Protein-DNA Interaction, Random Walks and Polymer Statistics

by

Michael Slutsky

Submitted to the Physics Department
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2005

Author . . . .
Physics Department
May 19, 2005

Certified by . . . .
Leonid A. Mirny
Assistant Professor
Thesis Supervisor

Accepted by . . . .
Thomas J. Greytak
Professor of Physics
Associate Department Head for Education

Protein-DNA Interaction, Random Walks and Polymer Statistics

by

Michael Slutsky

Submitted to the Physics Department
on May 19, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In Part I of the thesis, a general physical framework describing the kinetics of protein-DNA interaction is developed. Recognition and binding of specific sites on DNA by proteins is central for many cellular functions such as transcription, replication, and recombination. In the process of recognition, a protein rapidly searches for its specific site on a long DNA molecule and then strongly binds this site. Earlier studies have suggested that rapid search involves sliding of the protein along the DNA. I treat sliding as a one–dimensional diffusion in a sequence–dependent rough energy landscape. I demonstrate that, despite the landscape's roughness, rapid search can be achieved if one–dimensional sliding is accompanied by three–dimensional diffusion. I estimate the range of the specific and nonspecific DNA-binding energy required for rapid search and suggest experiments that can test the proposed mechanism. It appears that realistic energy functions cannot provide both rapid search and strong binding of a rigid protein. To reconcile these two fundamental requirements, a search-and-fold mechanism is proposed that involves the coupling of protein binding and partial protein folding. In this regard, I propose an effective energy landscape that incorporates longitudinal (sliding) and transversal (folding) dynamics. I also study the influence of finite correlation length in the binding potential profile on the one–dimensional diffusion. The proposed mechanism has several important biological implications for search in the presence of other proteins and nucleosomes, simultaneous search by several proteins, etc.

In Part II, I analyze the behavior of random walks in presence of smooth manifolds. First, I treat a random walk (or gaussian polymer) confined to a half-space using a field–theoretic approach. Using path integrals, I derive basic scaling relations and the probability distribution function for arbitrary coupling strength between the polymer and the manifold. Next, I consider self–avoiding polymers attached to the tip of an impenetrable probe. The scaling exponents γ_1 and γ_2, characterizing the number of configurations for the attachment of the polymer by one end, or at its midpoint, are shown to vary continuously with the tip's angle. These apex exponents are calculated analytically by ϵ–expansion and compared to numerical simulations in three dimensions. I find that when the polymer can move through the attachment point, it typically slides to one end; the apex exponents quantify the entropic barrier to threading the eye of the probe.

Thesis Supervisor: Leonid A. Mirny
Title: Assistant Professor

Acknowledgments

People who know what they are doing are rare, people who know why they are doing it are even rarer. Those who are able to convey this knowledge to others are almost unique. I was lucky to work with two people blessed with these virtues.

First, I wish to thank Professor Leonid Mirny, my thesis supervisor, who introduced me to biology – a field with more questions than answers and more exceptions than rules. The former makes biology attractive, but the latter may be the reason most theoretical physicists stay away from it. Leonid patiently guided me through intermittent periods of excitement and skepticism and stoically tolerated my whims, fussiness and occasional procrastination. Together, we managed to devise a couple of quantitative models, which I hope are both somewhat biologically relevant and still simple and general enough to retain some "physical appeal."

I also had the privilege of working with Professor Mehran Kardar. His outstanding ability, scientific and pedagogical, together with personal kindness and openness, turned my work with him into truly one of a kind experience. Whatever he does, he does it with taste and elegance and I hope I was able to absorb at least a little bit of each of these qualities.

At various stages of my research, I met many wonderful scientists. It is a bit ironic that I had to cross the ocean to start collaborating with Prof. Yacov Kantor of Tel-Aviv University, whom I knew when I was a student there and who helped me to get started in Boston. I also thank another collaborator – Roya Zandi – and wish her best of luck! I benefited from stimulating discussions with Alex van Oudenaarden, Shamil Sunyaev, Yariv Kafri, Antoine van Oijen, Babis Kalodimos and Bill Bialek. It was nice to be a part of the budding MIT Biophysics group and a slightly larger Condensed Matter Theory group; I thus thank the groups' graduate students and postdocs – Dmitry Abanin, Murat Acar, Roman Barankov, Dan Greenbaum, Dmitry Novikov, Juan Pedraza, Max Vavilov, Martin Zwierlein and many others – for many interesting discussions about physics, biology and life in general.

If there is anything I really like about MIT, it's the abundance and the quality of the courses. I had an especially great time working my tail off during the course on critical phenomena by Prof. Kardar and courses in Applied Math taught by Prof. Bazant and Dr. Margetis. I also enjoyed helping Mehran and Leonid to build a special course, Statistical Physics in Biology.

In my opinion, one of Leonid's principal achievements is the selection of wonderful labmates, for it's with them that I spent most of my time. Ivan Adzhubey, Vincent Bérubé, Juhi Chandalia, Carlos Goméz, Grisha Kolesov, Georgia Kringos, Joe Levine, Victor Spirin – thank you, ladies and gentlemen! Without you, my stay at MIT would be much less enjoyable. Special thanks are due to Zeba Wunderlich – for expressing interest in the projects I started and "picking up the torch" and for her patience in correcting my oral and written English.

A wise man once told me: "In the end, what is it you're left with? Friends and experiences..." I am deeply grateful to my dear friend, Dmitry Pushin, whom I've known since our secondary school years and with whom I was reunited at MIT. His support and understanding throughout one of the toughest periods of my life have been invaluable. It would also be unjust not to mention all those who periodically distracted themselves from professional pressures and created a warm and cheerful circle of friends. I therefore thank Igor Berezovsky, Hana and Gregory Chockler, Mark and Bozhena Gandelman, Eva and Eugene Kaminsky, Alex Schechter, Sonia and Serge Vilvovsky, Victoria Zaychik and Alex Baranchuk for helping me to survive the windy Boston climate.

My family was always there to support and encourage me. I thank my parents, who are largely responsible for bringing me to this point. I probably wouldn't have become a physicist at all, were it not for my father – where else should I mention this if not in the doctoral thesis? I thank Jane, the greatest of all sisters, my grandmother Zelda Gerchikova and my grandfather Moshe Slutsky for their love and trust in me. I am also very grateful to my more distant family, who are in fact very close people – Fanya and Alex Fabrikant and Ola and Alex Kogan.

Finally, I wish to thank my wife, Inna, and my children, Rachel and Ariel – for loving me and helping me to maintain a healthy perspective on life.

I dedicate this thesis to the loving memory of my late grandmother, Emilia Levinson, and my late grandfather, Alexander Gerchikov. The best part of what I've become, I owe to them.

לָכֵ֛ן הִנְנִי־שָׂ֥ךְ אֶת־דַּרְכֵּ֖ךְ בַּסִּירִ֑ים וְגָֽדַרְתִּי֙
אֶת־גְּדֵרָ֔הּ וּנְתִיבוֹתֶ֖יהָ לֹ֥א תִמְצָֽא׃

הוֹשֵׁעַ ב׃ח

Therefore, behold, I will hedge up thy way with thorns, and I will make a wall against her, that she shall not find her paths.

Hosea 2:8

Contents

I Biophysics of Protein-DNA Interaction 15

1 Kinetics of protein-DNA interaction: the search speed – stability paradox 17

1.1 Introduction: "Faster than diffusion" search 17

1.2 The Model 18

1.2.1 Search time 18

1.2.2 Protein-DNA energetics 19

1.3 Diffusion in a sequence-dependent energy landscape 22

1.3.1 The MFPT. 23

1.3.2 The time constant. 25

1.4 Optimal time of 3D/1D search 26

1.5 Non-specific binding 29

1.6 Speed versus stability 31

2 Folding and binding 34

2.1 The two-mode model 34

2.2 The two profiles 37

2.2.1 The numerics 38

2.2.2 Robustness of target location 40

2.3 Effective energy landscape 41

2.3.1 The Continuos Model 42

2.3.2 Dynamics on 2D Landscape 45

2.4 Biological implications 50

2.4.1 Specificity “for free”: kinetics vs. thermodynamics. 50
2.4.2 Coupling of folding and binding in molecular recognition . . . 50

3 The long reach of DNA heterogeneity. 52

3.1 Diffusion in a correlated random potential 52
3.1.1 The model 52
3.1.2 Mean First Passage Time 54
3.2 Typical versus average 55
3.2.1 Potential profile generation 56
3.2.2 Quantifying fluctuations 57
3.3 Examples from biology 61
3.3.1 DNA bending 61
3.3.2 DNA translocation through a nanopore 64

II Random Walks and Polymer Statistics 67

4 Diffusion in a half–space via path integrals 70

4.1 Example: unconstrained Gaussian chain 70
4.2 Anchored polymer: The partition function 72
4.3 Probability distribution 75

5 Apex exponents for polymer–probe interactions 79

5.1 Introduction 79
5.2 Analytic calculations 82
5.3 The numerics 86
5.3.1 Entropic competition 87
5.3.2 Dimerization 87
5.4 Conclusions 90

6 Concluding remarks and future directions 91

6.1 Protein-DNA interactions 91

6.1.1 Relevance to a bacterial cell 91

6.1.2 Implications for eukaryotes 94

6.1.3 Sequence, energy and folding 96

6.1.4 DNA conformation effects 97

6.2 Random walks and polymers 98

A Protein-DNA interaction energetics 102

A.1 Equilibrium Model for Interaction Energy 102

A.1.1 Berg - Von Hippel theory 103

A.1.2 Energy Gap 105

B Diffusion in a Half-Space - Classical results 107

B.1 Diffusion and random walks 107

B.1.1 Some history 107

B.1.2 Boundary conditions 108

B.1.3 Method of images 109

B.1.4 Counting walks on a lattice 109

List of Figures

1-1 Spectrum of binding energy for three different transcription factors and the Gaussian approximation (solid line). 21

1-2 The Model Potential. 22

1-3 (a) Dependence of the search time on the non-specific binding energy. (b) The parameter space. The dashed line corresponds to optimal parameters σ and E_{ns} connected by Eq. (1.39). 31

1-4 (a) Stability on the protein-DNA complex on the cognate site measured as the fraction of time in the bound state at equilibrium. (b) Optimal search time as a function of the binding profile roughness, for the range of parameters $10^{-4}\text{sec} \leq \tau_{3d} \leq 10^{-2}\text{sec}$, $10^{-10}\text{sec} \leq \tau_0 \leq 10^{-6}\text{sec}$. . . . 33

2-1 Cartoon demonstrating the two-mode search-and-fold mechanism. Top: search mode, bottom: recognition mode (a) two conformations of the protein bound to DNA: partially unfolded (top) and fully folded (bottom). (b) The binding energy landscape experienced by the protein in the corresponding conformations. (c) The spectrum of the binding energy determining stability of the protein in the corresponding conformations. 36

2-2 Mean target location time as a function of on-site conformational transition time for various values of σ_s. 40

2-3 Probability of conformational transition before dissociation as a function of on-site conformational transition time. 41

2-4 (a) A schematic view of a protein attached to DNA: harmonic potential combined with sequence–specific interaction. (b) A three–dimensional view of the energy landscape; note a specific site at $x = 63$ 43

2-5 (a) Spectrum of protein–DNA binding energies; (b) Potential profiles for various binding sites with different $U_s(x)$. 44

2-6 Average target location time t_{loc} as a function of D_x/D_z for $M = 10^3$ bp measured directly from MC simulations (squares) and estimated from the number of "jumps" (triangles). Inset: scaling of t_{loc} with genome size M for $\kappa = D_x/D_z = 10^2$; the slope of the line is equal 1. 47

2-7 Specific mode occupancy η as a function of D_x/D_z for a $M = 10^5$ bp stretch of *E. coli* genome. Different symbols correspond to different values of Δz. Inset: fraction of time spent at different sites for $M = 10^6$ bp, $\kappa = 16$; circles designate real binding sites. 48

3-1 Pseudoenergy probability density for a profile of length $L = 10000$, with $\sigma = 1.0$, $\xi_c = 20.0$. Insets: (a) Typical potential profile; (b) Potential profile correllator $g(r) = 1/2\langle[U(x) - U(x+r)]^2\rangle$; the averaging was performed over 1000 profile realizations. 56

3-2 Mean First Passage Times: typical versus average. Thick solid line is the result of averaging over 1000 realizations of correlated potential profiles with $\beta\sigma = 2.0$, $\xi_c = 20.0$. 58

3-3 MFPT standard deviation for $\beta\sigma = 1.0$ for correlated and uncorrelated potential profiles. 59

3-4 Probability density functions for MFPT calculated for 100,000 uncorrelated profile realizations at $\beta\sigma = 2.0$. 60

3-5 (a) Prokaryotic transcription factor sliding; (b) Nucleosome repositioning. 62

3-6 (a) Energy of local elastic deformation and (b) Potential profile correlator, as calculated from the data supplied by the server BEND.IT for a segment of *E. coli* genome. The deformed DNA sequence is assumed to be of length $L = 15$ bp. 64

3-7 ssDNA transport through the nanopore; on the right: charge density $q(x)$ and correlator $g(r) = \langle[\delta U(x) - \delta U(x+r)]^2\rangle/(2\langle\delta U^2(x)\rangle)$ as a function of the coordinate r. 65

4-1 The (normalized) scaling function $F(\hat{g}, \hat{z})$. Different curves are labeled by corresponding values of $\hat{g}$. The larger $\hat{g}$, the closer is the scaling function to a linear dependence. 78

5-1 Configurations of a polymer near an obstacle: (a) attached to the apex of a planar sector of angle α; (b) threaded through the eye of a cone with apex semi–angle β. 81

5-2 Diagrams contributing to renormalization of g to second order (a–c); to $\mathcal{Z}$ in first order (d,e) at the apex of a slice; and to $\mathcal{Z}_2$ in first order (f) at the eye of a conic needle. 83

5-3 The probability distributions $p(N_1)$ for two non-interacting segments of lengths N_1 and $N - N_1$ attached to the apex of a planar slice for diffrent values of angle α. The curves are the result of 10^9 MC steps for $N = 2000$. 88

5-4 Extrapolated values of the exponents $\Delta\gamma_1 = \gamma_0 - \gamma_1$ (circles) and $\Delta\gamma_2 = \gamma_0 - \gamma_2$ (diamonds) as a function of sector angle α from "entropic competition" (open symbols), and dimerization (full symbols). Error bars represent statistical uncertainties of individual estimates of the exponents, as well as the uncertainty in the extrapolation $N \to \infty$. . 89

6-1 Effect of DNA conformation on the effective diffusion distance: (a) Single globule; (b) Multi-domain conformation. 98

A-1 Energy spectrum of 20 bp sequence with unit discrimination energy. The squares are the results of computer simulation; the solid line is a quadratic fit. 105

A-2 Energy spectrum and energy profile for E. coli purine repressor ($PurR$). The weight matrix was built by analyzing 35 known binding sites for $PurR$. 106

B-1 The reflection principle. 110

Part I

Biophysics of Protein-DNA Interaction

The complex transcription machinery of cells is primarily regulated by a set of proteins, *transcription factors* (TFs), that bind DNA at specific sites [1, 2]. Every TF can have from one to several dozens of specific sites on the DNA. Upon binding to a specific site, the TF forms a stable protein-DNA complex that can either activate or repress transcription of nearby genes, depending on the actual control mechanism [3]. Fast and reliable regulation of gene expression requires (1) fast (∼1-10 sec) search and recognition of the specific site (referred to as the *target* or *cognate* site below) out of 10^6 - 10^9 possible sites on the DNA, and (2) stability of the protein-DNA complex ($K_d = 10^{-15} - 10^{-8}$ M). In spite of its apparent simplicity, such a mechanism is not understood in depth, either qualitatively or quantitatively. Here we focus on the simpler case of bacterial TFs recognizing their cognate sites on the naked DNA.

Currently, there are vast amounts of experimental data available, including the structures of protein-DNA complexes at atomic resolution in crystals and in solution [4, 5, 6, 7, 8], binding constants for dozens of native and hundreds of mutated proteins [9, 10], calorimetry measurements [11], and novel single-molecule experiments [12]. These experimental data are the most significant contribution to our present understanding of protein-DNA interaction since the early work of von Hippel, Berg *et al.* In a series of pioneering articles [13, 14, 15, 16], they created a conceptual basis for describing both the kinetics and thermodynamics of protein-DNA interaction, which became a starting point for practically every subsequent theoretical work on the subject.

