energies we make use of Eq. 2.32:

$$K^{\mathrm{obs}} = v_{\mathrm{w}} \exp\big(-\beta\epsilon\big), \tag{4.1}$$

where $v_w = 0.018\ \mathrm{L\,mol^{-1}}$ is the molar volume of water. Finally, we compare our findings with *in vivo* affinity data and determine whether *in vitro* measurements can be used to accurately predict the level of repression within a cell. To that end we are only interested in affinity constants at salt concentration of 0.2 M since this is the *in vivo* ionic condition [38]. Furthermore, when needed, we correct the experimental data for temperature and pH effects, using $K^{\mathrm{obs}}(\mathrm{pH})$ and $K^{\mathrm{obs}}(T)$, reported in literature. Although, *in vivo* temperature is 310 K, we will use data for $T \approx 297$ K, due to the non-linear dependence of K^{obs} on T.

4.1.1 Non-specific binding of LacI and RNAP to DNA

Below, in Table 4.1, we present the non-specific binding affinity results for the *lac* repressor. Although the corresponding binding energies have a rather wide standard deviation, we can safely say ϵ_{ns} is around $-13.5\ k_{\mathrm{B}}T$. All the researchers, listed in the first column, measured K^{obs} as function of salt concentration and obtained a linear dependence, as expected from Record *et al.*'s theory for transcription factor binding to DNA [13]. In the last two columns of the table the slopes, SK^{obs} ("S" stands for slope), and intercepts, ln $K^{\mathrm{obs}}_{1\ \mathrm{M}}$, obtained by these researchers, are listed.

Table 4.1: Non-specific binding of *lac* repressor to DNA. Average binding energy: $-13.44 \pm 0.71\ k_{\mathrm{B}}T$

Author	Year	$K^{\mathrm{obs}}_{\mathrm{NS}}$ [1/M]	ϵ_{NS} [$k_{\mathrm{B}}T$]	T [K]	SK^{obs}	$\ln K^{\mathrm{obs}}_{1\ \mathrm{M}}$
deHaseth *et al.* [39]	1977	1.29×10^4 [a]	-13.48	294	-10.00	-8.52
deHaseth *et al.* [40]	1977	7.54×10^3	-12.94	293	-11.94	-10.29
Revzin *et al.* [41]	1977	1.29×10^4	-13.48	293	-10.28	-7.07
Lohman *et al.* [42]	1980	6.03×10^3	-12.72	293	-10.70	-8.52
Ha *et al.* [25]	1992	3.80×10^4 [b]	-14.56	294	-9.80	-6.45

[a] Recalculated from data obtained at $[\mathrm{M}^+]$=0.13 M and pH=7.7
[b] Recalculated from data obtained at T=277 K and pH=7.9

As we already mentioned, the slope of $K^{\mathrm{obs}}\left([\mathrm{M}^+]\right)$ is a function of the number of ionic contacts formed between the transcription factor and the DNA strand $SK^{\mathrm{obs}} = -0.88Z$, where Z, as before, is the number of contacts formed/ions released into the solution. On the other hand, ln $K^{\mathrm{obs}}_{1\ \mathrm{M}}$ is the extrapolated intercept of this line at salt concentration equal to unity. From the data presented it is evident that when LacI and RNAP bind non-specifically to DNA approximately 12 ionic contacts are formed in both cases. Later in this chapter we will use this value and the averaged value for ln $K^{\mathrm{obs}}_{1\ \mathrm{M}}$ to calculate the standard deviation of the distributions for both proteins.

Table 4.2: Non-specific binding of RNAP to DNA. Average binding energy: $-15.98\ k_{\mathrm{B}}T$

Author	Year	$K^{\mathrm{obs}}_{\mathrm{NS}}$ [1/M]	ϵ_{NS} [$k_{\mathrm{B}}T$]	T [K]	SK^{obs}	$\lg K^{\mathrm{obs}}_{1\ \mathrm{M}}$
deHaseth *et al.* [43]	1978	1.41×10^5	-15.87	294	-10.80	-5.53
Lohman *et al.* [42]	1980	1.73×10^5	-16.08	293	-10.50	-4.84

[a] Recalculated from data obtained at $[\mathrm{M}^+]$=0.13 M and pH=7.7
[b] Recalculated from data obtained at T=277 K and pH=7.9

As it turns out the non-specific binding of RNAP to DNA is not a popular topic of study since we only managed to find two reports of experimental research on the topic.

Nevertheless, the values we calculate for the slope, intercept and binding energy are in good agreement.

4.1.2 *lac* repressor affinity for O_1

In a similar fashion as for the non-specific binding, we present experimental results for the specific binding of the *lac* repressor. We refrain ourselves from listing promoter-RNAP binding energies due to scarcity of reports on the matter and the large database of different promoters.

In most early works on the topic long patches of DNA were used, which resulted in calculating binding affinities on the order of 10^{11} (all results marked with 'a' in Table 4.3). This is due to the length of DNA patches used, which most probably contain secondary operator sites. Auxiliary sites facilitate DNA looping [27], which on the other hand leads to enhanced repression. To eliminate this enhancement, we corrected the data using results obtained by Oehler *et al.* [44], who found that the presence of either of the auxiliary sites causes an approximately 30-fold increase in repression.

Table 4.3: *lac* repressor affinity for O_1. Average binding energy: $-27.90 \pm 0.58\ k_BT$

Author	Year	K_1^{obs} [1/M]	ϵ_1 [k_BT]	T [K]	SK^{obs}	$\ln K_{1\,\text{M}}^{\text{obs}}$
Record *et al.* [45]	1977	2.21×10^{10} [a,b]	-27.83	293	-7.04	15.89
Lohman *et al.* [46]	1978	1.39×10^{10} [a,b]	-27.37	293	-7.04	15.43
Herrick [47]	1980	2.80×10^{10}	-28.07	298	-	-
O'Gorman *et al.* [48]	1980	1.81×10^{10}	-27.48	293	-1.59	21.07
Winter *et al.* [49]	1981	0.96×10^{10} [a]	-27.00	298	-6.76	15.50
Barkley *et al.* [26]	1981	1.55×10^{10} [a]	-27.48	293	-9.30	11.90
Whitson *et al.* [50]	1986	2.29×10^{10}	-27.87	298	- 6.41	13.54
Spotts *et al.* [51]	1991	3.00×10^{10} [c]	-28.14	293	-	-
Chakerian *et al.* [52]	1991	1.66×10^{10} [c]	-27.55	293	-	-
Zhang *et al.* [53]	1993	1.06×10^{10} [d]	-27.10	298	-	-
Bondeson *et al.* [54]	1993	6.86×10^{10} [d]	-28.97	293	-	-
Schlax *et al.* [55]	1995	3.90×10^{10}	-28.40	310	-	-
Swint-Kruse *et al.* [56]	2005	4.21×10^{10} [c]	-28.48	293	-	-
Wilson *et al.* [57]	2007	5.00×10^{10}	-28.65	298	-	-
Romanuka *et al.* [58]	2009	3.05×10^{10} [e]	-28.16	293	-	-

[a] Recalculated taking into account the auxiliary sites [44] and usage of KCl [26]
[b] As calculated by Barkley *et al.* [26]
[c] Recalculated using the linear relation obtained by O'Gorman *et al.* [48]
[d] Recalculated using the linear relation obtained by Whitson *et al.* [50]
[e] Recalculated to correct for pH, ionic conditions, and temperature [26, 50, 59]

Another seemingly contradictory feature of Table 4.3 we should explain is the presence of two linear relationships used to recalculate data when the usage of only one salt concentration was reported (studies carried out after 1991). Since a few researchers (the ones marked with 'c') utilized the experimental set up O'Gorman *et al.* used, we consider it is only natural to use $K^{\text{obs}}\left([\text{M}^+]\right)$ obtained by O'Gorman *et al.* to recalculate the affinity constant. What is more, as evident from the last two columns, the slope and intercept obtained by these researchers differ drastically from the values most often found in literature. Nevertheless, the linear dependence obtained by O'Gorman *et al.* yields reasonable values for the binding energy to O_1

4.1.3 *lac* repressor affinity for the synthetic and symmetric O_{id} and the auxiliary sites O_2 and O_3

Our next task is to present values for all other specific sites, that is O_{id}, O_2, and O_3. Data scarcity requires us to use affinity constant ratios K_1^{obs}/K_i^{obs}, where i is either 2 or 3, signifying that the ratio refers to either O_2 or O_3.