We start by reviewing the history of the problem and describing the paradox of the "faster than diffusion" association rate. Next, we present the classical model of protein-DNA "sliding" and explain how this model can resolve the paradox. We outline the problem that the sliding mechanism faces if the energetics of protein-DNA interactions are taken into account. Next, we introduce our novel quantitative formalism and undertake an in-depth exploration of possible mechanisms of protein-DNA interaction. We conclude by discussing biological implications of our model and propose a number of experiments to check our key findings.

Chapter 1

Kinetics of protein-DNA interaction: the search speed – stability paradox

1.1 Introduction: "Faster than diffusion" search

The problem of how a protein finds its target site on DNA has a long history. In 1970, Riggs et. al. [17, 18] measured the association rate of LacI repressor and its operator on DNA as $\sim 10^{10}$ M^{-1}s^{-1}. This astonishingly high rate (as compared to other biological binding rates) was shown to be much higher than the maximal rate achievable by three-dimensional (3D) diffusion. In fact, if a protein binds its site by 3D diffusion, it has to hit the right site on the DNA within $b = 0.34$ nm. (A shift by 0.34 nm would result in binding a site that is different from the native one by 1bp. Such a site can be very different, e.g. GCGCAATT vs. CGCAATTC). Using the Debye-Smoluchowski equation for the *maximal* rate of a bimolecular reaction (see e. g. [19, 20, 21]), with a protein diffusion coefficient of $D_{3d} \sim 10^{-7}$cm^2/s [22] we get

$$k_{DS} = 4\pi D_{3D} b \sim 10^8 \text{ M}^{-1}\text{s}^{-1} \qquad (1.1)$$

This value for the association rate, relevant for *in vitro* measurements, corresponds to target location *in vivo* on a time scale of a few seconds, when each cell contains up to several tens of TF molecules.

To resolve the discrepancy between the experimentally measured rate of 10^{10} $\mathrm{M}^{-1}\mathrm{s}^{-1}$ and the maximal rate of 10^{8} $\mathrm{M}^{-1}\mathrm{s}^{-1}$ allowed by diffusion, Riggs *et al.*, Richter *et al.* [19] and later Winter, Berg and von Hippel [13, 15] suggested that the dimensionality of the problem changes during the search process. They concluded that while searching for its target site, the protein periodically scans the DNA by "sliding" along it.

If a protein performs both 3D and 1D diffusion, then the total search process can be considered as a 3D search followed by binding DNA and a round of 1D diffusion. Upon dissociation from the DNA, the protein continues 3D diffusion until it binds DNA in a different place, and so on. Some experimental evidence supports this search mechanism. These include affinity of the DNA-binding proteins for any fragment of DNA (non-specific binding), single molecule experiments where 1D diffusion has been observed and visualized, and numerous other experiments where the rate of specific binding to the target site has been significantly increased by lengthening non-specific DNA surrounding the site [23]. What are the benefits and the mechanism of 1D diffusion and what limits the search rate? In this chapter, we present a rather general albeit simple way to quantitatively address this question.

1.2 The Model

1.2.1 Search time

In our model, the search process consists of N rounds of 1D search (each takes time of $\tau_{1d,i}, i = 1..N$) separated by rounds of 3D diffusion ($\tau_{3d,i}$). The total search time t_s is the sum of the times of individual search rounds:

$$t_s = \sum_{i=1}^{N} \left(\tau_{1d,i} + \tau_{3d,i}\right) . \qquad (1.2)$$

The total number N of such rounds occurring before the target site is eventually found is very large, so it is natural to introduce probability distributions for the essentially random entities in the problem. The first simplification that can be made is to replace $\tau_{3d,i}$ by its average $\bar{\tau}_{3d}$. As we discuss below, this approximation is valid when the distribution of 3D diffusion times inside the DNA nucleoid is sufficiently narrow. Each round of 1D diffusion scans a region of n sites (where n is drawn from some distribution $p(n)$). The time $\tau_{1d}(n)$ it takes to scan n sites can be obtained from the exact form of the 1D diffusion law. If, on average, $\bar{n}$ sites are scanned in each round, then the average number of such rounds required to find the site on DNA of length M is $N = M/\bar{n}$. Using average values, we get a total search time of

$$t_s\left(\bar{n}, M\right) = \frac{M}{\bar{n}}\left[\tau_{1d}\left(\bar{n}\right) + \bar{\tau}_{3d}\right], \qquad (1.3)$$

From (1.3) it is clear that in general, $t_s\left(\bar{n}, M\right)$ is large for both very small and very large values of $\bar{n}$. In fact if $\bar{n}$ is small, very few sites are scanned in each round of 1D search and a large number of such rounds (alternating with rounds of 3D diffusion) are required to find the site. On the contrary, if $\bar{n}$ is large, lots of time is spent scanning a single stretch of DNA, making the search very redundant and inefficient. An optimal value $\bar{n}_{\text{opt}}$ should exist that provides little redundancy of 1D diffusion and a sufficiently small number of such rounds. For a given diffusion law $\tau_{1d}(n)$, function $t_s\left(\bar{n}, M\right)$ can be minimized producing $\bar{n}_{\text{opt}}$, the optimal length of DNA to be scanned between the association and the dissociation events [1].

1.2.2 Protein-DNA energetics

While diffusing along DNA, a TF experiences the binding potential $U(\vec{s})$ at every site $\vec{s}$ it encounters. The energy of protein-DNA interactions is usually divided into two parts, *specific* and *non-specific* [16, 24]

$$U_i = U(\vec{s} = s_i, ..s_{i+l-1}) + E_{\text{ns}}, \qquad (1.4)$$

[1] Naturally, we assume here that $\tau_{1d}\left(\bar{n}\right)$ grows with $\bar{n}$ at least as $O(\bar{n}^{1+\alpha})$, with $\alpha > 0$.

where $\vec{s}$ describes a DNA sequence of length l. As its name suggests, the non-specific binding energy E_{ns} arises from interactions that do not depend on the DNA sequence that the TF is bound to, e. g. interactions with the phosphate backbone. The specific part of the interaction energy exhibits a very strong dependence on the actual nucleotide sequence. Here and below we use the term "energy" to refer to the change in the free energy related to binding, ΔG_b. This free energy includes the entropic loss of translational and rotational degrees of freedom of the protein and amino acids' side-chains, the entropic cost of water and ion extrusion from the DNA interface, the hydrophobic effect, etc.

The energy of specific protein-DNA interactions can be approximated by a weight matrix (also known as Position-Specific Scoring Matrix (PSSM), or "profile") where each nucleotide contributes independently to the binding energy [16]:

$$U(\vec{s} = s_i, ..s_{i+l-1}) = \sum_{j=1}^{l} \epsilon(j, s_j), \tag{1.5}$$

where s_j is a base-pair in position j of the site and $\epsilon(j, x)$ is the contribution of base-pair x in position j. Most of the known weight matrices of TFs $\epsilon(j, s_j)$ give rise to uncorrelated energies of overlapping neighboring sites, obtained by one base pair shift [24]. Figure 1-1 presents distributions of the sequence specific binding energy $f(U)$ obtained for different bacterial transcription factors at all possible sites in the corresponding genome. The weight matrices for these transcription factors have been derived using a set of known binding sites and a standard approximation [16, 25]. Notice that, for a sufficiently long site, the distribution of the binding energy of random sites (or genomic DNA) can be closely approximated (see Fig. 1-1) by a Gaussian distribution with a certain mean $\langle U \rangle$ and variance σ^2:

$$f(U_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(U_i - \langle U \rangle)^2}{2\sigma^2}\right]. \tag{1.6}$$

Binding energies calculated for bacterial TFs support this assumption. Other physical factors such as local DNA flexibility [26] can create a correlated energy landscape, which provides a different mode of diffusion that we describe in Chapter 3.

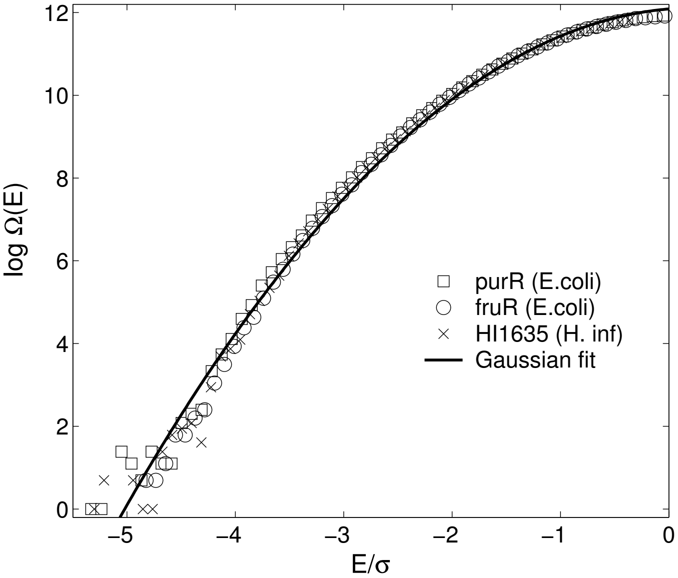


Figure 1-1: Spectrum of binding energy for three different transcription factors and the Gaussian approximation (solid line).

The whole DNA molecule can thus be mapped onto one-dimensional array of sites $\{\vec{s}_i\}$, each corresponding to a certain binding sequence comprising bases from the i-th to the $(i+l-1)$-th, l being the length of the motif (see Fig. 1-2). At each site, there is a probability p_i of hopping to site $i+1$ and a probability q_i of hopping to site $i-1$. These probabilities depend on the specific binding energies U_i and $U_{i\pm1}$ at the i-th site and at the adjacent sites, respectively, and are proportional to the corresponding transition *rates,* $\omega_{i,i+1}$ and $\omega_{i,i-1}$. For the latter, it is most natural to assume the regular activated transport form

$$\omega_{i,i\pm1} = \nu \times \begin{cases} e^{-\beta(U_{i\pm1}-U_i)} & \text{if } U_{i\pm1} > U_i \\ 1.0 & \text{otherwise} \end{cases} , \tag{1.7}$$

where ν is the effective attempt frequency, $\beta \equiv (k_BT)^{-1}$, k_B is the Boltzmann constant and T is the ambient temperature. The problem is thus related to a one-dimensional random walk with position-dependent hopping probabilities

$$p_i = \frac{\omega_{i,i+1}}{\omega_{i,i+1} + \omega_{i,i-1}}, \qquad q_i = 1 - p_i. \tag{1.8}$$

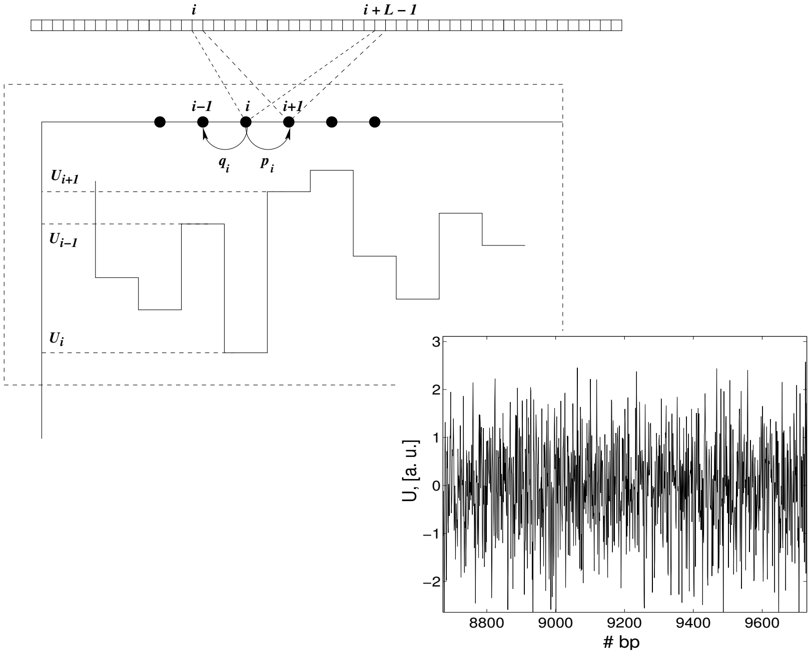


Figure 1-2: The Model Potential.

1.3 Diffusion in a sequence-dependent energy landscape

As has been shown in several papers in the last two decades, the properties of 1D random walks can vary dramatically depending on the actual choice of probabilities $\{p_i\}$ (for a review, see e.g. [27]). Here we employ the mean first-passage time (MFPT) formalism [28] to derive the diffusion law $\tau_{1d}(\bar{n})$ for protein sliding along the DNA given the sequence-dependent binding energy in Eq. (1.7).

The calculation consists of two steps, first, we describe the random walk along the DNA in terms of the number of steps. Next, we calculate the mean time between successive steps in a random energetic landscape which provides the time-scale for the problem. Such a decoupling, strictly speaking, does not hold when the number of steps is small, i.e. when the number of visited sites is small and the random quantities are not averaged properly. However, since we are dealing with large numbers of steps ($\sim 10^5 - 10^6$) this approach is valid, which is also confirmed by numerical simulations.

1.3.1 The MFPT.

To derive the diffusion law, we calculate the mean first passage time (MFPT) from site #0 to site #L, defined as the mean number of steps the particle is to make in order to reach the site #L *for the first time*. The derivation here follows the one in Ref. [28].

Let $P_{i,j}(n)$ denote the probability to start at site #i and reach the site #j in exactly n steps. Then, for example,

$$P_{i,i+1}(n) = p_i T_i(n-1), \tag{1.9}$$

where $T_i(n)$ is defined as the probability of returning to the i-th site after n steps *without* stepping to the right of it. Now, all the paths contributing to $T_i(n-1)$ should start with the step to the left and then reach the site #i in $n-2$ steps, not necessarily for the first time. Thus, the probability $T_i(n-1)$ can be written as

$$T_i(n-1) = q_i \sum_{m,l} P_{i-1,i}(m)\, T_i(l)\, \delta_{m+l,n-2}. \tag{1.10}$$

We now introduce generating functions

$$\tilde{P}_{i,j}(z) = \sum_{n=0}^{\infty} z^n \; P_{i,j}(n), \qquad \tilde{T}_i(z) = \sum_{n=0}^{\infty} z^n \; T_i(n). \tag{1.11}$$

One can easily show (see e. g. [29]) that

$$\tilde{P}_{0,L}(z) = \prod_{i=0}^{L-1} \tilde{P}_{i,i+1}(z). \tag{1.12}$$

Knowing $\tilde{P}_{i,i+1}(z)$, one calculates the MFPT straightforwardly as

$$\begin{aligned}\bar{t}_{0,L} = \frac{\sum_n nP_{0,L}(n)}{\sum_n P_{0,L}(n)} &= \left[\frac{d}{dz}\ln\tilde{P}_{0,L}(z)\right]_{z=1} \\ &= \sum_{i=0}^{L-1}\left[\frac{d}{dz}\ln\tilde{P}_{i,i+1}(z)\right]_{z=1}. \end{aligned} \tag{1.13}$$

Using (1.9) and (1.10), we obtain the following recursion relation for $\tilde{P}_{i,i+1}(z)$:

$$\tilde{P}_{i,i+1}(z) = \frac{zp_i}{1 - zq_i\tilde{P}_{i-1,i}(z)}. \tag{1.14}$$

To solve for $\bar{t}_{0,L}$, we must introduce boundary conditions. Let $p_0 = 1,\ q_0 = 0$, which is equivalent to introducing a reflecting wall at $i = 0$. This boundary condition clearly influences the solution for short times and distances. However, as numerical simulations and general considerations suggest, its influence relaxes quite fast, so that for longer times, the result is clearly independent of the boundary. The benefit of setting $p_0 = 1$ becomes clear when we observe that

$$\tilde{P}_{0,1}(1) = 1 \qquad \Rightarrow \qquad \forall\, i \qquad \tilde{P}_{i,i+1}(1) = 1. \tag{1.15}$$

Hence,

$$\bar{t}_{0,L} = \sum_{i=0}^{L-1}\tilde{P}'_{i,i+1}(1). \tag{1.16}$$

The recursion relation for $P'_{i,i+1}(1)$ is readily obtained from (1.14) :

$$\tilde{P}'_{i,i+1}(1) = \frac{1}{p_i} + \frac{q_i}{p_i}\tilde{P}'_{i-1,i}(1) = 1 + \alpha_i\left[1 + \tilde{P}'_{i-1,i}(1)\right], \tag{1.17}$$

with $\alpha_i \equiv q_i/p_i$. Thus, the expression for $\bar{t}_{0,L}$ is obtained in closed form

$$\bar{t}_{0,L} = L + \sum_{k=0}^{L-1}\alpha_k + \sum_{k=0}^{L-2}\sum_{i=k+1}^{L-1}(1+\alpha_k)\prod_{j=k+1}^{i}\alpha_j. \tag{1.18}$$

This solution expression gives the MFPT in terms of a given realization of disorder producing a certain set of probabilities $\{p_i\}$, whereas we are interested in the behavior averaged over all realizations of disorder. The cumulative products in (1.18) reduce to the two form $e^{\beta(U_i-U_j)}$, which after being averaged over *uncorrelated* Gaussian disorder produce a factor of $e^{\beta^2\sigma^2}$. After the summations are carried out, the expression for MFPT becomes for $L \gg 1$

$$\langle \bar{t}_{0,L} \rangle \simeq L^2 e^{\beta^2\sigma^2}. \tag{1.19}$$

Thus, the diffusion law appears to be the classical one, with a renormalized diffusion coefficient.