Table 4.4: *lac* repressor affinity for O_{id}, O_2, and O_3. Average binding energies: $\epsilon_{id} = -30.06 \pm 0.65\ k_BT$, $\epsilon_2 = -27.27 \pm 0.27\ k_BT$, and $\epsilon_3 = -21.06 \pm 0.70\ k_BT$

Site	Author	Year	K^{obs} [1/M]	ϵ [k_BT]	T [K]
O_{id}	Ha *et al.* [25]	1992	3.82×10^{11}	-30.69	296
	Frank *et al.* [60]	1997	1.05×10^{11}	-29.39	297
	Tsodikov *et al.* [61]	1999	2.14×10^{11}	-30.10	297
O_2	Winter *et al.* [49]	1981	1.04×10^{10}	-27.08	298
	Romanuka *et al.* [58]	2009	1.53×10^{10a}	-27.46	293
O_3	Winter *et al* [49]	1981	4.11×10^{7a}	-21.55	298
	Romanuka *et al.* [58]	2009	1.53×10^{7a}	-20.56	293

[a] Recalculated via affinity constant ratios

4.1.4 *In vivo* vs. *in vitro* binding data

After reviewing *in vitro* binding data for all commonly studied operator sites, we wish to see if these values agree with binding energies determined *in vivo* – after all, within a living cell there are far more factors which can affect *lac* repressor binding (other proteins adsorbed on DNA, supercoiling of the DNA strand, etc.) compared to the idealised *in vitro* experiment, where only one protein binds to a short strand of DNA. Since *in vivo* experiments yield energy differences rather than absolute values for the binding energies, we need to calculate the binding energy of each operator offset with respect to non-specific binding:

$$\Delta\epsilon_x = \epsilon_x - \epsilon_{ns} \tag{4.2}$$

Table 4.5: Comparison between *in vitro* and *in vivo* binding energies to different *lac* operators

Conditions	$\Delta\epsilon_1$ [k_BT]	$\Delta\epsilon_{id}$ [k_BT]	$\Delta\epsilon_2$ [k_BT]	$\Delta\epsilon_3$ [k_BT]
in vitro	-14.5	-16.6	-13.8	-7.6
in vivo [4]	-15.3	-17.0	-13.9	-9.7

As we compare *in vivo* and *in vitro* binding (Table 4.5), we see that for strongly binding sites there is a good agreement between values obtained via the two approaches. There are, indeed, negligible differences of 0.8 k_BT and less, which might as well be due to the high variance of the data we use. We cannot say the same for weakest operator site, O_3, though, because the two values for O_3 differ by more than 2 $k_\mathrm{B}T$. We can explain this deviation with the scarcity of *in vitro* binding data for this operator and the lack of precision in determining low affinity binding. Nevertheless, the close values obtained for the other three operator sites clearly indicates *in vitro* measurements, although rather simplified compared to the *in vivo* case, can give an accurate estimate of cellular energetics.

4.2 Extracting standard deviations from *in vitro* data

As we already discussed in Section 2.3, we can extract the standard deviation of a binding energy distribution, as long as we know the non-specific binding energy ϵ_ns and a few other biophysical parameters such as the area of hydrophobic contact, the binding energy at 1 M salt solution and the number of ionic contacts formed upon binding. The latter two parameters are easily extracted from binding affinity measurements of $K^\mathrm{obs}\left(\left[\mathrm{M}^+\right]\right)$. Hydrophobic area scales linearly with the number of water molecules released upon binding [62], but since we are working within the salt range, in which $K^\mathrm{obs}\left(\left[\mathrm{M}^+\right]\right)$ is linear, we can neglect the effect from water release [25]. Furthermore, Revzin and von Hippel have proven that non-specific binding is independent of hydrostatic pressure, a finding pointing to the idea that the volume change upon binding is (close to) zero [41]. Thus, we arrive at the conclusion that the process is driven by entropy gain from ion release and hydrophobic interactions do not play a major role. All that being said, we are now ready to extract the standard deviation from the slope and intercept of $K^\mathrm{obs}\left(\left[\mathrm{M}^+\right]\right)$, reported in literature. We start with the *lac* repressor, for which we established that 12 ions are released upon binding and at 1 M salt concentration $\ln K^\mathrm{obs}_\mathrm{1\,M} = -8.17$. We recall Eq. 2.38 and write it down for salt concentration of 1 M:

$$\ln\left(v_\mathrm{w}^{-1}K^\mathrm{obs}_\mathrm{1\,M}\right) = \ln K^\ominus_\mathrm{T} + Z\xi^{-1}\ln\left(\delta\gamma\right) \tag{4.3}$$

Here, we calculate the activity coefficient $\gamma = 0.74$ according to the Debye-Hückel theory for electrolyte solutions, a value in good agreement with [63]. In order to eliminate all electrostatic interactions, we use the same approach as Barkley *et al.* [26], and add $0.2Z$ to both sides of the equation and re-arrange it:

$$\ln K_\mathrm{ne} = \ln\left(v_\mathrm{w}^{-1}K^\mathrm{obs}_\mathrm{1\,M}\right) + 0.2Z - Z\xi^{-1}\ln\left(\delta\gamma\right), \tag{4.4}$$

where $\ln K_\mathrm{ne} = \ln K^\ominus_\mathrm{T} + 0.2Z$ is the purely non-electrostatic (ne) part of $\ln K^\ominus_\mathrm{T}$. Since we neglect hydrophobic effects, attribute all electrostatic interactions to the average binding energy and assume a Gaussian, it is safe to say that ϵ_ne is actually equal to the second term in Eq. 3.29. Therefore, we finally arrive at the expression (the subscript 'L' signifies that this is the standard deviation for the *lac* repressor distribution):

$$\frac{\beta^2\sigma_\mathrm{L}^2}{2} = \ln K^\mathrm{obs}_\mathrm{1\,M} - \ln v_\mathrm{w} - Z\left(\xi^{-1}\ln\left(\delta\gamma\right) - 0.2\right) = 0.76 \quad\Rightarrow\quad \sigma_\mathrm{L} = 1.24\ k_\mathrm{B}T \tag{4.5}$$

Carrying out the same procedure under the same assumptions for RNAP yields a standard deviation of $\sigma_\mathrm{R} = 2.74\ k_\mathrm{B}T$.

It is of particular interest to see how the standard deviation, the average energy, the number of ionic contacts and the site length all relate to each other. To that end we calculate $\langle\epsilon\rangle$, estimate quantities such as average energy per ionic pair and the ratio of the standard deviation to the site size, and present them in Table 4.6.

We see two very interesting trends forming: the average standard deviation contribution per base σ/L is close to 0.065 $k_\mathrm{B}T$ in both cases, a fact we should try and link to the number

Table 4.6: ϵ_{eff}, σ, and $\langle\epsilon\rangle$ and their ratios to the site size L (in base pairs) and the number of ionic contacts Z=12. All quantities are in $k_{\text{B}}T$, except $\sigma/\langle\epsilon\rangle$, which is dimensionless

Protein (L)	ϵ_{eff}	σ	$\langle\epsilon\rangle$	σ/L	$\langle\epsilon\rangle/Z$	$\lvert\sigma/\langle\epsilon\rangle\rvert$
RNAP (41)	-15.98	2.74	-12.22	0.067	1.02	22.4%
LacI (21)	-13.44	1.24	-12.67	0.059	1.06	9.8%

of non-electrostatic contacts formed per position of the adsorption site. On the other hand, the average mean energy contribution per ionic contact formed $\langle\epsilon\rangle/L$ is comparable to 1.05 $k_{\text{B}}T$, a value, which most probably follows from the way we model $\langle\epsilon\rangle$ – after all, we assumed the electrostatic interactions are a constant energy background (that is, the electrostatic component of the binding energy is constant for all sites) and neglected the energy gain from buried hydrophobic area. Carrying out similar calculations for other regulatory proteins may confirm the existence of such trends, which would then enable us to extract structural information (number and/or nature of non-electrostatic contacts) about non-specific complexes just by examining *in vitro* binding data.

4.3 Binding energy distributions via E-matrices

After looking into *in vitro* measurements and estimating the standard deviations of the binding energy distributions for LacI and RNAP, we turn our attention to *in vivo* techniques for genetic activity predictions. Since energy matrices are obtained via *in vivo* methods like *sort-seq*, we will focus on them. It is crucial to note that the elements in the **E**-matrices are not initially calculated in $k_{\text{B}}T$ due to the computational technique used to obtain them. The parallel tempering Monte Carlo technique, described by Kinney *et al.* [30], assumes values in arbitrary units (AU) for each matrix element and iterates until a best fit to *sort-seq* data is obtained. Rescaling the thus-obtained matrix to a $k_{\text{B}}T$ energy scale requires assuming a model for certain energy differences. In other words, for two sites of choice a known binding energy difference in $k_{\text{B}}T$ is compared to the energy difference in AU and a scaling factor is calculated from the ratio of the two differences. This approach is adopted by Brewster *et al.* [64], but in their work they do not take into consideration the width of the RNAP distribution, which ultimately leads to inaccurate scaling.

In this section we essentially follow these researchers' approach and we try to obtain the proper scaling for the RNAP binding energy distribution. We do this by applying the binding energy computation procedure explained in section 2.4.1 to the base sequences of all sites present on both strands of *E. coli* DNA. To that end we use 'The E. coli K-12 MG1655 Genome Sequence ECOLI version 2' kindly provided by [65], which reports the genetic sequence of this particular strain of *E. coli*. We choose K-12 because it is widely considered to be the wild-type bacterium strain ([66, 67] and references therein).

Our next step is to calculate the binding energy, which we do by starting from the 5'-end of the DNA strand, reading out the site sequence from the first to the L^{th} base (here, as usual, L is the length of the site) and feeding it into the **E**-matrix. Then we move on to the second base on the DNA strand, repeat the procedure, after that we slide to the third base and so on until we reach the L^{th} base counting from the 3'-end, which will form the last site. Since we are dealing with double stranded DNA, we carry out the procedure for both strands. Thus we calculate the binding energies for all possible sites. Transforming this set of energies into a distribution requires us to choose a bin size, in which we place sites with close energy. The bin size depends on the minimum and maximum binding energy values, or in other words, for wide distributions it is better to set a wider bin size.

While carrying this procedure for RNAP, however, we are confronted with a rather peculiar result. After hypothesising what the reason for our odd finding is, we move on to LacI, but we encounter a similar controversy. Both results point to the idea of *different modes of binding*, that is, differences in transcription factor conformation and bond nature depending on whether binding is specific or not. This notion is strongly supported by 2D NMR studies of specific and non-specific protein-DNA complex structures.

4.3.1 RNAP binding energies distribution

As we already pointed out, energy matrices entries are initially expressed in an arbitrary energy scale and only then, using either *in vivo* or *in vitro* values, are they converted to $k_\mathrm{B}T$. If we use the **E**-matrix verbatim, we usually obtain a wide distribution centred around a non-zero value. An example of one is the distribution, presented in Fig. 4.1, which is calculated from the original, non-rescaled RNAP **E**-matrix, reported by Brewster *et al.* (Supplementary Materials in [64]).