1.3.2 The time constant.

Consider a particle at site $\#i$. The particle will eventually escape to one of the neighboring sites $\#(i \pm 1)$, the escape rate being

$$r_i = \omega_{i,i+1} + \omega_{i,i-1}. \tag{1.20}$$

To calculate the characteristic diffusion time constant $\langle\tau\rangle$, this rate should be averaged over all configurations of disorder $\{U_i\}$. To obtain an analytic expression for the $\langle\tau\rangle$, we assume the form

$$\omega_{i,i\pm1} = \nu e^{-\beta(U_{i\pm1}-U_i)} \tag{1.21}$$

for both $U_{i\pm1} > U_i$ and $U_{i\pm1} < U_i$, as opposed to the form (1.7). Numerics show that this approximation introduces an up to ~ 15% error for small values of $\beta\sigma$ and is practically exact for $\beta\sigma > 2$. Thus,

$$r_i = \frac{1}{2\tau_0}\left(e^{-\beta(U_{i+1}-U_i)} + e^{-\beta(U_{i-1}-U_i)}\right), \tag{1.22}$$

where $\tau_0 = 1/(2\nu)$. The mean time between the successive steps can be calculated therefore as the average over all possible configurations of U_i, $U_{i\pm1}$ of the reciprocal of the escape rate, i. e.

$$\langle\tau\rangle = \left\langle \frac{1}{r_i} \right\rangle = 2\tau_0 \int_{-\infty}^{\infty} dU_i dU_{i+1} dU_{i-1} \frac{f\left(U_i\right) f\left(U_{i+1}\right) f\left(U_{i-1}\right)}{e^{-\beta(U_{i+1}-U_i)} + e^{-\beta(U_{i-1}-U_i)}}. \tag{1.23}$$

Assuming as above Gaussian energy statistics, this integral is evaluated as follows

$$\langle\tau\rangle = \frac{\tau_0 \ e^{\beta^2\sigma^2/2}}{\pi} \int_{-\infty}^{\infty} dxdy \frac{e^{-(x^2+y^2)/2}}{e^{-\beta\sigma x} + e^{-\beta\sigma y}}. \tag{1.24}$$

After the change of variables

$$s = \frac{1}{\sqrt{2}}(x+y), \qquad t = \frac{1}{\sqrt{2}}(x-y), \tag{1.25}$$

the integral factorizes leading to

$$\begin{aligned} \langle\tau\rangle &= \frac{\tau_0 \ e^{\beta^2\sigma^2/2}}{2\pi} \int_{-\infty}^{\infty} ds \ e^{-s^2/2+\beta\sigma s/\sqrt{2}} \int_{-\infty}^{\infty} dt \frac{e^{-t^2/2}}{\cosh(\beta\sigma t/\sqrt{2})} \\ &= \frac{\tau_0 \ e^{3\beta^2\sigma^2/4}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt \ e^{-t^2/2-\ln\left[\cosh(\beta\sigma t/\sqrt{2})\right]} \\ &= \frac{\tau_0 \ e^{3\beta^2\sigma^2/4}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dt \ e^{-t^2(1+\beta^2\sigma^2/2+\ldots)/2} \simeq \tau_0 \ e^{3\beta^2\sigma^2/4} \left[1+\beta^2\sigma^2/2\right]^{-1/2} \end{aligned} \tag{1.26}$$

Now, multiplying (1.19) by $\langle\tau\rangle$, we obtain the diffusion coefficient as

$$D_{1d}\left(\sigma\right) \simeq \frac{1}{2\tau_0}\left(1+\frac{\beta^2\sigma^2}{2}\right)^{1/2} e^{-7\beta^2\sigma^2/4}. \tag{1.27}$$

Hence, rapid diffusion of a protein along the DNA is possible only if the roughness of the binding energy landscape is small compared to $k_B T$ ($\beta\sigma < 1.5$). This requirement imposes strong constraints on the allowed energy of specific binding interactions.

1.4 Optimal time of 3D/1D search

When 1D scanning is combined with 3D diffusion, what is the optimal time a protein has to spend in each of the two regimes? To answer this question we compute the optimal number of sites the protein has to scan by 1D diffusion in order to get the fastest overall search. Results of this section are rather general and are not limited to the particular scenario of slow 1D diffusion on a rough landscape discussed above.

Each time the protein binds DNA, it performs a round of 1D diffusion. If the round lasts τ_{1d} then on average the protein scans [30]

$$\bar{n} = \sqrt{16 D_{1d} \tau_{1d} / \pi} \text{ bps.} \tag{1.28}$$

By plugging this relation into Eq. (1.3) for search time t_s, and minimizing t_s with respect to $\bar{n}$, we get the optimal total search time and the optimal number of sites to be scanned in each round:

$$t_s^{\text{opt}} = t_s(\bar{n}_{\text{opt}}) = \frac{M}{2}\sqrt{\frac{\pi \bar{\tau}_{3d}}{D_{1d}}} \quad \bar{n}_{\text{opt}} = \sqrt{\frac{16}{\pi} D_{1d} \bar{\tau}_{3d}} \tag{1.29}$$

This analysis brings us to the following conclusions.

First, and most importantly, we obtain that in the *optimal* regime of search

$$\tau_{1d}(\bar{n}_{\text{opt}}) = \tau_{3d}, \tag{1.30}$$

i.e. the protein spends equal amounts of time diffusing along non-specific DNA and diffusing in the solution. This result is very general, and is true irrespective of the values of diffusion coefficients D_{1d} or D_{3d}, or size of the genome M. In fact, it follows directly from the diffusion law $\bar{n} \sim \sqrt{\tau_{1d}}$. More importantly, this central result can be verified experimentally by either single-molecule techniques or by traditional methods. Also note that the optimal length of DNA scanned in a single round of 1D diffusion $\bar{n}_{\text{opt}}$ does not depend on M, i.e. it is the same irrespective of the size of the genomes to be searched for a specific site.

Second, the optimal 1D/3D combination reached at $\tau_{1d} = \tau_{3d}$ leads to a significant speed up of the search process. In fact, an optimal 1D/3D search is $\bar{n}_{\text{opt}}$ times faster than a search by 3D diffusion alone, and $M/\bar{n}_{\text{opt}}$ times faster than a search by 1D diffusion alone. For example, if the protein operates in the optimal 1D/3D regime and scans $\bar{n}_{\text{opt}} = 100$ bp during each round of DNA binding, then the experimentally measured rate of binding to the specific site can be 100 times greater than the rate achievable by 3D diffusion alone.

Third, we can estimate $\bar{n}_{opt}$, the maximal number of sites a protein can scan in each round of 1D search. If we set D_{1d} to its maximum, i.e. $D_{1d} \sim D_{3d}$ and estimate $\bar{\tau}_{3d}$ as a characteristic time of diffusion through a DNA globule of size l_{m}

$$\bar{\tau}_{3d} \sim l_{\text{m}}^2 / D_{3d}, \tag{1.31}$$

with $l_{\text{m}} \sim 0.1\mu$m, we get

$$\bar{n}_{\text{opt}}^{\text{max}} \sim 500 \text{ bp}. \tag{1.32}$$

For a smaller 1D diffusion coefficient, e. g. $D_{1d} \sim D_{3d}/100$, we get $\bar{n}_{\text{opt}}^{\text{max}} \sim 50$bp. Again, single molecule experiments can provide estimates of these quantities for different conditions of diffusion.

Finally, we obtain estimates of the shortest possible total search time. If $M \approx 10^6$ bp and 1D diffusion is at its fastest rate, i. e. $D_{1d} \sim D_{3d} = 10^{-7}\text{cm}^2/\text{s}$, then using Eq. (1.29) we get

$$t_s^{opt} \sim \frac{M}{2}\sqrt{2\pi\bar{\tau}_{3d}\tau_0} \sim 5 \text{ sec}, \tag{1.33}$$

where, given the inter-base distance $a_0 = 0.34$nm, we estimate $\tau_0 \sim a_0^2/D_{1d} \sim 10^{-8}$ sec.

One can also estimate the search time using *in vitro* experimentally measured binding rates in water $k_{on}^{\text{water}} \approx 10^{10}\text{M}^{-1}\text{s}^{-1}$ [17, 18]. The diffusion coefficient of a protein molecule in water can be estimated as [31]

$$D \simeq \frac{k_B T}{3\pi\eta d}, \tag{1.34}$$

where d is the diameter of the molecule and η is the water viscosity. Setting $\eta \sim 10^{-2}$ g/(sec · cm) and $d \sim 10$ nm, we obtain at room temperature

$$D \sim 10^2 \ \mu\text{m}^2/\text{sec}. \tag{1.35}$$

Diffusion coefficient measurements for GFP in *E. coli* [22] produce values of about $1 - 10\ \mu\mathrm{m}^2/\mathrm{sec}$. This difference in diffusion coefficients may account for more than order of magnitude difference in the theoretically calculated and measured target location times. Thus, the estimated *in vivo* binding rate is $k_{on}^{\mathrm{cytoplasm}} \approx 10^8 - 10^9 \mathrm{M}^{-1}\mathrm{s}^{-1}$. From this we obtain the time it takes for one protein to bind one site in a cell of $1\mu\mathrm{m}^3$ volume (i.e. $[\mathrm{TF}] \approx 10^{-9}\mathrm{M}$) as

$$t_s^{exp} = \left(k_{on}^{\mathrm{cytoplasm}}[\mathrm{TF}]\right)^{-1} \sim 1 - 10 \text{ sec.} \tag{1.36}$$

One can see a good agreement between our theoretical estimates and experimentally measured binding rates.

As we mentioned above, there are usually several TF molecules searching in parallel for the target site. Naturally, in this case, the search is sped up proportionally to the number of molecules.

1.5 Non-specific binding

While the diffusion of the TF molecules along DNA is controlled by the specific binding energy, the dissociation of the TF from the DNA depends on the total binding energy, i.e. on the non-specific and specific binding. Moreover, since the dissociation events are much less frequent than the hopping between neighboring base-pairs (roughly by a factor of $\bar{\tau}_{3d}/\langle\tau\rangle$), the non-specific energy E_{ns} makes a correspondingly larger contribution to the total binding energy.

For a TF at rest bound to some DNA site i, the dissociation rate r_i^{diss} would be given by the Arrhenius-type relation,

$$r_i^{\mathrm{diss}} = \frac{1}{\tau_0} e^{-\beta(E_{\mathrm{ns}} - U_i)}. \tag{1.37}$$

Given the specific U_i and the non-specific E_{ns} energy, one can calculate the average time τ_{1d} a protein spends before dissociating from the DNA

$$\tau_{1d} = \left\langle \frac{1}{r_i^{\text{diss}}} \right\rangle = \tau_0 e^{\beta E_{\text{ns}} + \beta^2 \sigma^2/2}. \tag{1.38}$$

Next we recall that, in the optimal regime, $\tau_{1d} = \bar{\tau}_{3d}$. Thus, to ensure optimal performance, τ_{1d} should be replaced by $\bar{\tau}_{3d}$ in Eq. (1.38):

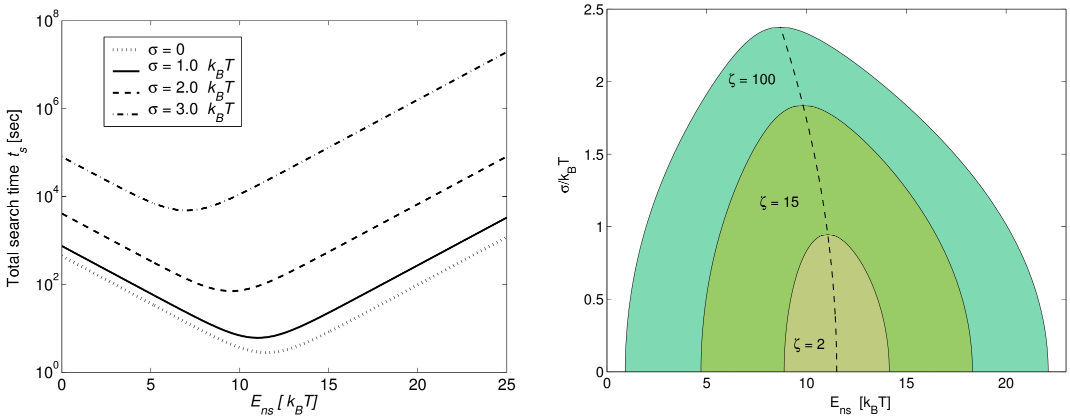
$$E_{\text{ns}} = k_B T \left[\ln\left(\frac{\tau_{3d}}{\tau_0}\right) - \frac{1}{2}\left(\frac{\sigma}{k_B T}\right)^2 \right]. \tag{1.39}$$

Since for a given value of σ, the non-specific binding controls the dissociation rate, the search time will deviate from the optimum if E_{ns} moves from this predetermined value. In Fig. 1-3a, we use Eqs. (1.3) and (1.28) to plot the search time as a function of the non-specific binding energy for different values of σ.

We now define the *tolerance factor* ζ as the ratio between the maximal acceptable value of the search time t_s and the minimal time t_{s0}. Experimental data suggest $\zeta \leq 5$, but we for the moment allow for much larger values of $\zeta \sim 10 - 100$ (this can be done when, for instance, there are many protein molecules searching in parallel). As we can see from Fig. 1-3a, for each value of σ, there is a range of possible values of E_{ns} such that the resulting search time is within the region of tolerance. This range is easily calculated producing the values of non-specific energy between

$$E_{\text{ns}}^{\pm}(\sigma, \zeta) = \frac{2}{\beta} \ln\left[\sqrt{\frac{D_{1d}(\sigma)\bar{\tau}_{3d}}{D_{1d}(0)\tau_0}} \left(\zeta \pm \sqrt{\zeta^2 - \frac{D_{1d}(0)}{D_{1d}(\sigma)}} \right) \right] - \frac{\sigma^2 \beta}{2} \tag{1.40}$$

Specifying ζ, we can define our parameter space, i. e. the values of specific and non-specific energy producing a total search time within the region of tolerance. In Fig. 1-3b, we consider three values of ζ. The most relaxed requirement $\zeta = 100$ provides a search time $t_s \leq 500$ sec. If 100 proteins are searching for a single site, then the first one will find it after ~ 5 sec, leading however to a fairly low binding rate of $k_{\text{on}} \approx 1/500 \text{ sec} \cdot 10^9 \text{ M}^{-1} = 2 \cdot 10^6 \text{ M}^{-1}\text{s}^{-1}$ (compared to experimentally measured $10^{10} \text{ M}^{-1}\text{s}^{-1}$ in water). Importantly, in order to comply with even this most relaxed



(a) (b)

Figure 1-3: (a) Dependence of the search time on the non-specific binding energy. (b) The parameter space. The dashed line corresponds to optimal parameters σ and E_{ns} connected by Eq. (1.39).

search time requirement, the characteristic strength of specific interaction must be smaller than $\sim 2.3\ k_BT$.

These results bring us to a very important conclusion that a protein cannot find its site in biologically relevant time if the roughness of the specific binding landscape is greater than $\sim 2\ k_BT$. Although an optimal 1D/3D combination can speed up the search, it cannot overcome the slowdown of 1D diffusion. Only fairly smooth landscapes ($\sigma \sim 1 k_BT$) can be effectively navigated by proteins.

1.6 Speed versus stability

While rapid search requires fairly smooth landscapes ($\sigma \sim 1 k_BT$), stability of the protein-DNA complex, in turn, requires a low energy of the target site ($U_{min} < 15\ k_BT$ for a genome of 10^6 bp).

In Fig. 1-4a, we present the equilibrium probability P_b of binding the strongest target site with energy $U_{\min} = U_0$ [24] as a function of σ/k_BT. In equilibrium, P_b equals the fraction of time the protein spends at the target site:

$$P_b = \frac{\exp\left[-\beta U_0\right]}{\sum_{i=0}^{M} \exp\left[-\beta U_i\right]}. \tag{1.41}$$

Since the target site is not separated from the rest of the distribution by a significant energy gap, P_b is comparable to 1 (which is the natural requirement for a good regulatory site) only at *σ much greater than k_BT.*

In fact, it is not hard to estimate analytically the (σ/k_BT) ratio for a genome of length M such that the probability of binding to the lowest site is comparable to the probability of binding to the rest of the genome, i.e. their contributions to the partition function are of the same order of magnitude. The partition sum for the Gaussian energy level statistics is

$$\begin{aligned}\Omega &= \frac{M}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\beta U - U^2/(2\sigma^2)}dU = Me^{\beta^2\sigma^2/2} \sim \\ &\sim \exp\left[-\beta U_{\min}\right] \sim \exp\left(\beta\sigma\sqrt{2\ln M}\right) \qquad (1.42)\end{aligned}$$

so that for $M = 10^6$

$$\sigma \sim k_BT\sqrt{2\ln M} \simeq 5\ k_BT. \qquad (1.43)$$

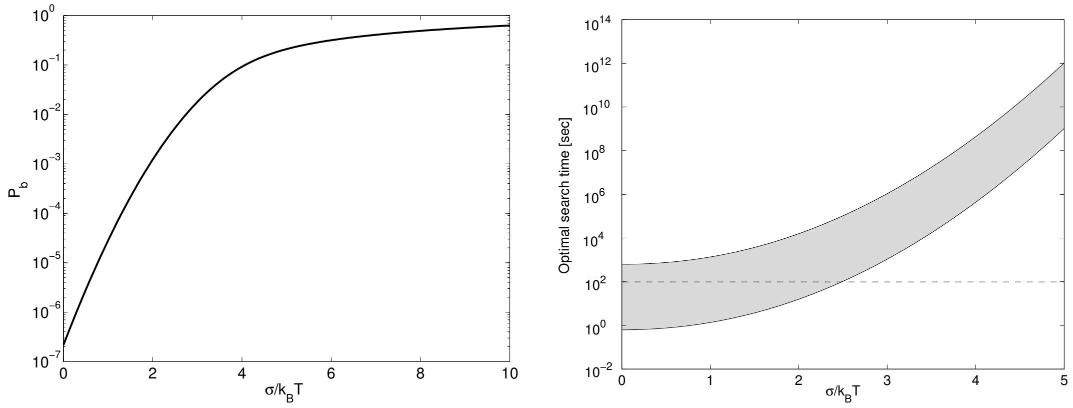
Strictly speaking, for a large though finite set of energy levels, the integration limits are cut off at $\pm\sigma\sqrt{2\ln M}$ so that for $\beta\sigma \gg \sqrt{\ln M}$ the partition function is dominated by the lower edge of the distribution. The estimate for $\beta\sigma$ gives therefore the crossover value between the regime of multiple-site contribution to Ω and the regime with single-site domination[2].

Figure 1-4b shows the optimal search time at the corresponding values of σ/k_BT. High roughness of $\sigma >> k_BT$ required for stability of the protein-DNA complex leads to astronomically large search times. In contrast, a protein can effectively search the target site *at σ smaller than* $1 - 2k_BT$.

From the above analysis, an obvious conflict arises: *the same energy landscape cannot allow for both rapid translocation and high stability of states formed at sites with the lowest energy.* This conflict is similar to the speed-stability paradox of protein folding formulated by Gutin *et al.* [33]: rapid search in conformation space requires a smooth energy landscape, but then the native state is unstable. In protein

[2] In the Random Energy Model [32], the analog of this effect is thermodynamical freezing.

folding, this conflict is resolved by the presence of a large energy gap between the native state and the rest of the conformations [34, 35].



(a)

(b)

Figure 1-4: (a) Stability on the protein-DNA complex on the cognate site measured as the fraction of time in the bound state at equilibrium. (b) Optimal search time as a function of the binding profile roughness, for the range of parameters 10^{-4}sec $\leq \tau_{3d} \leq 10^{-2}$sec, 10^{-10}sec $\leq \tau_0 \leq 10^{-6}$sec.

As evident from Fig. 1-1, no such energy gap separates cognate sites from the bulk of other (random) sites. In fact, the energy function in the form of (1.5) cannot, in principle, provide a significant energy gap.

Increasing the number of TFs cannot resolve the paradox either. If N_p proteins are searching and binding a single target site, then the probability of being occupied is given by

$$P(N_p) = 1 - (1 - P_b)^{N_p} \approx N_p P_b \tag{1.44}$$

where P_b is the probability of the site being occupied by a single protein and approximation is for $P_b << 1/N_p$. As evident from Fig. 1-4b, requirement of the rapid search is satisfied if $P_b(\sigma/k_BT \approx 1) \sim 10^{-5}$. An unrealistic number of copies of a single TF, $\sim 10^4$, is required to saturate such weak binding site. Thus, an alternative solution must be sought.

Chapter 2

Folding and binding

2.1 The two-mode model

The “search speed-stability” paradox has already been qualitatively anticipated by Winter, Berg and von Hippel [14], who concluded that a conformational change of some sort must exist to allow fast switching between “specific” and “non-specific” modes of binding. In the non-specific mode, the protein is “sliding” over an essentially equipotential surface (in our terms, $\sigma_{\mathrm{non-spec}} = 0$) whereas site-binding takes place in the “specific” mode ($\sigma_{\mathrm{spec}} \gg k_B T$). A protein in the non-specific binding mode is “unaware” of the DNA sequence it is bound to. Thus, it permanently alternates between the binding modes, probing the underlying sites for specificity.

This model naturally raises a question about the nature of the conformational change, which was originally described as a “microscopic” binding of the protein to the DNA accompanied by water and ion extrusion. However, numerous calorimetry measurements and calculations [11] show that such a transition is usually accompanied by a large heat capacity change ΔC. This ΔC cannot be accounted for, unless additional degrees of freedom, namely protein folding, are taken into account. On-site folding of the transcription factor may involve significant structural change [20, 21, 36] and take a time of $\sim 10^{-4} - 10^{-6}$ sec [37] (compared to a characteristic on-site time of $\tau_0 \sim 10^{-7} - 10^{-8}$ sec).

If the TF is to probe every site for specificity in this fashion, it would take hours to locate the native site. We note, however, that if there was a way to probe only a very limited set of sites, i.e. only those having high potential for specificity, the search time would be dramatically reduced. From the previous section it is clear that a relatively weak site-specific interaction (i.e. a smooth landscape, $\sigma \sim k_BT$) does not significantly affect the diffusive properties of the TF and the total search time. If this landscape, however, is correlated with the actual specific binding energy landscape (with $\sigma \sim 5-6\ k_BT$), the specific sites will be the strongest ones in both modes. The protein conformational changes should occur therefore mainly at these sites, which constitute "traps" in the smooth landscape. Since such sites constitute a very small fraction of the total number of sites, the transitions between the modes are very rare.

We therefore suggest that there are two modes of protein-DNA binding: the *search* mode and the *recognition* mode (Fig. 2-1). In the search mode, the protein conformation is such that it allows only a relatively weak site-specific interaction ($\sigma_{\rm s} \sim 1-2\ k_BT$). In the recognition mode, the protein is in its final conformation and interacts very strongly ($\sigma_{\rm r} \geq 5\ k_BT$) with the DNA (Fig 2-1 bottom). If two energy profiles are strongly correlated then the lowest lying energy levels ("traps") in the search mode ($\leq -5\ k_BT$) are likely to correspond to the strongest sites in the recognition mode, putatively, the cognate sites. The transitions between the two modes happen mainly when the protein is trapped at a low-energy site of the search landscape. In this fashion, the 1D diffusion coefficient D_{1d} is about 10–100 times smaller than the ideal limit, but the search time in the optimal regime is reduced only by a factor of $\sim 3-10$ (see Eq. (1.29)).

The coupling between the conformational change and association at a site with a low-energy trap is likely to take place through time conditioning. Namely, *the folding (or a similar conformation transition) occurs only if the protein spends some minimal amount of time bound to a certain site.* This statement is basically equivalent to saying that the free energy barrier that the protein must overcome to transform to the final state must be comparable to the characteristic energy difference that controls hopping to the neighboring sites.

The protein conformation in recognition mode should be stabilized by additional

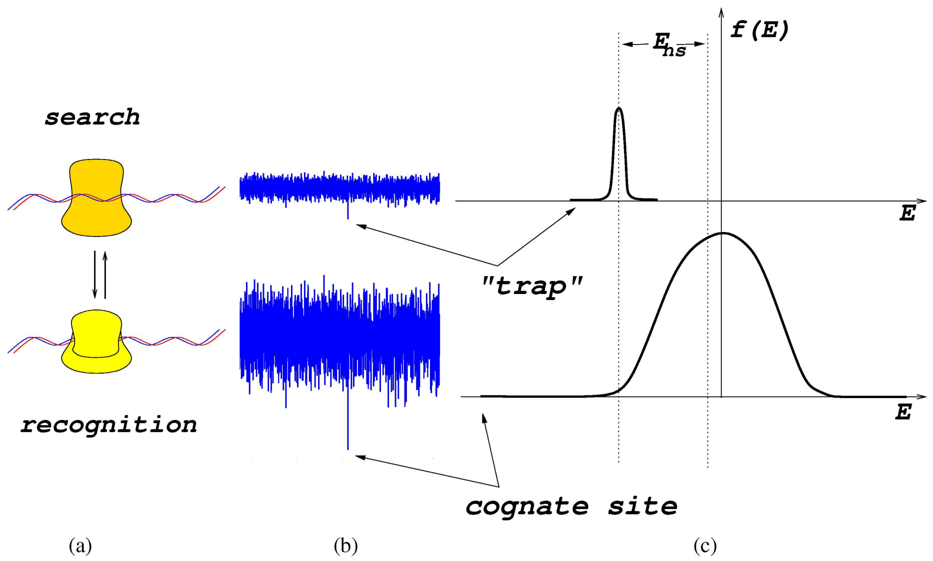


Figure 2-1: Cartoon demonstrating the two-mode search-and-fold mechanism. Top: search mode, bottom: recognition mode (a) two conformations of the protein bound to DNA: partially unfolded (top) and fully folded (bottom). (b) The binding energy landscape experienced by the protein in the corresponding conformations. (c) The spectrum of the binding energy determining stability of the protein in the corresponding conformations.

protein-DNA interactions. If these interactions are unfavorable, the folded structure is destabilized, then the search conformation is rapidly restored and the diffusion proceeds as before. If the new interactions are favorable, the folded structure is stable and the protein is trapped at the site for a very long time.

For this mechanism to work, transition between the two modes of search has to be associated with significant change in the free energy ($\sim 15k_BT$) of the protein-DNA complex (see Fig 2-1(c)). Such energy difference between the two states is required to make most of the high-energy sites in the recognition mode less favorable than in the search mode. So a protein would rather (partially) unfold than bind an unfavorable site. As a result, sites that lay higher in energy than a certain cutoff exhibit similar non-specific binding energy (i.e. switch into search mode of binding). Folding of partially disordered protein loops or helices can provide the required free energy difference between the two modes.

Efficiency of the proposed search-and-fold mechanism depends on the energy difference between the two modes, correlation between the energy profiles and the barrier between the two states. The barrier determines the rate of partial folding-unfolding transition. If the barrier is too low, then the protein equilibrates while on a single site having no effect on search kinetics. On the contrary, too high a barrier can lead to rare folding events and the cognate site can be missed. As we show in the following sections, a proper size of the barrier provides efficient search and stable protein-DNA complex.

2.2 The two profiles

A protein located at a site with locally minimal energy ΔU has the average residence time of

$$\tau_{res} \sim \tau_0 e^{\beta \Delta U}. \tag{2.1}$$

Then, the probability to undergo a conformational change there is

$$p_f(\Delta U) \sim \frac{\tau_{res}}{\tau_f} \sim \kappa^{-1} e^{\beta \Delta U}, \tag{2.2}$$

where ${\tau_f}^{-1}$ is the mean transition rate and $\kappa \equiv \tau_f/\tau_0$. Since on a single round of 1D diffusion the protein covers $\sim n$ sites and makes $\sim n^2$ steps, each site is revisited $\sim n$ times. Thus, the overall probability to locate the target site once the protein associates inside a region of size $\sim n$ containing the site is

$$p_{loc} \sim \min[1, np_f] \sim \min[1, p_f e^{\beta E_{ns}/2}]. \tag{2.3}$$

Now, we say that the location mechanism is *robust* if $p_{loc} \sim 1$. Also, we note that for a genome of size M, extreme values of a Gaussian distribution with variance σ_s^2 are approximately

$$\Delta U \sim \sigma_s \sqrt{2 \ln \frac{M}{\sqrt{2\pi}}}, \tag{2.4}$$

and they correspond to the outlier sites in our problem. Then, the protein is able to locate its cognate site robustly if

$$\kappa < \kappa_c \sim \exp\left[\beta\left(\frac{E_{ns}}{2} + \sigma_s\sqrt{2\ln\frac{M}{\sqrt{2\pi}}}\right)\right]. \tag{2.5}$$

Thus, for instance, for a genome size $M = 5 \times 10^6$bp, $\beta\sigma_s = 1.5$ and $E_{ns} \simeq 10k_BT$, we have

$$\kappa_c \sim 5 \times 10^5. \tag{2.6}$$

This corresponds to maximal mode transition (i.e. folding) times of the order of milliseconds and even slower. Note that the accepted point of view corresponds to $\sigma_s = 0$, which gives

$$\kappa_c \sim 10^2, \tag{2.7}$$

providing a much smaller degree of robustness in target location.

What happens if $\kappa > \kappa_c$? In this case, the protein dissociates from the region containing the site before a transition to the recognition mode occurs. To locate the site, the same region has to be scanned repetitively. In fact, for $p_{loc} \ll 1$, the protein has to return to the same region $\sim 1/p_{loc}$ times and the overall target location time grows proportionally to the number of returns. Thus, we come to a surprising conclusion: for given folding times, i.e. for a given TF, the *overall* target location time is shorter for slower sliding!

2.2.1 The numerics

To tackle the problem numerically, we use a version of the Gillespie algorithm [38, 39]. The protein at a given site can undergo 4 possible "reactions": it can move in positive or negative direction along the DNA, overcome a conformational change or dissociate from the DNA and reassociate at some other (random) lattice site. The rate of a reaction $\delta \to \gamma$ is calculated as according to

$$\omega_{\delta\to\gamma} = \frac{1}{\tau_0} \times \begin{cases} e^{-\beta\Delta U_{\delta\gamma}} & \text{if } \Delta U_{\delta\gamma} > 0 \\ 1.0 & \text{otherwise} \end{cases} , \tag{2.8}$$

where $\beta = 1/k_BT$, and $\Delta U_{\delta\gamma}$ is the energy barrier for the reaction. For movements along the DNA, this is the difference in the potential at the neighboring lattice sites; for dissociation from a site i, $\Delta U_{\delta\gamma} = E_{ns} - U_i$. Transitions from search to recognition mode are assumed to be governed by an energy barrier

$$\Delta G_{s\to r} = \max\left[U_i^r - U_i^s, k_BT \ln\left(\frac{\tau_f}{\tau_0}\right)\right], \tag{2.9}$$

where $U_i^{s,r}$ correspond to binding energies in the search and recognition modes, respectively. Reverse transitions have the barrier

$$\Delta G_{r\to s} = k_BT \ln\left(\frac{\tau_f}{\tau_0}\right) + U_i^s - U_i^r. \tag{2.10}$$

Each Monte-Carlo (MC) move starts from choosing the next reaction at random, with each reaction weighted proportionally to its rate (2.8). Then, the time to next reaction is drawn from the exponential distribution with a mean

$$\langle \Delta t \rangle = \left(\sum_\gamma \omega_{\delta\to\gamma}\right)^{-1}. \tag{2.11}$$

If the next reaction is dissociation, the total search time is incremented by a 3D diffusion time τ_{3d} and the protein is relocated to a new random position on DNA.