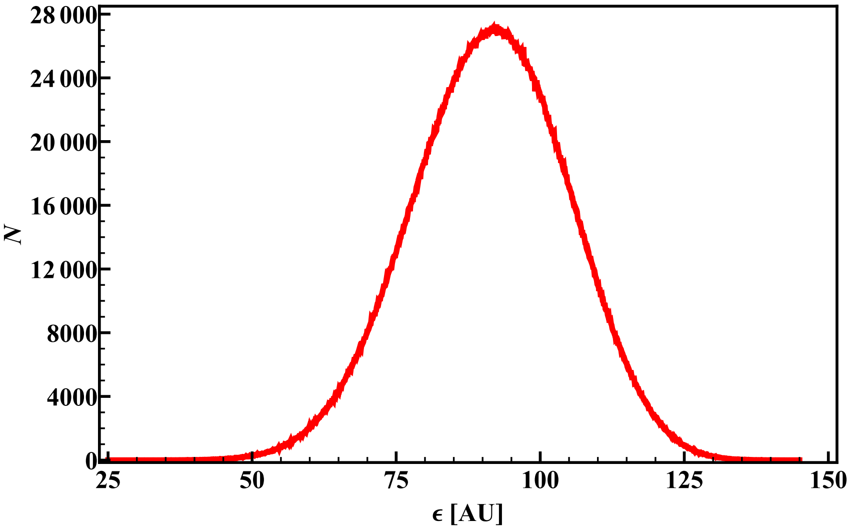


Figure 4.1: RNAP binding energy distribution as calculated with the non-rescaled **E**-matrix reported by [64]

We begin our attempt to scale this distribution by assuming there is a linear relationship between the two energy scales:

$$\epsilon\ [k_\mathrm{B}T] = \frac{\epsilon\ [\mathrm{AU}] - s\ [\mathrm{AU}]}{r\ [\mathrm{AU}/k_\mathrm{B}T]}, \tag{4.6}$$

where s [AU] and r [AU/$k_\mathrm{B}T$] are the shift, expressed in arbitrary units, and the scaling parameter (ratio) in units AU per $k_\mathrm{B}T$, respectively. We also assume these two parameters are constant for all sites present on the DNA strand, thus implying there is a single **E**-matrix, with which the binding energy of any site, be it specific or non-specific, can be calculated. The biophysical implications of this assumption are a lack of discerning between functionally different sites, formation of one and the same type and number of contacts (electrostatic, hydrogen bonding, etc.) regardless of the site, and constant conformation. Our next step is to take the average exponent of the energy in $k_\mathrm{B}T$ according to Eq. 4.6

(for brevity we drop all explicit dimensions safe for the ones distinguishing the two binding energies):

$$\langle \exp\left(-\beta\epsilon\ [k_{\mathrm{B}}T]\right)\rangle = \exp\left(\frac{s\beta}{r}\right)\left\langle \exp\left(-\frac{\beta\epsilon\ [\mathrm{AU}]}{r}\right)\right\rangle \tag{4.7}$$

Since β, s, and r are all constants, we are allowed to isolate the shift from the average exponent. Having pointed that out, we realise the second factor in Eq. 4.7 is nothing more than the MGF of the normally distributed variable ϵ [AU] with the real constant $-\beta/r$. Then, we simply re-write Eq. 3.26 up to the second term with β/r instead of β and all energies expressed in AU:

$$\langle \exp\left(-\beta\epsilon\ [k_{\mathrm{B}}T]\right)\rangle = \exp\left(-\frac{\beta\left(\langle\epsilon\rangle\ [\mathrm{AU}] - s\right)}{r} + \frac{\left(\beta\sigma\ [\mathrm{AU}]\right)^2}{2r^2}\right) = \exp\left(-\beta\epsilon_{\mathrm{eff}}\right), \tag{4.8}$$

an expression from which we can easily isolate the link between the effective energy in $k_{\mathrm{B}}T$ and the cumulants in AU:

$$\epsilon_{\mathrm{eff}}\ [k_{\mathrm{B}}T] = \frac{\langle\epsilon\rangle\ [\mathrm{AU}] - s}{r} - \frac{\beta\left(\sigma\ [\mathrm{AU}]\right)^2}{2r^2} \tag{4.9}$$

As we already established in Chapter 3, the effective energy actually expresses the binding to the non-specific sites and it is the quantity one should use when calculating energy offsets with respect to non-specific binding. To arrive at the scaling factor r we shall use one such offset, namely the binding energy difference between the wild-type *lac* promoter and non-specific sites. Thus, we essentially follow Brewster *et al.*'s approach with the main difference that we use the proper energy offset. A well-established value from literature [27] for this difference is $-5.35\ k_{\mathrm{B}}T$. The last quantity we need in order to calculate r is the binding energy of the wild-type promoter in $k_{\mathrm{B}}T$, which we shall express via the binding energy in AU. Obtaining this value is rather straightforward – knowing the base sequence of the promoter site, we apply the **E**-matrix to the sequence which results in $\epsilon_{\mathrm{wt}}\ [\mathrm{AU}] = 53.4$ AU. We are finally able to construct the equation expressing the energy difference in $k_{\mathrm{B}}T$:

$$\begin{aligned}\epsilon_{\mathrm{wt}}\ [k_{\mathrm{B}}T] - \epsilon_{\mathrm{eff}}\ [k_{\mathrm{B}}T] &= \frac{\epsilon_{\mathrm{wt}}\ [\mathrm{AU}] - s}{r} - \frac{\langle\epsilon\rangle\ [\mathrm{AU}] - s}{r} + \frac{\beta\left(\sigma\ [\mathrm{AU}]\right)^2}{2r^2} \\ &= \frac{\epsilon_{\mathrm{wt}}\ [\mathrm{AU}] - \langle\epsilon\rangle\ [\mathrm{AU}]}{r} + \frac{\beta\left(\sigma\ [\mathrm{AU}]\right)^2}{2r^2} = -5.35\ k_{\mathrm{B}}T\end{aligned} \tag{4.10}$$

Before proceeding with solving it, it is worth noting two important features of Eq. 4.10:

1. Since the shift s is common for all sites, it dropped out of the equation. This not only makes it solvable, because we now have only one variable, but it proves our assumption that the shift is arbitrary and we can easily center the distribution at zero, without affecting the physics of binding.

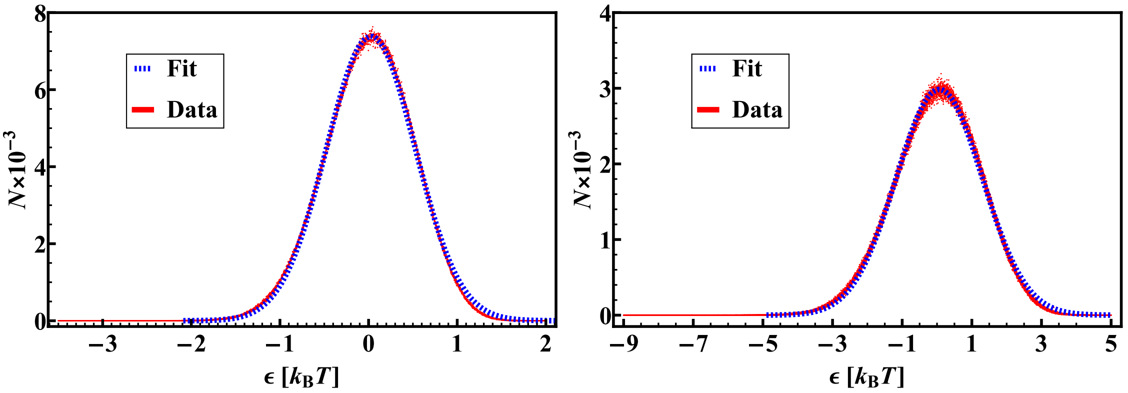
2. Our final equation resembles the one Brewster *et al.* use to calculate the scaling factor. The only difference is the second term in the second line of Eq. 4.10, which takes into account the shape of the distribution.

All that being said, we numerically solve Eq. 4.10 and are confronted with a surprising result – the scale factor is a complex number $r = 3.53 \pm 2.10\mathrm{i}$. Clearly there is something wrong with the assumptions we made in the course of deriving Eq. 4.10. But our only presupposition is the initial hypothesis there is a single set of parameters s and r and one matrix, which can yield binding energies for all sites on the DNA strands. The complex nature of r suggests the falsity of the hypothesis and points to conformational and binding differences depending on the type of site RNAP is bound to. Several reports in the literature support this notion: [68–71] and references therein.

After proving the RNAP **E**-matrix cannot be scaled using the energy difference between a specific and a non-specific site, we speculate this re-scaling should be done using promoter sequence energies. Furthermore, one should choose the promoters in such a way that RNAP's mode of binding to them is one and the same. We will illustrate this in the next subsection in which we try to scale the *lac* repressor energy matrix.

4.3.2 *lac* repressor binding energy distribution

After establishing the RNAP **E**-matrix, reported in literature, is not fitting for obtaining the non-specific site energy distribution, we carry out the same procedure for the *lac* repressor, as well. As before, we start with the non-scaled distribution depicted in Fig. 4.2a, which is calculated using the LacI **E**-matrix, kindly provided by D. Jones of Caltech, Pasadena, CA, USA [72], according to to the procedure described in the previous subsection with a bin size of 0.001. We obtain a distribution with a standard deviation $\sigma_{\mathrm{L}}^{\star} = 0.5$ AU and zero mean. It may seem controversial that the RNAP distribution peaks at roughly 28 000 sites, while for LacI the maximum is at less than 8000 sites. The reason for that is the bin size chosen – since RNAP has a wide distribution a bin size of 0.1 was used.



(a) Non-rescaled LacI distribution

(b) Rescaled LacI distribution

Figure 4.2: (a) LacI binding energy distribution as calculated with the non-rescaled **E**-matrix, kindly provided by D. Jones of Caltech, Pasadena, CA, USA [72] (b) LacI binding energy distribution as calculated with the rescaled **E**-matrix (see text)

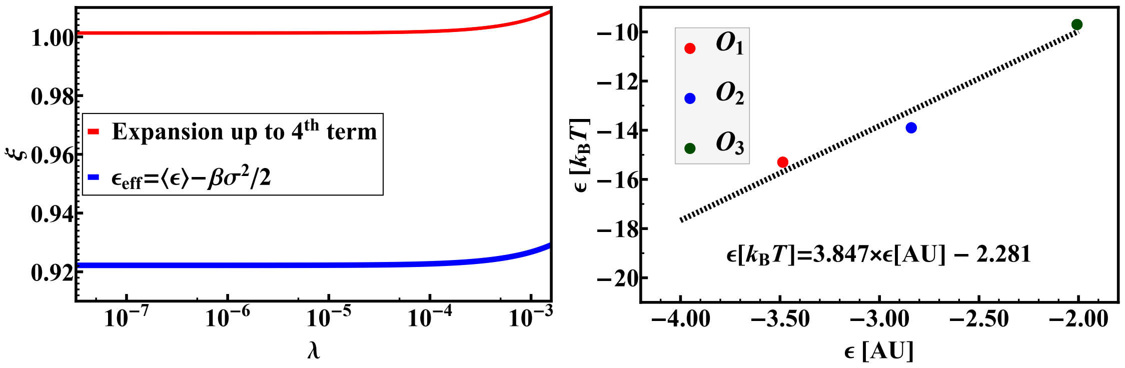
Since the distribution is centred around zero and we already established the position is irrelevant, we can rescale the matrix by simply multiplying it by the ratio of the standard deviation found from *in vitro* data to $\sigma_{\mathrm{L}}^{\star}$. We then slide the re-scaled matrix along both strands of DNA and plot the resulting distribution in Fig. 4.2b. Fitting the distribution to a Gaussian yields $\sigma = 1.24\ k_{\mathrm{B}}T$, which is indeed identical to the one found from *in vitro* measurements. It is now interesting to see how well our analytical formula predicts the partition function for this distribution. To that end we plot $\xi\left(\lambda\right)$ and present the function in Fig. 4.3a.

Even at first glance a major inconsistency that draws our attention is evident, namely, the abnormally low ξ value, which is smaller than unity regardless of the fugacity. This result most probably stems from the value of ϵ_{eff} we used: instead of calculating the first few cumulants of the distribution, we assumed it is a Gaussian (in agreement with our reasoning in section 2.3) and used non-linear fitting to obtain the mean and variance. Checking if this is indeed correct requires the calculation of the cumulants according to Eq. 2.51. The results, presented in Table 4.7, clearly point to the fact that we are not dealing with normally distributed energies, evident from the presence of a skew and excess kurtosis.

Table 4.7: First four cumulants of the distribution in Fig 4.2b obtained via non-linear fitting to a Gaussian and Eq. 2.51. The two approaches yield different effective energies.

Approach	κ_1	κ_2	κ_3	κ_4	ϵ_{eff}
Non-linear fit	0.064	1.539	0.000	0.000	-0.706
Eq. 2.51	0.012	1.512	-0.305	-0.165	-0.788

From the values presented in Table 4.7 it becomes clear why the fit parameters fail to give a reasonable value of ξ. Since we assumed a Gaussian, we ruled out the option of having cumulants of order higher than two, thus overestimating the effective energy by roughly 10%, which leads to an underestimation of the partition function by approximately 9% (these estimates can be obtained by plugging in the values for the effective energy in Eq. 3.30 and expanding the logarithm). Using the correct ϵ_{eff} yields $\xi(\lambda) \approx 1$ for the fugacity range we are interested in, as is evident from the red line in Fig. 4.3a.



(a) LacI distribution is skewed and platykurtic

(b) Scaling of specific site energies

Figure 4.3: (a) $\xi(\lambda)$ for the binding energy distribution in Fig 4.2b. Fitting the distribution to a Gaussian and using the fit parameters to calculate the effective energy leads to an underestimation of Ξ_{Th}. Using Eq. 3.29 up to the fourth term yields the proper value for the function. (b) Linear relationship between the two energy scales for the three operator sites.

Having said that, we should ask ourselves what causes this contradictory result. Both the negative value for the excess kurtosis and the presence of a skew dismiss the possibility of observing a Student's t distribution rather than a Gaussian since the t distribution has no skew and the kurtosis for 20 degrees of freedom is 0.375. Upon closer examination of Fig. 4.2 we see that the left tail is slightly heavier compared to the one predicted by the fit, while the right tail is overestimated by the fit. This points to the idea that the matrix generates more favourable sites than necessary. To explain this finding we recall the conclusions we drew from our attempt to scale the RNAP **E**-matrix. Do these results for LacI imply different modes of binding for this protein, as well? To answer this question we investigate the matrix even further and, at least for now, we neglect the non-Gaussian nature of the distribution and simply assume it is indeed normal with an effective energy equal to the one following from the detailed expansion of the CGF.

The next step in our analysis is to check the applicability of the rescaled matrix to specific sites. As we already mentioned, we make the naïve assumption that one matrix is able to predict both specific and non-specific binding, a notion, which we test by calculating the binding energy of O_1, O_2, and O_3 using the re-scaled matrix. The values we determine are off compared to the ones known from literature – using this approach we obtain $O_1 = -8.63\ k_BT$, $O_2 = -7.03\ k_BT$, and $O_3 = -4.97\ k_BT$, which are clearly incorrect. In order to compute the true scaling relation for the specific sites we again assume there is a linear relation between the energy in arbitrary units and on the k_BT scale, *i.e.*, $\epsilon\,[k_BT] = (\epsilon\,[\mathrm{AU}] - s)/r = r^{-1} \times \epsilon\,[\mathrm{AU}] - s/r$. The easiest way to calculate the slope and the intercept of this line is to plot a point with coordinates $(\epsilon[\mathrm{AU}], \epsilon[k_BT])$ for each operator site and then make a linear regression (Fig. 4.3b). As we suspected, the scaling parameters, which this approach yields, differ considerably from the simple scaling relation we employed to obtain the proper standard deviation of the distribution. Not only does the slope of the line differ from our scaling ratio (3.85 vs. 2.48), but we now have an intercept, as well, which indicates a non-zero mean. While the latter is not a major problem since we are able to shift the distribution however we want (the mean is arbitrary), the former discrepancy poses a serious question, namely 'Is there, indeed, only one matrix, with which binding can be modelled?'

In a final attempt to gain some more insight into the problem, we combine *in vitro* and *in vivo* measurements. To that end we turn our attention to *in vitro* binding of the *lac* repressor to non-specific sites with well-defined base sequence and we are mainly interested in the interaction with the alternating polymer poly(dG-dC)· poly(dG-dC). There are two distinct binding sites, *i.e.*, a 21-base long site starting with G (G-site) and a complementary one starting with C (C-site):

$$\begin{aligned}\text{G-site}&: \ \text{GCGCGCGCGCGCGCGCGCGCG}\\ \text{C-site}&: \ \text{CGCGCGCGCGCGCGCGCGCGC}\end{aligned}$$

Since we are working with more than one site we must calculate the effective energy of the poly(dG-dC)· poly(dG-dC) strand $\epsilon_{\text{GC-GC}}$ using Eq. 3.43

$$\beta\epsilon_{\text{GC-GC}} = -\ln\left[\frac{1}{2}\exp(-\beta\epsilon_{\mathrm{G}}) + \frac{1}{2}\exp(-\beta\epsilon_{\mathrm{C}})\right] \tag{4.11}$$

To obtain the two binding energies we resort to the energy matrix instead of using *in vitro* measurements because we need separate values for the two sites, a piece of information which *in vitro* experiments can not provide. A reasonable question here should be "Why not use single-stranded DNA?". The reason for that is the secondary structure of double-stranded DNA which affects binding affinity.