To build the recognition mode profile, we employ the standard weight–matrix method [16, 40, 25] using a known set of $PurR$ transcription factor binding sites (see Appendix A). The search mode profile is built by rescaling the recognition mode profile.

2.2.2 Robustness of target location

First, we simulate the process of target location on a short stretch of DNA ($M = 1000$ bp). The process starts when a protein is bound at a random site in the search mode and ends when it is bound at the cognate site in the recognition mode. Figure 2-2 shows the values of the mean total search time t_{loc} as a function of τ_f for various values of σ_s. We see that each graph consists of a plateau region and a linear growth region. The former corresponds to the values of τ_f, for which $p_{loc} \simeq 1$, whereas the latter is realized when $p_{loc} < 1$. In this regime

$$t_{loc} \simeq \frac{2M\tau_{3d}}{np_{loc}} \sim \frac{2M\tau_{3d}}{n}\left(\frac{\tau_f}{\tau_{res}}\right). \tag{2.12}$$

In the plateau region, the smaller σ_s is, the faster target location is. The slowdown for $\beta\sigma_s \leq 1.5$ is less than by a factor of 10 (note that in the optimal regime, the slowdown factor for $\beta\sigma_s = 1.5$ is $e^{-7\beta^2\sigma_s^2/8} \simeq 0.14$). In the linear growth region, for a given τ_f, the relation is reversed. For $\sigma_s = 0$, the residence time τ_{res} is small ($< 10^{-7}$ sec) and it grows exponentially with σ_s, and so does the extent of the plateau region.

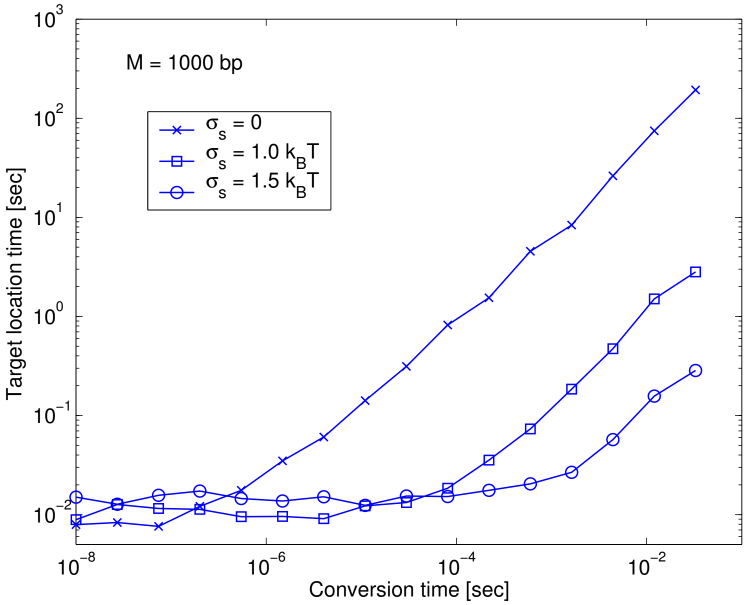


Figure 2-2: Mean target location time as a function of on-site conformational transition time for various values of σ_s.

To have a more accurate measure of p_{loc}, we perform the following set of simulations. For a given search profile, we place the protein in the *search* mode at the cognate site and measure the probability of finding it at the same site in the *recognition* mode before it dissociates from the DNA. Figure 2-3 presents the dependence of p_{loc} on τ_f for various values of σ_s. One can see a remarkable correspondence between the behavior of the target location time and that of p_{loc}.

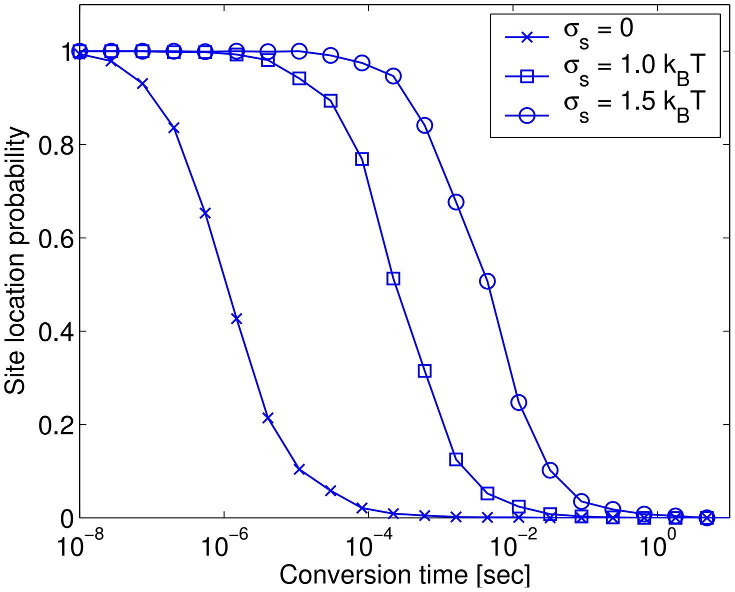


Figure 2-3: Probability of conformational transition before dissociation as a function of on-site conformational transition time.

Thus, we conclude that if the search landscape is correlated with the recognition one, target location is significantly more robust for search landscapes of finite roughness.

2.3 Effective energy landscape

The mechanism proposed above allows to bridge the timescale gap between 1D diffusion and protein conformational changes by the virtue of a "search" potential profile which is correlated with the "recognition" profile. The latter creates the required distribution of waiting times, so that conformational transitions are very likely to occur at a preselected set of sites corresponding to potential minima. This hypothesis is still to be tested experimentally, but, it appears that search speed–stability paradox cannot be resolved without at least one additional reaction coordinate describing two binding modes.

In this section we explore yet another possibility, namely, that this reaction coordinate is continuous rather than discrete. We suggest that protein–DNA interaction can be effectively described by a two–dimensional (2D) energy landscape. We then model the dynamics of protein–DNA complex as a random walk on this landscape. Finally, we discuss the influence of slow transversal dynamics on the kinetic and equilibrium properties of the landscape.

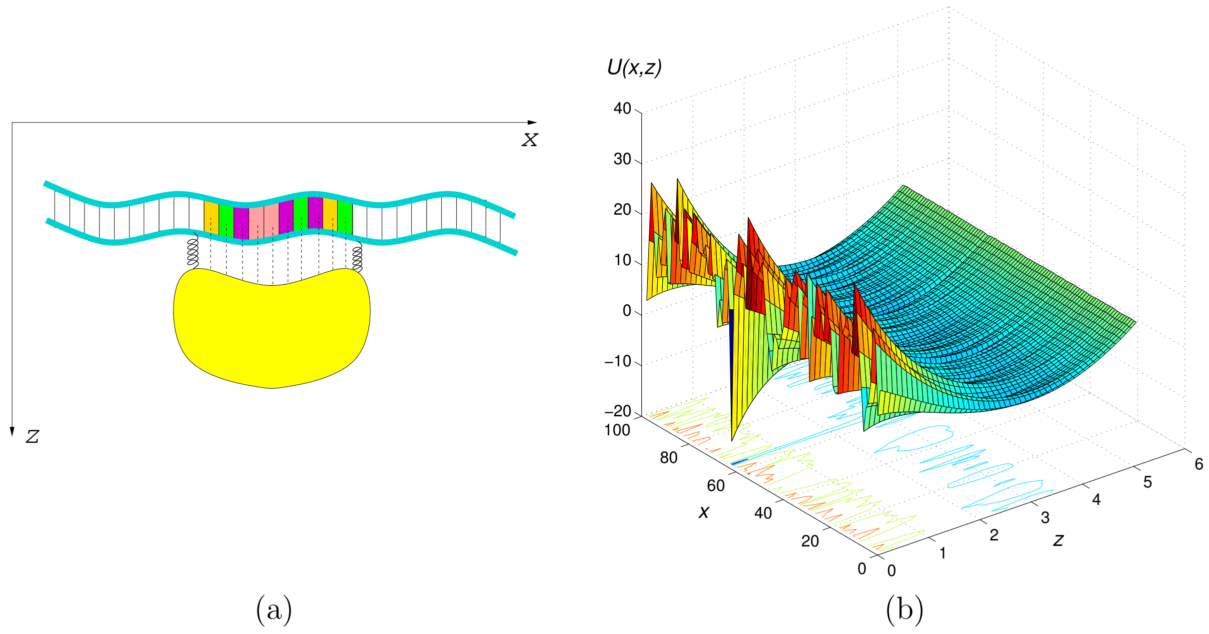
2.3.1 The Continuos Model

Treating chemical reactions as stochastic dynamics on energy landscapes is a very common approach in chemical physics [41]. In molecular biophysics, e.g. in protein or RNA folding problems, the energy landscapes are defined in a multidimensional space [42] and the computational effort associated with modelling such dynamics is enormous.

In our approach, the effective energy landscape $U(x,z)$ is two–dimensional: x is the position of the protein along the DNA and z is a continuous reaction coordinate that describes the internal dynamics of the protein–DNA complex [43, 44]. The potential has a general form

$$U(x,z) = U_{\mathrm{spec}}(x,z) + U_{\mathrm{non-spec}}(z), \tag{2.13}$$

where $U_{\mathrm{non-spec}}(z)$ is sequence–independent and arises mainly from electrostatic interactions with DNA backbone, the solvent etc.; it has a minimum away from the DNA. $U_{\mathrm{spec}}(x,z)$ depends on the sequence and the state of the complex; it is assumed to decay rapidly in z but to be strong enough to provide a net nonzero force for finite z. In this fashion, cognate sites can reduce the energy barrier to formation of the specific complex.



(a)

(b)

Figure 2-4: (a) A schematic view of a protein attached to DNA: harmonic potential combined with sequence–specific interaction. (b) A three–dimensional view of the energy landscape; note a specific site at $x = 63$

We thus assume probably the simplest possible functional form that has the required properties:

$$U(x, z) = U_{\mathrm{s}}(x)e^{-z} + \frac{\alpha}{2}(z - z_0)^2. \tag{2.14}$$

Here, α and z_0 are the potential parameters and $U_{\mathrm{s}}(x)$ is the specific potential profile that can be taken from one of the standard models [16, 25, 40, 45, 10]; also, $\overline{U_{\mathrm{s}}(x)} = 0$. Figure 2-5 illustrates this toy model. Note, that in the original picture by Winter *et. al.* [14], the "hard" and fully folded protein switches between the non–specific and the specific modes by "docking" to DNA and expelling water and ions. In this case, the reaction coordinate z is merely the distance from DNA. However, the internal protein–DNA complex dynamics involve multiple degrees of freedom [46], so that the explicit meaning of the *ad hoc* reaction coordinate z is not obvious. Nevertheless, in what follows, we assume that these dynamics can be effectively described by a single coordinate.

For suitably chosen α and z_0, all sites can be divided into three categories. Most non–cognate sites have an equilibrium position at nonzero z. Cognate sites have a single minimum at $z = 0$. The third group consists of the so-called “traps,” a set of sites that differ in sequence from the cognate ones at a few nucleotides. The traps have equilibrium positions separated by a barrier, at both $z = 0$ and $z \neq 0$. A protein moving in such a landscape would spend most of the time sliding inside a “gutter” of a nearly parabolic cross-section that varies slightly from site to site and would quite rarely get stuck at $z = 0$. The gutter thus corresponds to the non–specific binding mode, while $z = 0$ describe the specific one. Both α and z_0 can be readily estimated from few simple considerations. First, we note that in the non–specific mode, the variations in the potential along x should be small enough, which places a lower bound on z_0. Second, we use energetic parameters known from the experiments.

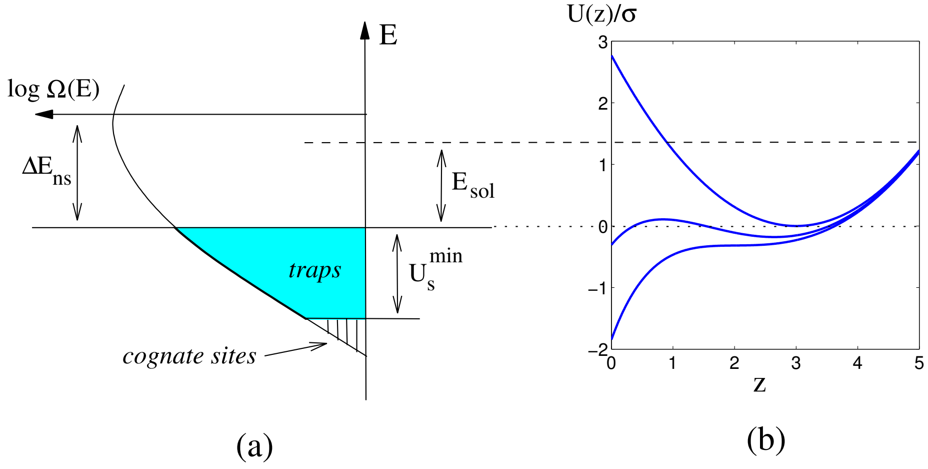


Figure 2-5: (a) Spectrum of protein–DNA binding energies; (b) Potential profiles for various binding sites with different $U_s(x)$.

A typical situation is shown in Fig. 2-5. The spectrum of specific binding energies is described by a Gaussian with standard deviation $\sigma \simeq 6 - 6.5 k_B T$. Cognate sites reside $U_s \simeq 15 - 20 k_B T$ below this threshold [24]. One can also estimate the change in the non–specific binding energy ΔE_{ns} as the protein converts between the two modes [46] which is usually positive and amounts to $\sim 15 - 20 k_B T$. Naturally, this observed one–dimensional spectrum should be obtainable from the 2D landscape by projecting the latter along z in some way, as shown in Fig. 2-5. If e^{-z_0} is small enough, we can estimate

$$\Delta E_{ns} \approx \frac{1}{2}\alpha z_0^2. \tag{2.15}$$

Also, to discriminate between the traps and the cognate sites, we require that for a cognate site the force is always in the negative z–direction

$$\frac{\partial U(x,z)}{\partial z} \geq 0 \qquad \text{for } z > 0, \tag{2.16}$$

where the equality occurs for some $z < z_0$ and x such that $U(x,0) = U_s^{min}$. For given ΔE_{ns} and $U_s^{\min}$, we solve

$$\alpha \left[1 - \ln\left(\frac{\alpha}{U_s^{min} + \Delta E_{ns}}\right)\right]^2 = 2\Delta E_{ns} \tag{2.17}$$

for α and then obtain z_0 from Eq. (2.15).

2.3.2 Dynamics on 2D Landscape

As we mentioned above, there is much experimental evidence in favor of describing protein–DNA association kinetics as intermittent rounds of 1D and 3D diffusion. The 1D diffusion distance n is controlled by a free energy barrier E_{ns} (see Chapter 1), which is merely a difference between the free energy of a protein in solution (or cytoplasm) and that of a nonspecifically bound protein

$$n \sim \exp\left(\frac{E_{ns}}{2k_BT}\right). \tag{2.18}$$

Thus, for $E_{ns} \sim 9 - 12k_BT$, we have $n \sim 100 - 400$ bp.

Within our model, the observed 1D dynamics of the protein are a projection of its motion on the 2D energetic landscape. We assume that the motion in both x– and z–directions is overdamped with well–defined diffusion coefficients D_x and D_z, respectively. To tackle the problem numerically, we put the landscape on a lattice and use a version of the Gillespie algorithm [38, 39]. The protein at a given lattice site can undergo 5 possible "reactions:" it can move in positive or negative x– or z–

directions or it can dissociate from the DNA and reassociate at some other (random) lattice site. The rate of a reaction $\delta \to \gamma$ is calculated as according to

$$\omega_{\delta\to\gamma} = \frac{1}{\tau_{\delta\gamma}} \times \begin{cases} e^{-\beta\Delta U_{\delta\gamma}} & \text{if } \Delta U_{\delta\gamma} > 0 \\ 1.0 & \text{otherwise} \end{cases}, \tag{2.19}$$

where $\beta = 1/k_BT$, $\tau_{\delta\gamma} = \tau_{\gamma\delta}$ is the relevant time constant and $\Delta U_{\delta\gamma}$ is the energy barrier for the reaction. For movements on the landscape, this is the difference in the potential at the neighboring lattice sites; for dissociation from a site (i,j), $\Delta U_{\delta\gamma} = E_{\text{sol}} - U(i,j)$. Longitudinal motion and dissociation reactions have a timescale $\tau_{\delta\gamma} = \tau_x = (2D_x)^{-1}$, whereas for transversal motion, $\tau_{\delta\gamma} = \tau_z = (\Delta z)^2/2D_z$, Δz is the lattice spacing in the z-direction. Each Monte–Carlo (MC) move starts from choosing the next reaction at random, with each reaction weighted proportionally to its rate (2.19). Then, the time to next reaction is drawn from the exponential distribution with a mean

$$\langle \Delta t \rangle = \left(\sum_{\gamma} \omega_{\delta\to\gamma} \right)^{-1}. \tag{2.20}$$

If the next reaction is dissociation, the total search time is incremented by a 3D diffusion time τ_{3d} [47].

The potential profile $U_s(x)$ of *E. coli* genome was built by a standard weight-matrix method [16, 25], using a known set of *PurR* transcription factor binding sites [48]. Landscape parameters were chosen to fit $|U_s^{\min}| \simeq 12k_BT$, $\Delta E_{ns} \simeq 18k_BT$, i.e. $\alpha = 4.0k_BT$ and $z_0 = 3.0$. Also, from the previous work [24, 47] it is known that $D_x \sim 1-10\mu\text{m}^2/\text{sec}$ and $\tau_{3d} \sim 10^{-3}$sec. However, no reliable order–of–magnitude estimates are possible for D_z unless the collective coordinate z has been defined explicitly in terms of coordinates and masses of all participating particles. Thus, in what follows, we fix D_x at some predefined value and study various aspects of the target location kinetics for different values of $\kappa \equiv D_x/D_z$.

As we mentioned above, any effective model should allow for both rapid target location and stable cognate complexes. Figure 2-6 shows the dependence of the mean target location t_{loc} time as a function of κ for a short ($M = 10^3$ bp) stretch of DNA.

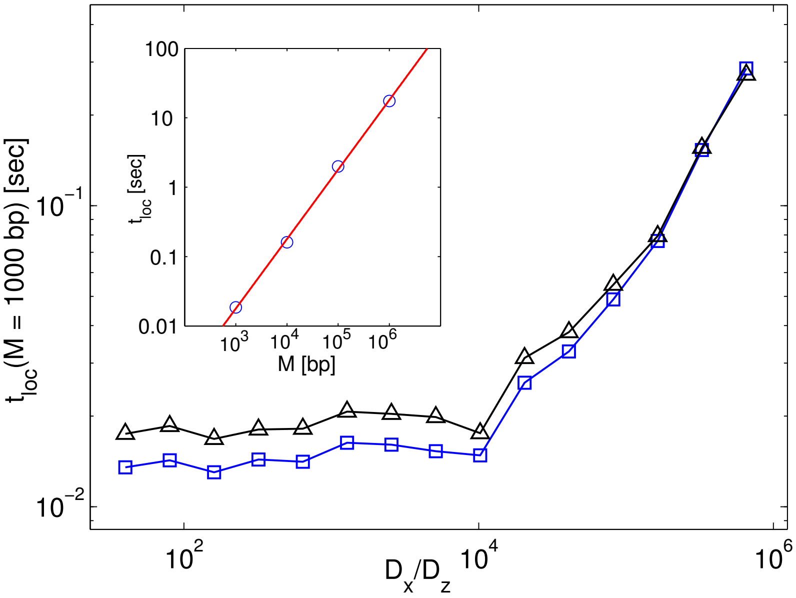


Figure 2-6: Average target location time t_{loc} as a function of D_x/D_z for $M = 10^3$ bp measured directly from MC simulations (squares) and estimated from the number of "jumps" (triangles). Inset: scaling of t_{loc} with genome size M for $\kappa = D_x/D_z = 10^2$; the slope of the line is equal 1.

One can see that $t_{\rm loc}$ is practically constant for $\kappa < 10^4$ and blows up very fast when $\kappa > 10^4$. This behavior can be explained as follows. Before the protein finds its target, a significant part of the genome becomes "covered" by segments of effectively one–dimensional diffusion, each containing $n = 100 - 200$ base pairs. Most of the time, the protein is sliding in the nonspecific mode. If the diffusion in the z–direction is fast enough, the protein is able to locate the target each time its segment "covers" the target site. The protein returns to the same location inside each segment 10^2 times and thus the requirement of fast transverse diffusion is relaxed significantly. It is clear that there must be some characteristic κ_c, at which the protein starts missing the target site before it dissociates from the segment and should wait for another return to the vicinity of the target. To demonstrate this point, we monitor the mean number of dissociation–reassociation events (or "jumps") n_j before target location. The mean time to location can be estimated as

$$t_{\rm loc} \simeq n_j(\tau_{3d} + \overline{\tau_{sl}}), \qquad (2.21)$$

where $\overline{\tau_{sl}}$ is the mean time the protein spends sliding over the landscape before dissociation. As one can see from Fig. 2-6, this estimate is very close to the mean location time measured from numerical simulations. To estimate κ_c, we compare the characteristic time of transverse diffusion

$$t_z \sim k_B T/(\alpha D_z), \tag{2.22}$$

to the longitudinal diffusion time $t_x \sim n^2/D_x$. At $\kappa \sim \kappa_c$, we should have $t_x \sim t_z$ and thus

$$\kappa_c \sim \alpha n^2/(k_B T). \tag{2.23}$$

Plugging in the numbers, we get $\kappa_c \sim 10^4 - 10^5$, as observed. Figure 2-6 also shows that target location time scales linearly with the size of the genome M (see inset), and thus the above argument holds for more realistic genome sizes of $M \sim 10^6$ bp.

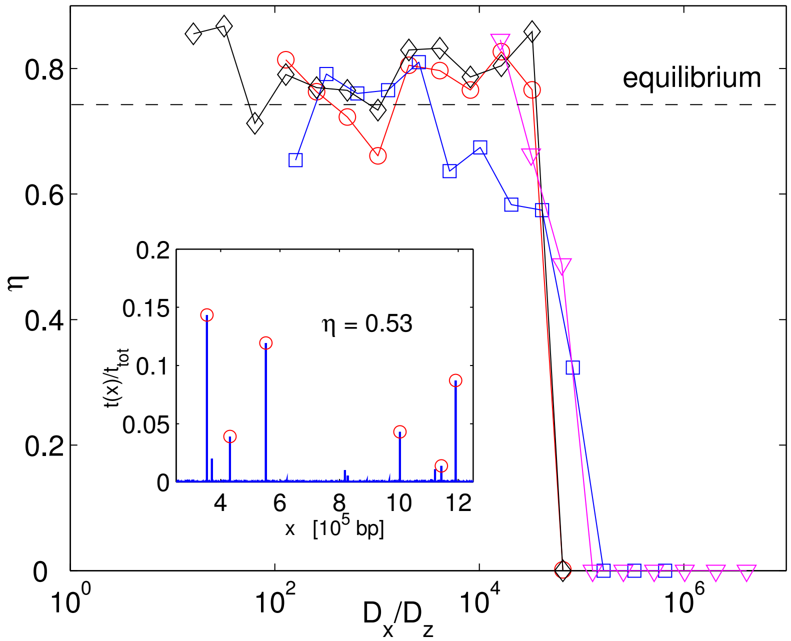


Figure 2-7: Specific mode occupancy η as a function of D_x/D_z for a $M = 10^5$ bp stretch of *E. coli* genome. Different symbols correspond to different values of Δz. Inset: fraction of time spent at different sites for $M = 10^6$ bp, $\kappa = 16$; circles designate real binding sites.

Next, we explore the stability of cognate complexes in our model. For that purpose, we define the specific mode occupancy η as the fraction of the time spent in the specific mode ($z \leq 0.25$). Figure 2-7 shows η calculated for a set of runs with $M = 10^5$ bp. On this stretch, there was a single cognate site for *PurR* binding. The calculated total duration of each run was at least 60 sec; the runs were performed for different values of lattice spacing Δz. We see that for $\kappa < \kappa_c$, the protein was able to both locate the site and explore a large part of the "genome", so that η fluctuates close to its equilibrium value. Above κ_c, the protein cannot accomplish the search in 60 sec and thus $\eta = 0$. We therefore conclude that for finite times, *measured* stability is also influenced by the exact value of κ. Figure 2-7 also shows a typical occupancy profile for $M = 10^6$ bp at $\kappa = 16$. The calculated total duration of the run was ~ 30 sec. We see that the protein has effectively located many cognate sites[1] and that the total specific mode occupancy $\eta \simeq 0.53$ is close to the equilibrium one ($\eta_{\text{eq}} \simeq 0.58$). The average occupancy of each site, especially in the case of few tens to hundreds of protein molecules per cell, is thus sustained at a near–equilibrium limit. However, in this limit of fast diffusion (and equilibration) along the z–direction, each single protein molecule stays at a cognate site for just a few seconds. This is probably not the best property from the functional point of view and thus we conclude that at $\kappa \sim 10^2 - 10^3$ we may expect a more robust though slightly slower performance. In real time units this corresponds to mode interconversion times $t_z \sim 10^{-5} - 10^{-4}$ seconds. It is noteworthy that, in principle, it is possible to have $\kappa \ll 1$. But that limit is virtually inaccessible to PC simulations of reasonable duration and is more appropriate for molecular dynamics studies. Also, for the reasons mentioned above, this limit is unlikely to be of practical interest.

[1] The efficiency of *real* cognate sites discovery is at most as good as in the underlying model for $U_s(x)$.

2.4 Biological implications

2.4.1 Specificity "for free": kinetics vs. thermodynamics.

The proposed mechanism of specific site location is akin to kinetic proofreading [49], which is a very general concept for a broad class of high-specificity biochemical reactions. In kinetic proofreading, the required specificity is achieved through formation of an intermediate metastable complex that paves the way for an irreversible enzymatic reaction. If the reaction is much slower than the life-time of the complex, then substrates that spend enough time in the complex are subject to the enzymatic reaction, while substrates that form short-lived complexes are released back to the solvent before the reaction takes place. In other words, the substrates are selected by kinetic partitioning.

2.4.2 Coupling of folding and binding in molecular recognition

Several DNA- and ligand-binding proteins are known to have partially unfolded (disordered) structures in the unbound state. The unstructured regions fold upon binding to the target. Does binding-induced folding provide any biological advantage?

The idea of coupling between local folding and site binding has been around for some time and was recently reassessed in the much broader context of intrinsically unstructured proteins [50, 51, 52]. Induced folding of these proteins can have several biological advantages. First, flexible unstructured domains have intrinsic plasticity allowing them to accommodate targets/ligands of various size and shape. Second, free energy of binding is required for compensation for entropic cost of ordering of the unstructured region. A poor ligand that doesn't provide enough binding free energy cannot induce folding and, hence, can not form a stable complex. Williams *et al.* have suggested that unstructured domains can be result of evolutionary selection that acts on the bound (structured) conformation, while ignoring the unbound (unstructured) conformation [53]. Partial unfolding can also increase protein's radius of gyration and, hence, increase the binding rate [54, 55]

Here we propose a mechanism that suggests a role for induced folding in providing rapid and specific binding. Induced folding (or other sorts of two-state conformational transitions) allows a protein to search and recognize DNA in two different conformations, providing rapid binding to the target site. Importantly, this mechanism reconciles rapid search for the target site with stable bound complex. The rate of induced folding can also play a role in determining the specificity of recognition.

Structural and thermodynamic data argue in favor of distinct protein conformations for search along non-cognate DNA and for recognition of the target site. Proteins such as λ cI, Eco RV and GCN4 apparently do not fold their unstructured regions while bound to non-cognate DNA [56, 57, 58], supporting our hypothesis.

Heat capacity measurements on a vast variety of protein-DNA complexes report a large negative heat capacity change in site-specific recognition, which is a clear indication of a phase transition. These measurements supplemented by X-ray crystallography and NMR structural data were interpreted by Spolar *et al*. [11] mainly in terms of hydrophobic and conformational contributions to entropy. Thus, folding-binding coupling is now considered a well-established feature of a large set of transcription factors.

However, real-time kinetic measurements were not performed until recently, so the question of the actual mechanism was left open. Major advances in this direction were made by Kalodimos *et al*. [59, 60, 36], who observed a two-step site recognition by dimeric *Lac* repressor. The H/D-exchange NMR data unambiguously demonstrate site pre-selection by α-helices bound in the major groove followed by folding of hinge helices that bind to the minor groove elements and complete the specific site recognition. Though the experiments in this field were performed with a single model system, their implications are likely to have a general character.

Chapter 3

The long reach of DNA heterogeneity.

Until now, we assumed that the potential profile of protein-DNA interaction is uncorrelated, i.e. the values of the sequence–specific binding energy at two neighboring sites are independent of each other. In this chapter, we analyze the influence of correlations of arbitrary range in the potential profile on the 1D diffusion.

3.1 Diffusion in a correlated random potential

3.1.1 The model

The random walk is characterized by the set of hopping probabilities that are derived from the random potential U_i, which is the sequence-dependent component of the potential energy. The latter is basically a sum of many random contributions and can therefore be considered to be normally distributed [24]. Thus, in the absence of correlations, the probability for realization of a certain profile $U(x)$ of length L is (in the continuum limit)

$$P[U(x)] \propto \exp\left[-\frac{\alpha}{2}\int_0^L dx\, U^2(x)\right]. \tag{3.1}$$

This is the well-known Random Energy Model [61], which has been applied successfully to various biophysical problems, from protein folding [62] to protein-DNA interaction [24]. It assumes no correlations between energies of different sites. One can think of a more general form of potential profile

$$P[U(x)] \propto \exp\left[-\frac{1}{2}\int_0^L\int_0^L dydx\; U(x)G(x-y)U(y)\right]. \tag{3.2}$$

Taking for example, $G(x-y) \propto \partial^2_{xy}\delta(x-y)$, we obtain the Random Force Model [63], which describes an energy landscape as a random walk with linearly growing correlations. This model was studied during the last decades in the context of heteropolymer dynamics [64, 65], glassy systems [66, 67] and quite recently to describe DNA denaturation dynamics [68]. Characteristic features of the Random Force Model are logarithmically slow ("Sinai's") diffusion [69, 70] and aging [67, 68]. More generally, G is related to the correllator of U by $\langle U(x)U(y)\rangle = G^{-1}(x-y)$.

To include finite-range correlations into Eq. (3.1), we must incorporate a limitation on the acceptable forces. The ensemble of energy profiles is therefore naturally described by the following probability density

$$P[U(x)] \propto e^{-\mathcal{H}[U]}, \tag{3.3a}$$

with *pseudoenergy*

$$\mathcal{H}[U] = \frac{1}{2}\int_0^L dx\;\left[\alpha U^2(x) + \gamma\left(\frac{dU}{dx}\right)^2\right]. \tag{3.3b}$$

Energy level statistics for this kind of potential profile is also Gaussian, as can be seen from the average

$$\left\langle e^{ikU}\right\rangle = \frac{\displaystyle\int \mathcal{D}[U]e^{ikU}e^{-\mathcal{H}[U]}}{\displaystyle\int \mathcal{D}[U]e^{-\mathcal{H}[U]}} = \exp\left(-\frac{k^2}{4\sqrt{\alpha\gamma}}\right), \tag{3.4}$$

which is the characteristic function for Gaussian distribution with zero mean and variance

$$\sigma^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{dq}{\alpha+\gamma q^2} = \frac{1}{2\sqrt{\alpha\gamma}}. \tag{3.5}$$

The correlator of the potential profile is readily calculated as

$$g(r) \equiv \frac{1}{2}\langle[U(x)-U(x+r)]^2\rangle = \sigma^2\left(1-e^{-|r|/\xi_c}\right), \tag{3.6}$$

where $\xi_c = \sqrt{\gamma/\alpha}$ is the correlation length.