Although the sequence is well-defined, it differs considerably from the specific site sequences, which makes it indistinguishable from a random-sequence site to the LacI. Therefore, we expect the values, calculated via the **E**-matrix to lie within roughly one standard deviation away from the average energy, an educated guess, which is actually confirmed:

$$\epsilon_{\mathrm{G}} \approx \langle\epsilon\rangle \tag{4.12}$$

$$\epsilon_{\mathrm{C}} \approx \langle\epsilon\rangle + 1.2\sigma \tag{4.13}$$

We will express all energies via binding distribution parameters, that is $\langle\epsilon\rangle$ and σ, without explicitly writing them down in $k_{\mathrm{B}}T$ values, to illustrate how this approach can also be used for calibration of energy matrices. Plugging Eq. 4.12 and 4.13 into Eq. 4.11 we calculate the effective energy of our model lattice:

$$\begin{aligned}\beta\epsilon_{\text{GC-GC}} &= -\ln\frac{1}{2} - \ln\left[\exp\left(-\beta\langle\epsilon\rangle\right) + \exp\left(-\beta\langle\epsilon\rangle - 1.2\beta\sigma\right)\right] \\ &= \ln 2 - \ln\left[\mathrm{e}^{-\beta\langle\epsilon\rangle}\left(1 + \mathrm{e}^{-1.2\beta\sigma}\right)\right] \\ &= \ln 2 + \beta\langle\epsilon\rangle - \ln\left(1 + \exp\left(-1.2\beta\sigma\right)\right)\end{aligned} \tag{4.14}$$

Our next step is to compare the affinity constants for a truly non-specific site and for the lattice under consideration. Sadly, there is no such investigation made at least to our knowledge. Riggs *et al.* [73] however point out that poly dA· poly dT binds LacI roughly 8 times stronger than poly(dG-dC)· poly(dG-dC). On the other hand Revzin and von Hippel [41] have found that poly dA· poly dT binds to LacI seven-fold tighter compared to non-specific DNA. Wang *et al.* [74], however, report an affinity ratio of only 3. It is fair to say, then, that poly dA· poly dT binds the repressor 5 times more tightly than non-specific DNA. Therefore:

$$\frac{K_{\text{GC-GC}}}{K_{\text{A-T}}} \approx \frac{1}{8} \text{ and } \frac{K_{\text{ns}}}{K_{\text{A-T}}} \approx \frac{1}{5} \Rightarrow \frac{K_{\text{ns}}}{K_{\text{GC-GC}}} \approx 1.6 \tag{4.15}$$

If we now take the natural logarithm on both sides of the equation and keeping in mind that the equilibrium constant is a function of the binding energy (Eq. 2.32), we obtain:

$$\ln\frac{K_{\text{ns}}}{K_{\text{GC-GC}}} = \ln\frac{v_{\mathrm{w}}\exp(-\beta\epsilon_{\text{eff}})}{v_{\mathrm{w}}\exp(-\beta\epsilon_{\text{GC-GC}})} = \beta\epsilon_{\text{GC-GC}} - \beta\epsilon_{\text{eff}} = \ln 1.6 \tag{4.16}$$

Substituting the last line of Eq. 4.14 and Eq. 3.29 up to the second term into Eq. 4.16, we obtain:

$$\frac{\beta^2\sigma^2}{2} - \ln(1 + \exp(-1.2\beta\sigma)) = \ln 0.8 \tag{4.17}$$

This equation has two roots, one of which is non-physical. Therefore, we can conclude that the standard deviation of the distribution is:

$$\sigma \approx 0.69\ k_{\mathrm{B}}T \tag{4.18}$$

Unfortunately, this value is off by a factor of roughly 2 compared to σ_{L}, a result which we already expected given the inaccurate predictions for the operator binding energies. As a last resort measure let us see what are the possible bounds for the standard deviation by using 1/3 and 1/7 for the affinity constant ratio $K_{\mathrm{NS}} \times K^{-1}_{\text{GC-GC}}$. Following the same procedure as described above we find $\sigma_{\text{upper}} = 1.04\ k_{\mathrm{B}}T$ when the ratio is 1:3 and $\sigma_{\text{lower}} = 0.2\ k_{\mathrm{B}}T$ for $K_{\mathrm{NS}} \times K^{-1}_{\text{GC-GC}} = 1/7$. Although σ_{upper} comes close to σ_{L} we are inclined to question this result since Wang *et al.* are using a conceptually different experimental procedure. Furthermore, their experiments are carried out at $T > 310$ K, and as we know from the studies of Frank *et al.* [60] and deHaseth *et al.* [39] repressor affinity diminishes upon increasing the temperature. With this final result we come to the conclusion that the repressor **E**-matrix is not fitting for non-specific site binding energy predictions. If it were, it should have been able to yield the proper standard deviation in our last set of calculations.

Let us now ponder what the cause for this discrepancy may be, while keeping in mind we already encountered a similar situation when dealing with RNAP. Recalling our explanation for RNAP, we ask ourselves if there is more than one binding mode for the *lac* repressor, as well. To answer this question we turn our attention to NMR studies done by R. Boelens and his group, who investigated the problem of the repressor binding to operator and non-specific sites thoroughly. In 2004 they managed to obtain results which clearly show major differences in both the contacts formed between DNA and the TF, and in the conformation of the DNA-TF complexes (Fig. 7 in [75], also [58, 76, 77] and the references therein). Their study supports our assumption of differentiated binding modality and confirms hydrophobic contacts do not play a major role in non-specific binding. Furthermore, it suggests what kind of interactions may be the cause of the variability of the binding energy, that is, hydrogen bonds formed between the LacI and the DNA backbone.

We now arrive at the conclusion that energy matrices currently reported in literature can not be used to calculate the standard deviation of the non-specific energy distribution due to the different number, geometry and nature of the formed complexes. This, although implied by the way these matrices are obtained, is not*a priori* evident since some of these matrices are used to model the specific-non-specific binding energy difference [64]. Obtaining **E**-matrices from a wide variety of non-specific sites using similar (if not the same) experimental techniques will shed more light on the matter and plausibly help us to better understand what makes a specific site specific.

Chapter 5

Discussion and Outlook

In the preceding few chapters we discussed a variety of topics spanning from genetics (the description of the *lac* operon), through biophysics (the linear relationship between affinity and salt concentration) and statistical mechanics (applications of the grand canonical ensemble in genetic regulation modelling) to statistics and tensor calculus. Then, we looked into a problem rarely discussed, namely the distribution of regulatory protein-DNA binding energies, and with the help of this theoretical background we managed to develop a simple, yet powerful theory, which is able to accurately calculate the grand canonical partition function of a lattice like DNA. Finally, we tested our theory with experimental data from both *in vitro* and *in vivo* measurements and drew two conclusions with biophysical relevance. On one hand, we extracted the standard deviation of binding energies for LacI and RNAP and saw it scales with the number of bases these proteins occupy when adsorbed. On the other hand, the idea of binding mode differentiation explained the controversial results we faced while trying to apply energy matrices to non-specific binding. Now is the proper time to evaluate our findings in a more general manner and seek links to phenomena outside the scope of biophysics. Furthermore, we ought to discuss what future research can focus on to provide more insight into the topic of binding energy distributions.

The effective energy ϵ_{eff} and the standard deviation σ

As we derived in Chapter 3, we are able to collapse a whole distribution, with all its different energies and numbers of sites having these energies into a single value, which, within reasonable approximation, reflects the entire information the distribution carries:

$$\epsilon_{\text{eff}} = \sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^{n-1}}{n!} = \langle\epsilon\rangle - \frac{\beta\sigma^2}{2} + \frac{\beta^2\gamma_1\sigma^3}{6} - \frac{\beta^3\gamma_2\sigma^4}{24} + \mathcal{O}\left(\sigma^5\right) \tag{5.1}$$

It is worth noting that, although in rigorous mathematical terms the ϵ_{eff} is an expansion in the cumulants κ_n, it can also be deemed an expansion in the standard deviation. We simply need to define a 'standardised' mean $\bar{\langle\epsilon\rangle} = \langle\epsilon\rangle/\sigma$ (we put standardised in quotations since this is not the usual meaning of the word in statistics). Then:

$$\epsilon_{\text{eff}} = \bar{\langle\epsilon\rangle}\sigma - \frac{\beta\sigma^2}{2} + \frac{\beta^2\gamma_1\sigma^3}{6} - \frac{\beta^3\gamma_2\sigma^4}{24} + \mathcal{O}\left(\sigma^5\right) \tag{5.2}$$

What is more, as we demonstrated in Chapter 4 the average energy is completely arbitrary and we can set it to zero, thus reducing the effective energy to a function of the standard deviation alone. This reasoning now explains why we put so much effort in obtaining it – σ turns out to be the parameter, which properly scales all cumulants. On top of that, in our initial model for the distribution, we assumed the energies obey a Gaussian. Setting the average to zero renders the effective energy a function of only one variable, σ.

But what is it that makes the standard deviation so important in biophysical terms? To answer this question we ponder its physical meaning by considering a normal distribution. 3σ intervals to the left and right of the average cover 99.7% of the entire Gaussian. If we now split these 6σs in smaller intervals (or bins) of fixed size we obtain, say, N bins. Let us now think about two other Gaussians, one wider and one narrower than our initial distribution. If we carry out the same task for both of them, while keeping the bin size the same, the number of bins we obtain is different – wider distributions yield more bins than narrower, or in other words when the standard deviation increases, the number of bins goes up, as well. With this simple example we wish to illustrate protein's ability to discern between sites with respect to their energy, *i.e.*, the wider the distribution, the larger the number of sites the protein can distinguish between. To support our reasoning we recall the standard deviations we calculated for the two regulatory proteins we studied, RNAP and LacI, with $\sigma_{\mathrm{R}} = 2.74\ k_{\mathrm{B}}T$ and $\sigma_{\mathrm{L}} = 1.24\ k_{\mathrm{B}}T$, respectively. RNAP, being a multivalent protein, with thousands of promoter sites, needs to be able to distinguish between sites with a similar base sequence, which ultimately means similar energies. On the other hand, the repressor, whose functional sites are only a few, does not need to be that selective when it comes to non-specific sites. To fully grasp the reasoning behind this, we need to recall which energy contribution to ϵ_{NS} we have not modelled up to now. While electrostatic interactions can be described via the DNA-protein binding theory devised by Record *et al.* and hydrophobic interactions are hardly relevant for non-specific complexes, modelling hydrogen bonding, which is the sequence-specific part of ϵ_{NS}, is not straightforward. We can speculate that σ^2 is a measure of the number of hydrogen bonds a protein forms when binding to a NS site. Then highly specific proteins like LacI will have a smaller σ due to the small number of specific contacts they form with NS sites. On the other hand, RNAP will 'recognise' parts of a NS site as specific and will be prone to form more hydrogen bonds. All that being said we realise the standard deviation is most probably characteristic of the protein, not the DNA it binds to, and it is an indirect measure of its specificity. To verify this conclusion, however, an extensive study of many proteins binding non-specifically to different DNA strands should be carried out, a task which is beyond the scope of our work and which we leave to future research.

Energy matrices and modes of binding

In our endeavour to verify the obtained standard deviations for RNAP and the *lac* repressor, we stumbled upon a rather intriguing inconsistency, namely the energy matrices, commonly used in literature to model specific site binding energies, are unfitting for non-specific sites. This finding led us to the idea of binding modality, that is, proteins bind differently to functionally different sites, a notion supported by NMR studies. This result is important not only because it theoretically and independently predicts an experimental fact, but suggests what further experiments should be done. The first and by far the most important is obtaining non-specific energy matrices, which accurately predict protein-DNA interactions and possibly answer some of the questions posed by the standard deviation dependence on site size. On the other hand 2D NMR studies, which are a powerful structure determination tool, can be used to examine the interactions in a wide variety of non-specific complexes, such as the ones formed between TFs and alternating polymeric DNA (like the one studied in Section 4.3). These two approaches combined will help us understand the function of the non-specific sites even better and gain even more insight into the problem of DNA recognition processes.

Chapter 6

Conclusion

The recent development of a variety of high-throughput experimental techniques for genetic processes investigation has lead to a boom in the field of quantitative biology, and more specifically in gene regulation studies. The accumulation of *in vivo* and *in vitro* results on the matter has naturally paved the way towards theoretical prediction of a cell's behaviour under certain conditions. That being said, the rapid expansion of the newly emerged field of computational biology seems only natural. Using statistical mechanics and biophysics as tools, researchers are developing models, which aim to either explain currently available results or predict experimental outcomes.

Though extensively studied for more than 5 decades, the *lac* operon still remains the system of choice when it comes to genetic regulation modelling. The great number of possible regulatory frameworks provides fertile ground for theoretical studies. While most researchers focus on the problem of lactose control by examining how regulatory proteins affect the process of transcription, few have looked into the interactions between these proteins and non-regulatory DNA. The current work deals with exactly this topic and is an attempt to devise a theory of non-specific DNA-protein interactions.

The main results of our work can be summarized as follows:

1. The problem of binding energy distribution for non-specific adsorption sites on a one dimensional lattice was discussed under the framework of the grand canonical ensemble. With the help of the cumulant-generating function an effective energy was defined, with which one quickly and accurately predicts the value of the partition function for a distribution of choice.

2. Using *in vitro* affinity measurements, we extracted the average energies and the standard deviations of the RNAP and *lac* repressor binding energy distributions.

3. The inconsistency between standard deviations calculated from *in vivo* and *in vitro* experiments led us to the concept of binding mode differentiation.

Although we acknowledge some of our results and conclusions require more in-depth investigation, we should point out that our research is still in progress and we strive to gain a better understanding of the matter. Nevertheless, we hope the current work provides novel insight into transcription factor-DNA interactions and serves as inspiration for future work in the field, both theoretical and experimental.

Acknowledgements

So ends my story on genetic regulation, seemingly posing more questions than it has answered. Or in other words, the yellow brick road did not lead me to the Emerald City, where Oz the Great and Almighty answers all my question, but rather to a crossroads, each path being a new exciting exploration. And much like Dorothy, I had quite a few trusted companions, who guided me and supported me throughout my journey.

First and foremost, I would like to thank Willem for giving me the opportunity to delve into a topic, which, before meeting him, I would have brushed off as being 'simply biology, therefore boring'. A topic, which greatly expanded my scientific horizons and proved to be an interdisciplinary challenge, which tested all my wits. His supervision always guided me towards the answers I was looking for (sometimes even towards those I did not know I was seeking) and is one of the main reasons this thesis came to be the way it is. Last, but not least, I am thankful for the chances he gave me to present my work to a wider scientific community.

But it was not only Willem who ushered me in the right direction - Jasper, through his daily supervision, helped immensely in the shaping of this work. The long discussions, both on scientific and not so scientific topics, the extensive explanations of how LaTeX, Mathematica and Illustrator work (and the reasons why they did not from time to time) and his constant support in battling problems were what often drove me forward. I would dare say I had a supervisor for no more than two or three months, who, then, turned into a friend, tending to all my questions and scientific struggles. Thank you for being the Samwise to my quest, Jasper!

I should also mention all the other PhD students who helped me, in one way or the other. Special thanks go to Samia, who has been acting as sort of my 'project unrelated questions' advisor for the last few months - apart from the various pieces of advice about the Honours master, Erasmus application and what it is to be a PhD student, I am thankful for the velouté recipe and the oh-so endearing nickname, which always puts a smile on my face. Also, I would like to thank Chris for his help in developing the binding energy distribution code, answering every Mathematica question I had for him and conceiving the idea of the convergence criterion. A huge thanks goes, also, to all people I shared a laugh with over a beer at borrels, barbecues, and other social events - you have made my stay at FCC enjoyable and memorable.

Отправям специални благодарности и към всички, които ме подкрепяха от разстояние - някои от 200, други от 2200 км. Към моите "белгийски" приятели - благодаря ви, че сте моят safe haven, винаги когато ме налегне носталгията (а и не само). Ще продължавам да идвам в Льовен и да нападам лютеницата на Ромина, да знаете! À tous mes amis в България - две минаха, остават още четири, след това отново ще има съботно шкембе с бира. Благодаря и на Крис и семейството му, които ме приеха тъй топло - планината наистина ражда Хора. Но тези, които имат най-големи заслуги за това да пиша тези редове, са родителите ми - знам, че е трудно да изпратиш единствения си син далеч от род и родина, но си струваше! Благодаря, че ми дадохте възможността да го разбера!

Last but not least I would like to thank the people with whom I interacted the most during the past one year - my fellow master students, and in particular Daniël, Joeri and Stijn. Being part of 'Shaken, not vortex mixed' is tons of fun and I am always looking forward to the next game night. Your help whenever I encounter problems of any sort means a lot to me and I will cherish all the fun (and most often downright ridiculous) moments we shared. You will always be welcome in Delft, Bulgaria, or wherever I may roam!

Thank you all for being there for me!

Zeist, the Netherlands, 28th Feb, 2016

Bibliography

[1] F. Jacob and J. Monod, *Journal of molecular biology*, 1961, **3**, 318–356.

[2] J. M. Berg, J. L. Tymoczko and L. Styerl, *Biochemistry*, W.H. Freeman & Company, 5th edn., 2002.

[3] R. Phillips, J. Kondev and J. Theriot, in *Physical Biology of the Cell*, Garland Science, New York, 2008.

[4] H. G. Garcia and R. Phillips, *Proceedings of the National Academy of Sciences*, 2011, **108**, 12173–12178.

[5] U. Gerland, J. D. Moroz and T. Hwa, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12015–12020.

[6] M. Sheinman, O. Bénichou, Y. Kafri and R. Voituriez, *Reports on progress in physics. Physical Society (Great Britain)*, 2011, **75**, 026601.

[7] Y. M. Wang, R. H. Austin and E. C. Cox, *Physical Review Letters*, 2006, **97**, 1–4.

[8] A. Esadze, C. A. Kemme, A. B. Kolomeisky and J. Iwahara, *Nucleic Acids Research*, 2014, **42**, 7039–7046.

[9] A. M. Sengupta, M. Djordjevic and B. I. Shraiman, *Proceedings of the National Academy of Sciences*, 2002, **99**, 2072–2077.

[10] Y. Zhao, D. Granas and G. D. Stormo, *PLoS Computational Biology*, 2009, **5**, 1000590.

[11] M. Lässig, *BMC bioinformatics*, 2007, **8 Suppl 6**, S7.

[12] M. Razo-Mejia, J. Q. Boedicker, D. Jones, A. DeLuna, J. B. Kinney and R. Phillips, *Physical biology*, 2014, **11**, 026005.

[13] M. T. Record, C. F. Anderson and T. M. Lohman, *Quarterly reviews of biophysics*, 1978, **11**, 103–178.

[14] M. Lewis, *Comptes Rendus - Biologies*, 2005, **328**, 521–548.

[15] S. Busby and R. H. Ebright, *Journal of molecular biology*, 1999, **293**, 199–213.

[16] C. L. Lawson, D. Swigon, K. S. Murakami, S. A. Darst, H. M. Berman and R. H. Ebright, *Current Opinion in Structural Biology*, 2004, **14**, 10–20.

[17] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev and R. Phillips, *Current Opinion in Genetics and Development*, 2005, **15**, 116–124.

[18] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman and R. Phillips, *Current Opinion in Genetics and Development*, 2005, **15**, 125–135.

[19] J. P. Peters, N. a. Becker, E. M. Rueter, Z. Bajzer, J. D. Kahn and L. J. Maher, *Quantitative methods for measuring DNA flexibility in vitro and in vivo*, 2011, vol. 488, pp. 287–335.

[20] J. Q. Boedicker, H. G. Garcia and R. Phillips, *Physical Review Letters*, 2013, **110**, 1–5.

[21] J. Landman, Private communication, 2015.

[22] I. Langmuir, *Journal of the American Chemical Society*, 1918, **40**, 1361–1403.

[23] F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips and W. K. Kegel, *Physical Review Letters*, 2014, **113**, 1–5.

[24] G. Manning, *Accounts of Chemical Research*, 1979, **12**, 443–449.

[25] J. H. Ha, M. W. Capp, M. D. Hohenwalter, M. Baskerville and M. T. Record, *Journal of molecular biology*, 1992, **228**, 252–264.

[26] M. D. Barkley, P. a. Lewis and G. E. Sullivan, *Biochemistry*, 1981, **20**, 3842–3851.

[27] T. Kuhlman, Z. Zhang, M. H. Saier and T. Hwa, *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**, 6043–6048.