3.1.2 Mean First Passage Time

The hopping probabilities are defined as

$$p_i = \frac{\omega_{i,i+1}}{\omega_{i,i+1}+\omega_{i,i-1}}, \qquad\qquad q_i = 1-p_i, \tag{3.7}$$

where, as before

$$\omega_{i,i\pm1} = \nu \times \begin{cases} e^{-\beta(U_{i\pm1}-U_i)} & \text{if } U_{i\pm1} > U_i \\ 1.0 & \text{otherwise} \end{cases}. \tag{3.8}$$

The disorder-averaged version of the MFPT is readily obtained after we note that the sequential products in Eq. (1.18) reduce to

$$\prod_{j=k}^{i}\alpha_j = \exp\left[\beta(U_i-U_k)\right]. \tag{3.9}$$

For an uncorrelated potential profile, this exponential factorizes into independent exponentials; after the ensemble averaging and the summations are carried out, we obtain for $N \gg 1$

$$\langle \bar{t}_{0,N}\rangle = N^2 e^{\beta^2\sigma^2}, \tag{3.10}$$

where, for the uncorrelated potential ($\gamma = 0$)

$$\sigma^2 = \frac{1}{\alpha a}. \tag{3.11}$$

Note that this expression cannot be obtained by simply putting $\gamma = 0$ in Eq. (3.5), since the discrete nature of the underlying lattice (the DNA) starts to matter when γ becomes small. The integration in the Fourier space in Eq. (3.5) extends only up to $|q_{\max}| = \pi/a$, and thus,

$$\sigma^2|_{\gamma\to 0} = \int_{-\pi/a}^{\pi/a} \frac{dq}{2\pi\alpha} = \frac{1}{\alpha a}. \tag{3.12}$$

Returning to the case of a finite correlation length, we calculate

$$\begin{aligned}\left\langle e^{\beta(U(x)-U(y))}\right\rangle &= \frac{\displaystyle\int \mathcal{D}[U] e^{\beta(U(y)-U(x))} e^{-\mathcal{H}[U]}}{\displaystyle\int \mathcal{D}[U] e^{-\mathcal{H}[U]}} \\ &= \exp\left[\frac{\beta^2\xi_c}{2\gamma}(1 - e^{-|x-y|/\xi_c})\right]. \end{aligned}\tag{3.13}$$

For $|x-y| \ll \xi_c$, Eq. (3.13) reduces to $\exp(\beta^2|x-y|/2\gamma)$, so that for $N \ll \xi_c$ we have

$$\langle \bar{t}_{0,N}\rangle \sim N^2 \exp(\beta^2\sigma^2 N/\xi_c) = N^2 \exp(\beta^2 N/2\gamma). \tag{3.14}$$

(Here and in what follows, we measure distances in units of a, unless specified otherwise.) This kind of exponential creep is quite expected, since for $\alpha \to 0$, $\xi_c \to \infty$ our model (3.3) reduces to the Random-Force Model.

In the opposite limit $|x - y| \gg \xi_c$, we can neglect the exponent $e^{-|x-y|/\xi_c}$, so that Eq. (1.18) produces an ordinary diffusion law, with a disorder-renormalized diffusion coefficient:

$$\langle \bar{t}_{0,N}\rangle = N^2 e^{\beta^2\sigma^2}. \tag{3.15}$$

3.2 Typical versus average

Large deviations from the average are characteristic to many disordered systems. In this section, we therefore explore the *typical* properties of random walks as compared to the disorder-averaged ones.

3.2.1 Potential profile generation

To numerically explore the diffusion in correlated potentials, we set up the following procedure for generating potential profiles.

Given the pseudoenergy partition function

$$Z(\lambda) = \int \mathcal{D}[U] e^{-\lambda \mathcal{H}[U]}, \tag{3.16}$$

the average pseudoenergy is

$$\langle \mathcal{H} \rangle = -\frac{\partial}{\partial \lambda} \ln Z(\lambda) \bigg|_{\lambda=1}, \tag{3.17}$$

and the variance is

$$\langle (\Delta \mathcal{H})^2 \rangle = \langle \mathcal{H}^2 \rangle - \langle \mathcal{H} \rangle^2 = \frac{\partial^2}{\partial \lambda^2} \ln Z(\lambda) \bigg|_{\lambda=1}. \tag{3.18}$$

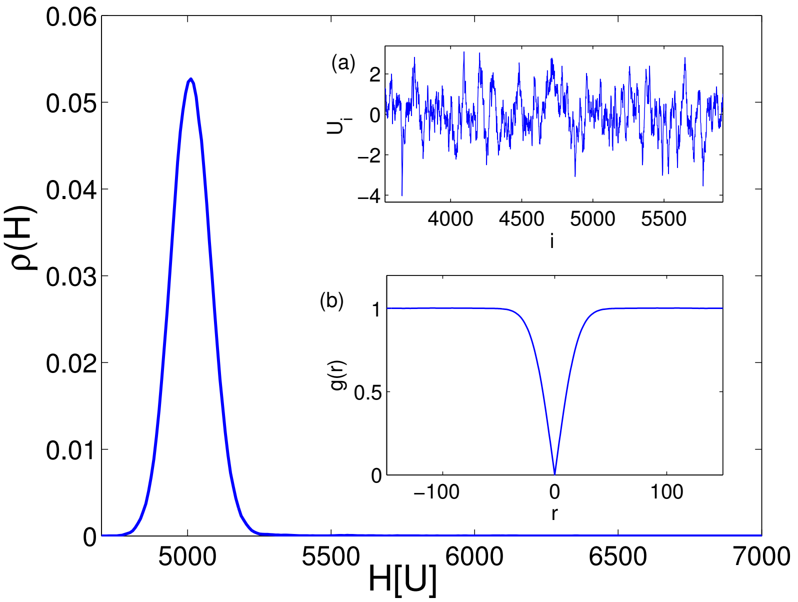


Figure 3-1: Pseudoenergy probability density for a profile of length $L = 10000$, with $\sigma = 1.0$, $\xi_c = 20.0$. Insets: (a) Typical potential profile; (b) Potential profile correllator $g(r) = 1/2\langle[U(x) - U(x + r)]^2\rangle$; the averaging was performed over 1000 profile realizations.

Straightforward calculation for the pseudoenergy given by Eq. (3.3) yields

$$Z(\lambda) = \prod_q \frac{1}{\sqrt{2\pi\lambda\,(\alpha + \gamma q^2)}}. \tag{3.19}$$

Since a discrete chain of length L has exactly L modes, each contributing a factor of $\lambda^{-1/2}$, we have

$$\ln Z(\lambda) = -\frac{L}{2}\ln\lambda + A, \tag{3.20}$$

where A does not depend on λ. Thus,

$$\langle\mathcal{H}\rangle = L/2, \qquad \langle(\Delta\mathcal{H})^2\rangle = L/2. \tag{3.21}$$

Hence, typical potential profiles have pseudoenergies in the range $L/2 \pm \sqrt{L/2}$. This result together with Gaussian statistics of energy levels of Eq. (3.4) forms the basis of the algorithm we employ for building the energy profiles. First, a random and uncorrelated potential profile obeying Gaussian statistics with the required variance σ^2 is generated on a one-dimensional lattice. Next, we look for a permutation of lattice sites that produces a typical pseudoenergy $\mathcal{H}[U]$ for a given correlation length ξ_c (or, equivalently, for given values of α and γ). This is accomplished by a Metropolis-type algorithm that converges to a prescribed value of pseudoenergy picked at random from Gaussian distribution around $\langle\mathcal{H}\rangle$ (see Fig. 3-1).

3.2.2 Quantifying fluctuations

After the potential profile is generated, we calculate the MFPT using Eq. (1.18). Fig. 3-2 presents the mean first passage times calculated for various realizations of $U(x)$ at biologically relevant temperature ($\sigma \simeq k_BT$). It is clear that although the ensemble-averaged MFPT does behave as prescribed by Eq. (3.15), typical MFPT exhibits high variability from one profile to another. The stepwise shape of typical curves suggests that a random walk in such a profile consists of intermittent regions of subdiffusion (vertical "steps") and superdiffusion (plateaus).

To quantify the sample dependence of the MFPT, we calculate its variance over

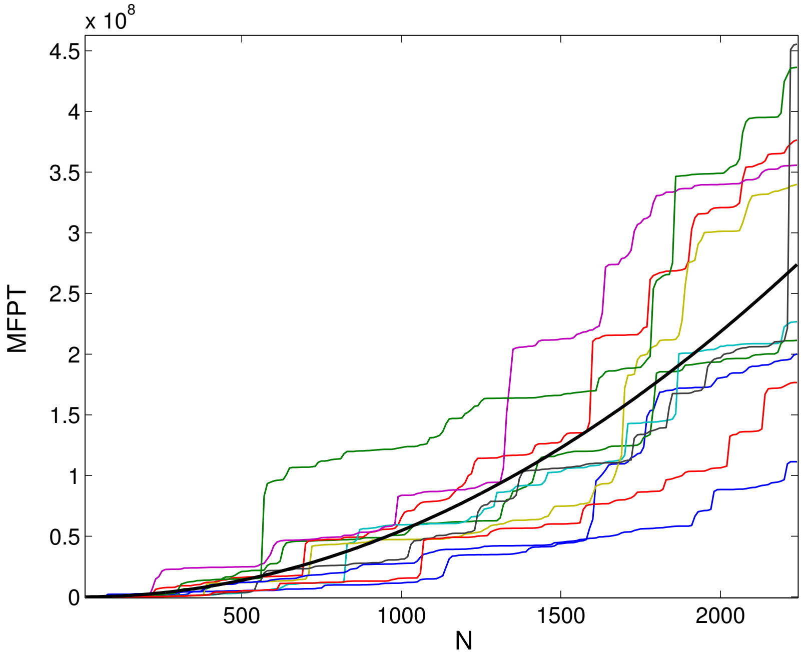


Figure 3-2: Mean First Passage Times: typical versus average. Thick solid line is the result of averaging over 1000 realizations of correlated potential profiles with $\beta\sigma = 2.0$, $\xi_c = 20.0$.

the ensemble of potential profiles. Fig. 3-3 presents the standard deviation in $\bar{t}_{0,N}$ as a function of N for correlated as well as uncorrelated potential profiles. We observe that the variance scales as N^3 for all profiles. This dependence can be obtained analytically in a quite straightforward fashion. The MFPT is given by Eq. (1.18); then

$$\langle(\Delta\bar{t}_{0,N})^2\rangle \simeq 4\int_0^N dx \int_x^N dy \int_0^N dx' \int_{x'}^N dy' \left[\langle e^{\beta(U(x)-U(y)+U(x')-U(y'))}\rangle \right.$$
$$\left. -\langle e^{\beta(U(x)-U(y))}\rangle\langle e^{\beta(U(x')-U(y'))}\rangle\right]. \qquad (3.22)$$

We now recall that energies at points separated by distances larger than ξ_c are essentially independent. Therefore, to estimate the averages, we assume the energies to be equal for points within one correlation length and independent otherwise. The first average in the integral produces

$$
\begin{aligned}
\langle e^{\beta(U(x)-U(y)+U(x')-U(y'))}\rangle &\simeq \langle e^{\beta(U(x)-U(y))}\rangle\langle e^{\beta(U(x')-U(y'))}\rangle \\
&+ \xi_c\delta(x-x')\langle e^{\beta(2U(x)-U(y))}\rangle\langle e^{-\beta(U(y'))}\rangle \\
&+ \xi_c\delta(y-y')\langle e^{-\beta(2U(y)-U(x))}\rangle\langle e^{\beta(U(x'))}\rangle \\
&+ \xi_c^2\delta(x-x')\delta(y-y')\langle e^{-2\beta(U(y)-U(x))}\rangle + ... \\
&\simeq \langle e^{\beta(U(x)-U(y))}\rangle\langle e^{\beta(U(x')-U(y'))}\rangle \\
&+ \xi_c e^{3\beta^2\sigma^2}\left[\delta(x-x')+\delta(y-y')\right] \\
&+ \xi_c^2 e^{4\beta^2\sigma^2}\delta(x-x')\delta(y-y') + ... \,. \qquad (3.23)
\end{aligned}
$$

Plugging this expression into Eq. (3.22) and performing the integrations, we obtain the leading term

$$\langle(\Delta\bar{t}_{0,N})^2\rangle \sim \xi_c N^3 e^{3\beta^2\sigma^2}. \qquad (3.24)$$

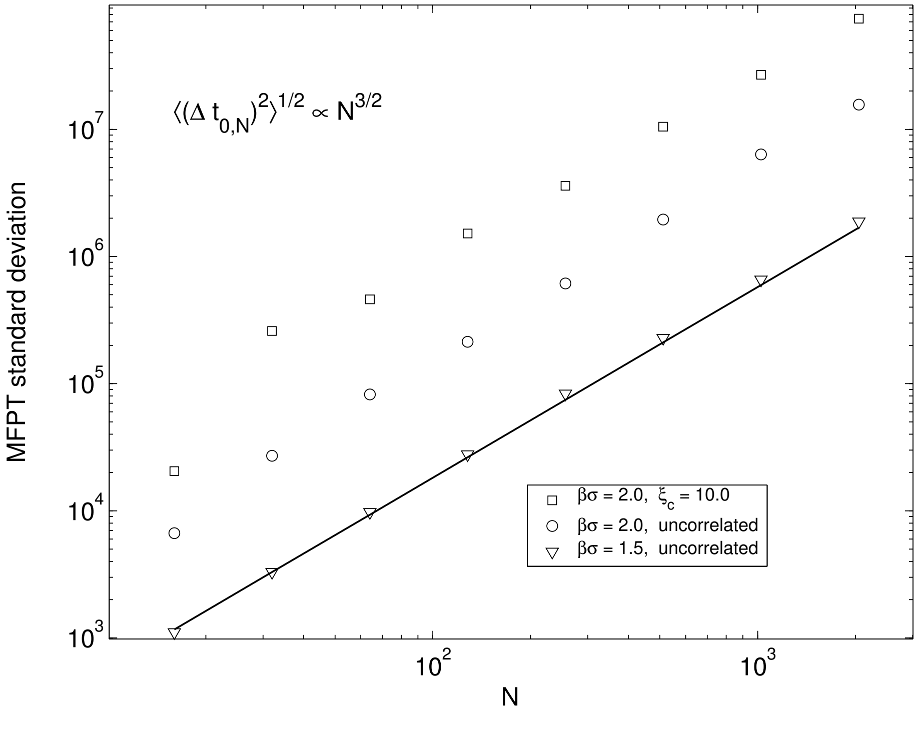


Figure 3-3: MFPT standard deviation for $\beta\sigma = 1.0$ for correlated and uncorrelated potential profiles.

Comparing the expressions for the variance with the corresponding expressions for disorder-averaged MFPT, we see that for any temperature, there is a characteristic distance N_c, below which there is no self-averaging and the typical MFPT is determined by fluctuations. This length is

$$N_c \sim \xi_c e^{\beta^2\sigma^2}. \tag{3.25}$$

This effect is akin to "freezing" in the Random Energy Model [61] – for low enough temperatures, typical passage times for distances below N_c are dominated by high barriers. This is more pronounced for correlated profiles due to amplification by a factor of $\sim \xi_c$, as sites within a correlation length give similar contributions. Figure 3-4 demonstrates the lack of self-averaging for uncorrelated potential profiles at short distances and low temperatures: the *median* MFPT (defined as the 50th percentile of a sample) shows large deviations from the average at distances shorter than N_c and coincides with it at distances larger than N_c.

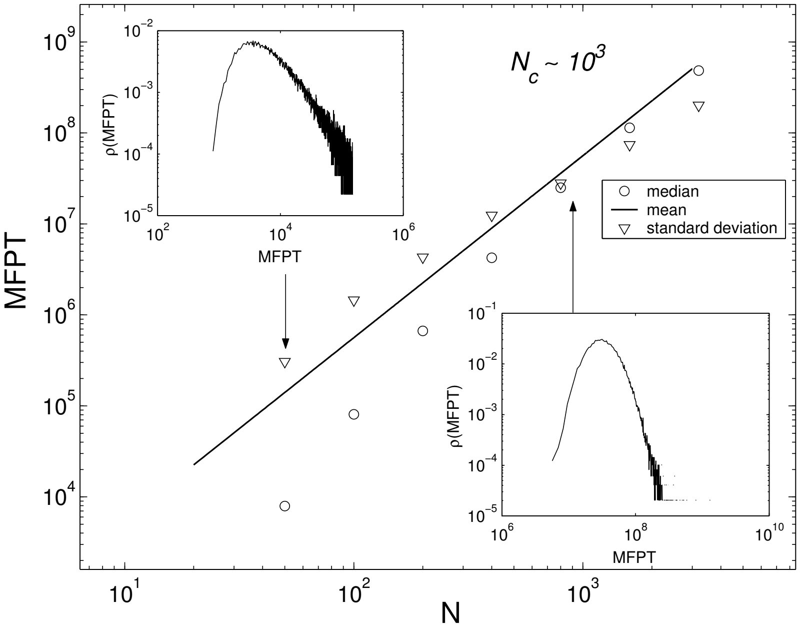


Figure 3-4: Probability density functions for MFPT calculated for 100,000 uncorrelated profile realizations at $\beta\sigma = 2.0$.