[28] M. Perros and T. a. Steitz, *Science*, 1996, **274**, 1929–1930.

[29] G. D. Stormo, T. D. Schneider, L. Gold and A. Ehrenfeucht, *Nucleic Acids Research*, 1982, **10**, 2997–3011.

[30] J. B. Kinney, A. Murugan, C. G. Callan and E. C. Cox, *Proceedings of the National Academy of Sciences*, 2010, **107**, 9158–9163.

[31] D. Pollard, *Introduction to Statistics*, Yale University Lecture Notes, 1998.

[32] H. D. Ursell, *Proceedings of the Cambridge Philosphical Society*, 1927, **23**, 685–697.

[33] L. T. DeCarlo, *Psychological Methods*, 1997, **2**, 292–307.

[34] G. Lebanon, *Technical Notes on Statistics*, Georgia Institute of Technology Lecture Notes, 2006.

[35] M. Abramowitz and I. A. Stegun, in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1972, p. 930.

[36] K. E. Rudd, *RegulonDB 9.0*, 2015.

[37] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chávez, A. Hernández-Alvarez, E. Morett and J. Collado-Vides, *Nucleic Acids Research*, 2013, **41**, 203–213.

[38] Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble and P. H. Von Hippel, *Proceedings of the National Academy of Sciences*, 1977, **74**, 4228–32.

[39] P. L. DeHaseth, T. M. Lohman and M. T. Record, *Biochemistry*, 1977, **16**, 4783–4790.

[40] P. L. DeHaseth, C. A. Gross, R. R. Burgess and M. T. Record, *Biochemistry*, 1977, **16**, 4777–4783.

[41] A. Revzin and P. H. von Hippel, *Biochemistry*, 1977, **16**, 4769–4776.

[42] T. M. Lohman, C. G. Wensley, J. Cina, R. R. Burgess and M. T. Record, *Biochemistry*, 1980, **19**, 3516–3522.

[43] P. L. DeHaseth, T. M. Lohman, R. R. Burgess and M. T. Record, *Biochemistry*, 1978, **17**, 1612–1622.

[44] S. Oehler, E. R. Eismann, H. Krämer and B. Müller-Hill, *The EMBO journal*, 1990, **9**, 973–979.

[45] M. T. Record, P. L. DeHaseth and T. M. Lohman, *Biochemistry*, 1977, **16**, 4791–4796.

[46] T. M. Lohman, P. L. DeHaseth and M. T. Record Jr, *Biophysical Chemistry*, 1978, **8**, 281–294.

[47] G. Herrick, *Nucleic Acid Research*, 1980, **8**, 3721–3728.

[48] R. B. O'Gorman, M. Dunaway and K. S. Matthews, *Journal of Biological Chemistry*, 1980, **255**, 10100–10106.

[49] R. B. Winter and P. H. von Hippel, *Biochemistry*, 1981, **20**, 6948–6960.

[50] P. A. Whitson, J. S. Olson and K. S. Matthews, *Biochemistry*, 1986, **25**, 3852–8.

[51] R. O. Spotts, a. E. Chakerian and K. S. Matthews, *Journal of Biological Chemistry*, 1991, **266**, 22998–23002.

[52] A. E. Chakerian and K. S. Matthews, *Journal of Biological Chemistry*, 1991, **266**, 22206–22214.

[53] X. Zhang and P. A. Gottlieb, *Biochemistry*, 1993, **32**, 11374–11384.

[54] K. Bondeson, Å. Frostell-Karlsson, L. Fägerstam and G. Magnusson, *Analytical Biochemistry*, 1993, **214**, 245–251.

[55] P. J. Schlax, M. W. Capp and M. T. Record, *Journal of molecular biology*, 1995, **245**, 331–350.

[56] L. Swint-Kruse, H. Zhan and K. S. Matthews, *Biochemistry*, 2005, **44**, 11201–11213.

[57] C. J. Wilson, H. Zhan, L. Swint-Kruse and K. S. Matthews, *Biophysical Chemistry*, 2007, **126**, 94–105.

[58] J. Romanuka, G. E. Folkers, N. Biris, E. Tishchenko, H. Wienk, A. M. J. J. Bonvin, R. Kaptein and R. Boelens, *Journal of Molecular Biology*, 2009, **390**, 478–489.

[59] M. I. Moraitis, H. Xu and K. S. Matthews, *Biochemistry*, 2001, **40**, 8109–8117.

[60] D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski and M. T. Record, *Journal of molecular biology*, 1997, **267**, 1186–1206.

[61] O. V. Tsodikov, R. M. Saecker, S. E. Melcher, M. M. Levandoski, D. E. Frank, M. W. Capp and M. T. Record, *Journal of molecular biology*, 1999, **294**, 639–655.

[62] M. G. Fried, D. F. Stickle, K. V. Smirnakis, C. Adams, D. MacDonald and P. Lu, *Journal of Biological Chemistry*, 2002, **277**, 50676–50682.

[63] M. Utsumi, *Nippon Kagaku Kaishi*, 1937, **58**, 297–300.

[64] R. C. Brewster, D. L. Jones and R. Phillips, *PLoS Computational Biology*, 2012, **8**, e1002811.

[65] K. E. Rudd, *EcoGene 2.0*, `http://www.ecogene.org/old/SequenceDownload.php`, 2007.

[66] F. R. Blattner, G. I. Plunkett, A. C. Bloch, T. N. Perna, V. Burland, M. Riley, J. Collado-Vides, D. J. Glasner, K. C. Rode, F. G. Mayhew, J. Gregor, W. N. Davis, A. H. Kirkpatrick, A. M. Goeden, J. D. Rose, B. Mau and Y. Shao, *Science*, 1997, **277**, 1453–1462.

[67] K. F. Jensen, *Journal of Bacteriology*, 1993, **175**, 3401–3407.

[68] P. L. DeHaseth, M. L. Zupancic and M. T. Record, *Journal of Bacteriology*, 1998, **180**, 3019–3025.

[69] R. M. Saecker, O. V. Tsodikov, K. L. Mcquade, P. E. S. Jr, M. W. Capp and M. T. Record, *Journal of molecular biology*, 2002, **2836**, 649–671.

[70] K. Gezvain and R. Landick, *The bacterial chromosome*, 2004, 283–296.

[71] A. Robinson and A. M. van Oijen, *Nature reviews. Microbiology*, 2013, **11**, 303–15.

[72] D. Jones, Private communication, 2015.

[73] A. D. Riggs, S. Lin and R. D. Wells, *Proceedings of the National Academy of Sciences*, 1972, **69**, 761–764.

[74] A. C. Wang, A. Revzin, A. P. Butler and P. H. von Hippel, *Nucleic Acid Research*, 1977, **4**, 1579–1593.

[75] C. G. Kalodimos, R. Boelens and R. Kaptein, *Chemical reviews*, 2004, **104**, 3567–3586.

[76] C. G. Kalodimos, N. Biris, A. M. J. J. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens and R. Kaptein, *Science (New York, N.Y.)*, 2004, **305**, 386–389.

[77] E. R. Zuiderweg, R. M. Scheek, R. Boelens, G. W. Van and R. Kaptein, *Biochimie*, 1985, **67**, 707–715.

Appendices

Appendix A

List of Abbreviations and Symbols

Abbreviations

RNAP	RNA polymerase
LacI	*lac* repressor
CRP	cAMP receptor protein
CAP	catabolite activator protein, another name for CRP
TF	transcription factor
O_1	main operator site of LacI
O_2 and O_3	auxiliary operator sites of LacI
MGF	moment-generating function
CGF	cumulant-generating function
PDF	probability density function
CDF	cumulative density function
D	DNA binding site
DTF	DNA-TF complex

General thermodynamics

Symbol	**Units**	**Description**
T	K	Absolute temperature
k_B	$\mathrm{J\,K^{-1}}$	Boltzmann constant
N_A	$\mathrm{mol^{-1}}$	Avogadro's number
$\beta = (k_\mathrm{B}T)^{-1}$	$\mathrm{J^{-1}}$	Inverse thermal energy
μ	$k_\mathrm{B}T$	Chemical potential
$\lambda = \mathrm{e}^{\beta\mu}$	–	Fugacity
ΔG	$k_\mathrm{B}T$	Change of Gibbs' free energy
K	–	Thermodynamic equilibrium constant
x_i	–	Molar fraction of species 'i'
a_i	–	Activity of species 'i'
γ_i	–	Activity coefficient of species 'i'

Statistical mechanics

Symbol	**Units**	**Description**
ϵ	$k_\mathrm{B}T$	Binding energy
$\Delta\epsilon$	$k_\mathrm{B}T$	Specific-non-specific binding energy difference

ϵ_{eff}	$k_{\mathrm{B}}T$	Effective binding energy (Eq. 3.29)
Φ	$k_{\mathrm{B}}T$	Grand potential
N_{total}	–	Total number of non-specific sites on DNA
P	–	Total number of molecules adsorbed on DNA
p	–	Promoter occupation number
Z_{i}	–	Statistical weight of state 'i', canonical partition function
Ξ	–	Grand canonical partition function
Ξ_{Th}	–	Grand canonical partition function, according to Eq. 3.11
Ξ_{Num}	–	Grand canonical partition function, according to Eq. 3.1
$\xi = \ln \Xi_{\mathrm{Th}} \times \ln \Xi_{\mathrm{Num}}^{-1}$	–	Ratio of $\ln \Xi_{\mathrm{Th}}$ to $\ln \Xi_{\mathrm{Num}}$
$\xi^{\star}$	–	Convergence criterion, maximum acceptable value of ξ
$\lambda^{\star}$	–	Divergence point, λ at which ξ exceeds $\xi^{\star}$
$\lambda_{\max}$	–	Theoretical counterpart of the divergence point (Eq. 3.53)

Electrochemistry and Biophysics

Symbol	**Units**	**Description**
e	C	Elementary charge
ε_0	$\mathrm{F\,m^{-1}}$	Vacuum permittivity
ε	–	Permittivity of the medium
λ_{B}	m	Bjerrum length
b	m	Axial charge spacing on DNA
$\xi_{\mathrm{el}} = \lambda_{\mathrm{B}} \times b^{-1}$	–	Characteristic scale of DNA charge-charge interaction
$q_{\mathrm{eff}} = e \times \xi_{\mathrm{el}}^{-1}$	C	Effective charge of a phosphate group on DNA
I	$\mathrm{mol\,L^{-1}}$	Ionic strength
$\kappa = \sqrt{8\pi N_{\mathrm{A}} I \lambda_{\mathrm{B}}}$	$\mathrm{m^{-1}}$	Debye-Hückel screening parameter
$\tilde{\kappa} = \kappa \times b$	–	Reduced axial charge separation
$\delta = \sqrt{8\pi N_{\mathrm{A}} \lambda_{\mathrm{B}} c^{\ominus}}$	$\mathrm{L^{1/2}\,mol^{1/2}}$	$\tilde{\kappa}$ scaled with the ionic strength
v_w	$\mathrm{L\,mol^{-1}}$	Molar volume of water
$c_w = v_w^{-1}$	$\mathrm{mol\,L^{-1}}$	Molar concentration of water
K^{obs}	$\mathrm{L\,mol^{-1}}$	Observed affinity constant
Z	–	Number of charges in the active center of the TF

Statistics

Symbol	**Units**	**Description**
$\langle\epsilon\rangle$	$k_{\mathrm{B}}T$	Average binding energy
σ	$k_{\mathrm{B}}T$	Standard deviation
$\mathcal{N}(-\beta\langle\epsilon\rangle, \beta^2\sigma^2)$	–	Mathematical notation for a normal distribution with mean $-\beta\langle\epsilon\rangle$ and standard deviation $\beta\sigma$
μ_n	$(k_{\mathrm{B}}T)^n$	n^{th} moment of a distribution
κ_n	$(k_{\mathrm{B}}T)^n$	n^{th} cumulant of a distribution
$M_{\epsilon}(-\beta)$	–	MGF of a random variable ϵ with parameter $-\beta$
$K(-\beta)$	–	CGF of a random variable ϵ with parameter $-\beta$
$\varphi(-\beta) = M(-\mathrm{i}\beta)$	–	Characteristic function with parameter $-\beta$
γ_1	$k_{\mathrm{B}}T$	Skewness of a distribution
γ_2	$k_{\mathrm{B}}T$	Excess kurtosis of a distribution
s	$k_{\mathrm{B}}T$	Shape parameter of a Laplace distribution

Appendix B

Dependence of the affinity constant on salt concentration

We start our derivation by recalling Eq. 2.37 and the two approximations made in Section 2.3:

$$\ln(v_{\mathrm{w}}^{-1}K^{\mathrm{obs}}) = \ln K_{\mathrm{T}}^{\ominus} + \ln a_{\mathrm{M}^+}^{-\nu} + \ln \frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} \tag{B.1}$$

With the two approximations we can estimate the sodium ions released upon binding. Prior to binding, the DNA strand has N phosphate groups with θN Na^+ condensed on them. After binding, the number of phosphates is reduced by the charge of the transcription factor and the number of condensed ions drops to $\theta(N - Z)$. Therefore, the number of ions released is proportional to the charge in the active center of the protein:

$$\nu = \theta N - \theta(N - Z) = Z\theta \tag{B.2}$$

In order to tackle the last term in Eq. B.1, we recall the physical meaning of the activity coefficient - it represents the energy excess $\mu^{\star}$ in an non-ideal system, compared to the ideal case:

$$\gamma = \mathrm{e}^{-\beta\mu^{\star}} \tag{B.3}$$

Therefore, it is more convenient to work with energy excess, rather than activity coefficients. To that end we re-write the last term in Eq. B.1 in terms of chemical potentials:

$$\ln \frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} = \beta(\mu_{\mathrm{DTF}}^{\star} - \mu_{\mathrm{D}}^{\star} - \mu_{\mathrm{TF}}^{\star}) \tag{B.4}$$

We can now break down the energy excess for each species into several contributions:

1. $\mu_{\mathrm{TF}}^{\star}$ can be split into two parts - one for the active site $\mu_{\mathrm{as}}^{\star}$ and one for the remaining part of the protein $\mu_{\mathrm{r}}^{\star}$. Thus we assume that major changes take place only at the active site. Furthermore, we assume the non-ideality of the active site is proportional to its charge and arises from non-ideality of the solution itself:

$$\mu_{\mathrm{as}}^{\star} = Z\mu_{\mathrm{M}^+}^{\star} \tag{B.5}$$

2. We suppose non-ideality in the DNA strand is solely due to the electrostatic interactions between the phosphate groups, which we have already expressed

in Eq. 2.24:

$$\beta\mu^{\star}_{\mathrm{D}} = \beta G^{\mathrm{el}} = -\frac{N}{\xi_{\mathrm{el}}}\ln(\tilde{\kappa}) = -\frac{N}{\xi_{\mathrm{el}}}\left[\frac{\ln\overline{[\mathrm{M}^+]}}{2} + \ln\delta\right], \tag{B.6}$$

where $\delta = b\sqrt{8\pi N_{\mathrm{A}}\lambda_{\mathrm{B}}c^{\ominus}}$ and $\overline{[\mathrm{M}^+]} = [\mathrm{M}^+]/c^{\ominus}$ is the molar concentration of cations divided by the standard molar concentration $c^{\ominus}$ of 1 M.

3. We model $\mu^{\star}_{\mathrm{DTF}}$ as the sum of $\mu^{\star}_{\mathrm{TF}}$ and $\mu^{\star}_{\mathrm{D}}$ after binding. We presume $\mu^{\star}_{\mathrm{TF}}$ loses the contribution from the active site due to binding but retains $\mu^{\star}_{\mathrm{r}}$. On the other hand, from Manning's theory we know the excess energy of the DNA is proportional to the number of charges on the strand. After binding we have effectively neutralised Z phosphates from the strand. In that sense, we estimate a decrease in the DNA component of $\mu^{\star}_{\mathrm{DTF}}$ equal to the fraction of phosphate groups neutralised upon binding:

$$\mu^{\star}_{\mathrm{DTF}} = \mu^{\star}_{\mathrm{r}} + \mu^{\star}_{\mathrm{D}} - \frac{Z}{N}\mu^{\star}_{\mathrm{D}} \tag{B.7}$$

We are now ready to re-write Eq. B.4 in terms of the contributions we recognized:

$$\begin{aligned}\ln\frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}} &= \beta\left(\mu^{\star}_{\mathrm{r}} + \mu^{\star}_{\mathrm{D}} - \frac{Z}{N}\mu^{\star}_{\mathrm{D}} - \mu^{\star}_{\mathrm{D}} - \mu^{\star}_{\mathrm{r}} - \mu^{\star}_{\mathrm{as}}\right) = -\beta\left(\mu^{\star}_{\mathrm{as}} + \frac{Z}{N}\mu^{\star}_{\mathrm{D}}\right)\\ &= Z\left(-\beta\mu^{\star}_{\mathrm{M}^+} + \frac{\ln\overline{[\mathrm{M}^+]}}{2\xi_{\mathrm{el}}} + \frac{\ln\delta}{\xi_{\mathrm{el}}}\right)\\ &= Z\ln\gamma_{\mathrm{M}^+} + \frac{Z}{\xi_{\mathrm{el}}}\left(\frac{\ln\overline{[\mathrm{M}^+]}}{2} + \ln\delta\right)\end{aligned} \tag{B.8}$$

Thus Eq. B.1 transforms to:

$$\begin{aligned}\ln(v_{\mathrm{w}}^{-1}K^{\mathrm{obs}}) &= \ln K^0_{\mathrm{T}} + \underbrace{Z\left(\ln\gamma_{\mathrm{M}^+} + \frac{\ln\overline{[\mathrm{M}^+]}}{2\xi_{\mathrm{el}}} + \frac{\ln\delta}{\xi_{\mathrm{el}}}\right)}_{\ln\frac{\gamma_{\mathrm{D}}\gamma_{\mathrm{TF}}}{\gamma_{\mathrm{DTF}}}} \underbrace{- Z\left(\theta\ln\gamma_{\mathrm{M}^+} - \theta\ln\overline{[\mathrm{M}^+]}\right)}_{-\nu\ln a_{\mathrm{M}^+}}\\ &= \ln K^0_{\mathrm{T}} + Z\ln\overline{[\mathrm{M}^+]}\left(\frac{1}{2\xi_{\mathrm{el}}} - \theta\right) + Z\ln\gamma_{\mathrm{M}^+}(1-\theta) + Z\frac{\ln\delta}{\xi_{\mathrm{el}}}\\ &= \ln K^0_{\mathrm{T}} + Z\ln\overline{[\mathrm{M}^+]}\left(\frac{3}{2\xi_{\mathrm{el}}} - 1\right) + Z\frac{\ln\gamma_{\mathrm{M}^+}}{\xi_{\mathrm{el}}} + Z\frac{\ln\delta}{\xi_{\mathrm{el}}}\\ &= \ln K^0_{\mathrm{T}} + Z\frac{\ln(\delta\gamma_{\mathrm{M}^+})}{\xi_{\mathrm{el}}} - Z\ln\overline{[\mathrm{M}^+]}\left(1 - \frac{3}{2\xi_{\mathrm{el}}}\right)\end{aligned} \tag{B.9}$$