Large differences between the median and the average values are a signature of a broad ("fat-tailed") asymmetric probability distribution. The insets of Fig. 3-4 present two probability density functions for MFPT, at $N \ll N_c$ and $N \gg N_c$. For the short distance, the distribution is very broad and spans several orders of magnitude. For $N \gg N_c$, the system is self-averaging, in the sense that the MFPT distribution is much narrower with almost identical median and average values.

3.3 Examples from biology

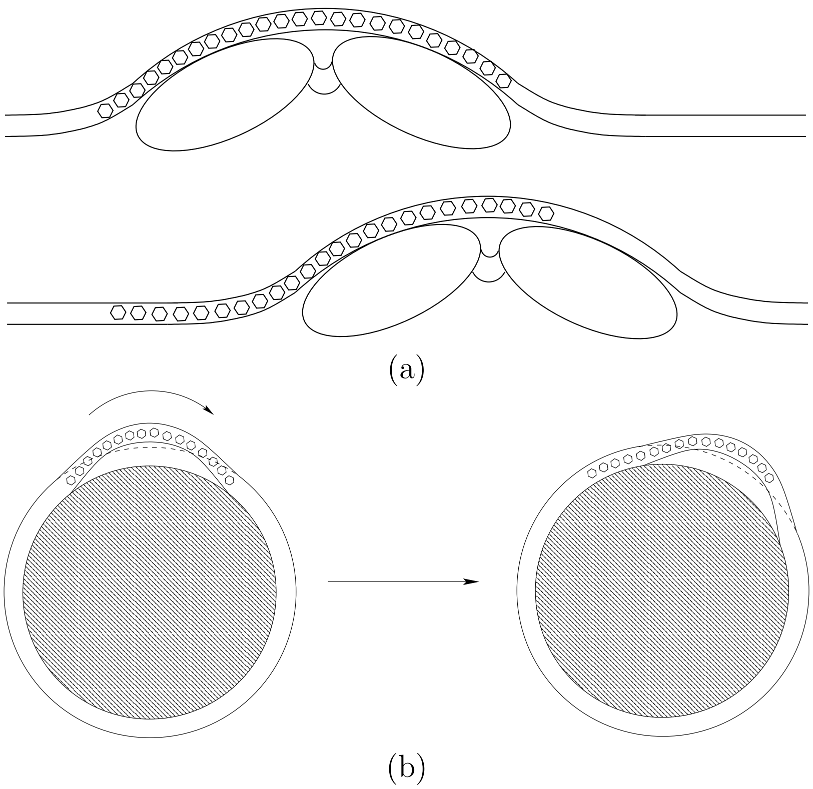
In this section, we study a few examples that demonstrate how energy landscapes with finite correlation length may appear in living cells.

3.3.1 DNA bending

Recent scanning force microscopy experiments by Erie *et al.* [26] clearly demonstrate DNA bending by the *Cro* repressor protein, both at operator and at non-operator sequences.[1] Since local DNA elasticity is known to be highly sequence-dependent [71], the energy of protein bound at random locations should have a random component, correlated at length scales of the order of the protein binding domain size; see Fig. 3-5a. This sequence-dependent interaction energy component appears in addition to possible local uncorrelated sequence-dependent contributions from amino acid-base pair contacts.

To estimate the significance of the random component of the elastic energy, we use DNA elasticity data supplied by the BEND.IT server [72], that incorporates DNase I based bendability parameters [73] and the consensus bendability scale [74]. We assume that the protein-DNA complex in Fig. 3-5a has a fixed geometry, i.e. the protein is "hard." Then, the elastic contribution to the protein-DNA interaction energy at the i-th sequence has a random component proportional to the random

[1] DNA bending by transcription factors is a well-known phenomenon, though practically all the available experimental data focus on proteins bound to operator sequences.



(a)

(b)

Figure 3-5: (a) Prokaryotic transcription factor sliding; (b) Nucleosome repositioning.

component of the Young's modulus δE_i

$$U_i = \left[1 + \frac{\delta E_i}{\bar{E}}\right] \left(\frac{\ell_p \theta^2}{2L}\right) k_B T, \tag{3.26}$$

where $\ell_p \simeq 50$ nm is the DNA persistence length, $\theta \simeq 60°$ is the curvature angle [26], $L = 10 - 20$ bp is the bent sequence length and $\bar{E} \simeq 3.4 \times 10^8$ N/m is the average Young's modulus. The resulting potential profile is plotted in Fig. 3-6a. The standard deviation of the elastic energy induced by the Young's modulus variations (10-15% typically) for biologically relevant parameters is $\langle(\delta U)^2\rangle^{1/2} \sim 0.5 - 1.5\ k_B T$, so that disorder appears to be relevant for this problem. Figure 3-6b shows the normalized energy-energy correlator for the random energy component

$$g(r) = \frac{1}{2\langle \delta U^2(x)\rangle} \langle [\delta U(x) - \delta U(x+r)]^2 \rangle, \tag{3.27}$$

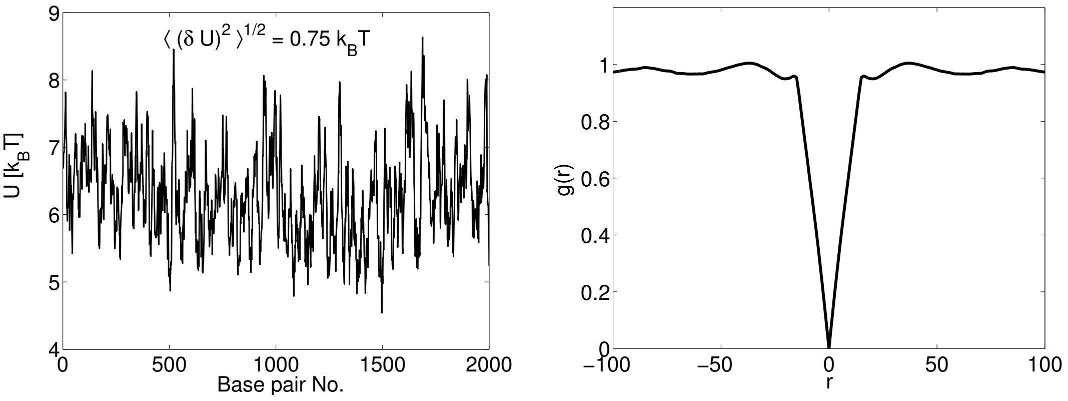
averaged over 10000 DNA sequences. Saturation to $g(r) \simeq 1$ is clearly observed on the scale of 15 base-pairs, which is the correlation length of this potential profile.

Another interesting example, also from the field of protein-DNA interaction, was considered recently by Schiessel *et al.* [75] and deals with nucleosome repositioning by DNA reptation. It was argued that chromatin remodeling [76, 77] can be readily understood in terms of intranucleosomal loop diffusion; the size of the loop resulting mainly from a compromise between elastic energy and nucleosome-DNA binding energy. Here again, for a given size of the loop, the elastic energy is sequence-dependent [77] and therefore has a random component with finite correlation length; see Fig. 3-5b. For nucleosome repositioning, this effect may be even more pronounced than for prokaryotic protein-DNA interaction; the bending angles θ and the sequence lengths L are 2-3 times larger so that the net effect may be twice as strong as for the *Cro* repressor [75].

It is known that DNA can have an *intrinsic* curvature arising from the stacking interactions between base pairs. Such sequence-dependent curvature can play a role similar to sequence-dependent DNA bendability in providing a correlated landscape. The bending energy of an intrinsically curved region is easier, requiring a smaller angular deformation $\theta = \theta_{\text{complex}} - \theta_{\text{intrinsic}}$ by the DNA-protein complex. Such sequence-dependent intrinsic curvature was suggested to be involved in positioning nucleosomes [78].

Aside from DNA bendability and curvature, local correlations in nucleotide composition, known to be present in eukaryotic genomes, (AT/GC-rich isochores) can result in a correlated landscape of the protein-DNA binding energy. This effect becomes especially pronounced when a DNA-binding protein has a strong preference toward a particular AT/GC composition of its site. However, in this case, variations take place over much longer scales, and are not quantitatively relevant in the specific contexts addressed in this paper.

Both above examples can be viewed as specific cases of DNA reptation by means of a propagating defect (or "slack") of a fixed size. Elastic energy associated with the slack creation is sequence-dependent and correlated on the scale of the slack



(a) (b)

Figure 3-6: (a) Energy of local elastic deformation and (b) Potential profile correlator, as calculated from the data supplied by the server BEND.IT for a segment of *E. coli* genome. The deformed DNA sequence is assumed to be of length $L = 15$ bp.

size. The propagating defect is well localized and samples the energies of well-defined subsequent DNA segments. As was pointed out by Cule and Hwa [65], short-range correlated randomness of this kind has no effect on the scaling of the reptation time. However, the defect motion itself is strongly influenced by the disorder and has non-trivial behavior at different length scales, as we demonstrate in Chapter 3.

3.3.2 DNA translocation through a nanopore

Consider a piece of single-stranded DNA (ssDNA) passing through a large membrane channel. If the potential difference across the membrane is zero, the motion of the ssDNA is governed by thermal fluctuations. Since the channel width differs from the ssDNA external diameter only by few Ångstroms[2], it is reasonable that local interactions between the nucleotides and the amino acids of the channel take place. These interactions may have a local base-dependent component. In addition, longer-range terms are likely to appear in the presence of a voltage difference. In the cytoplasm, the DNA negative charge is almost completely screened out at distances of few nanometers by the counterion cloud. When the DNA molecule enters the pore, most of the counterions are likely to be “shaved off,” though some of them may remain stuck to

[2] For α-haemolysin, the diameter of the limiting aperture is about 15 Å.

the DNA; see Fig. 3-7. Thus, the linear charge density inside the pore acquires a random and basically uncorrelated component:

$$q(x) = \bar{q}(x) + \delta q(x), \qquad \langle \delta q(x)\ \delta q(y) \rangle = \rho^2 a \delta(x-y), \tag{3.28}$$

where $a = 0.34$nm is the interbase distance. The potential energy of the DNA segment inside the pore in the presence of a voltage difference of V_0 is

$$U(x) = \frac{V_0}{h} \int_0^h x' q(x+x') dx'. \tag{3.29}$$

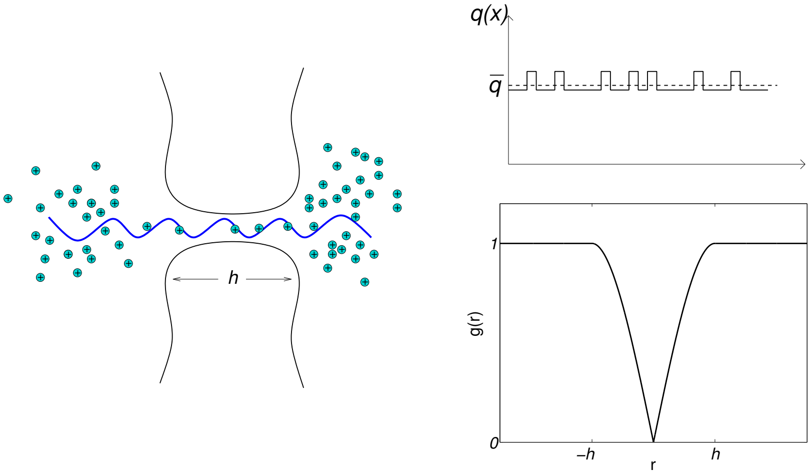


Figure 3-7: ssDNA transport through the nanopore; on the right: charge density $q(x)$ and correlator $g(r) = \langle [\delta U(x) - \delta U(x+r)]^2 \rangle / (2\langle \delta U^2(x) \rangle)$ as a function of the coordinate r.

Since the average charge density $\bar{q}(x)$ is in general nonzero, DNA transport is driven by the average force $V_0 \bar{q}(x)/h$. The correlation function of the random component of $U(x)$ is readily calculated to be

$$\langle \delta U(x) \delta U(x+y) \rangle = \frac{V_0^2 \rho^2 a}{3h^2} (h - |y|)^2 \left(h + \frac{|y|}{2} \right) H(h - |y|), \tag{3.30}$$

where $H(x)$ is the Heaviside function. Thus, the potential profile for DNA motion has a random component with correlation length of h. Taking $V_0 \sim 100$ mV, $\rho \sim e/h$ (e is the elementary charge), $h \sim 10$ nm, we obtain $\delta U \sim k_B T$.

Although this example differs from the above ones in that a nonzero average driving force is present, large random fluctuations of the energy landscape may have significant effect on the distribution of translocation times – a problem that has attracted much interest lately [79].

Part II

Random Walks and Polymer Statistics

Random walks are often used to model long polymer molecules. Proper description includes a nontrivial requirement: the random walk must be self–avoiding, i.e. no point in space should be visited more than once. This requirement introduces long–range correlations and makes the problem tractable only approximately. However, if the self–avoidance is removed, the problem becomes much simpler. Such a polymer is called *phantom*; alternative names are *ideal* or *Gaussian* chain. The distribution function for the end–to–end radius vector $\mathbf{r}$ of a phantom polymer of length N obeys the diffusion equation (B.1).

Rapid developments in polymer physics and the theory of critical phenomena have revealed a number of *universal* properties that arise in all polymer chains beyond a certain level of coarse–graining [80, 81]. These properties are characterized by *scaling relations*. Perhaps the most widely known scaling law relates the mean end–to–end distance R of a polymer chain to its length: $R \propto N^{\nu}$. The number $1/\nu$ thus plays the role of the polymer's fractal dimension. It is universal in that it depends only on the dimensionality of the embedding space, e.g. in 3D, $\nu \approx 0.59$ for a self–avoiding polymer and $\nu = 1/2$ for a phantom one. Another important scaling relation describes the number of different configurations $\mathfrak{N}$ of a polymer

$$\mathfrak{N} = \text{const} \times \zeta^{N} N^{\gamma-1},$$

where ζ is the "effective coordination number" that depends on the microsopic details and γ is a universal exponent. The factor ζ^N can be thought of as counting the configurations of an unconstrained N–step random walk with ζ options available at each step, whereas $N^{\gamma-1}$ accounts for constraints such as self–avoidance, obstacles present etc.

If the distribution function $G(\mathbf{r}, N)$ is known, the exponent γ can be obtained in a very straightforward way – by simply integrating $G(\mathbf{r}, N)$ over the whole space. Thus, a phantom polymer has $\gamma = 1$. It turns out that incorporating self–avoidance constraint leads to $\gamma \simeq 1.16$. This can be interpreted as the enhancement of available space for a self–avoiding polymer that appears "swollen" compared to a phantom one.

In Chapter 4, we show how to calculate the probability distribution function for a Gaussian polymer confined to a half–space using field theory methods. To say the least, this is not the most effective way of doing it. For instance, the above mentioned analogy to random walks can be exploited to obtain the answer in just a few lines, as we demonstrate in Appendix B. However, beside its clear educational value, the path integral analysis of random walks is a much more flexible and powerful tool when it comes to real systems, like polymers in solvents. The self-avoidance constraint introduces long-range interactions that make the traditional approaches lose their power and elegance. For example, the factorization of $G(\mathbf{r}, N)$ into transverse and longitudinal parts is no longer valid. Integrating out one set of degrees of freedom introduces non-local interactions into the other. Although lattice walks are useful for Monte Carlo simulations and diffusion equations could perhaps be modified to include mean-field corrections, field theory, specifically the renormalization group, is presently the only analytical tool for obtaining universal scaling relations and phase diagrams using controlled approximations [82]. As an example, we consider in Chapter 5 the problem of self–avoiding polymers attached to the tip of an impenetrable probe. We find that the scaling exponents γ_1 and γ_2, characterizing the number of configurations for the attachment of the polymer by one end, or at its midpoint, vary continuously with the tip's angle. These apex exponents are calculated analytically by ϵ-expansion and numerically by simulations in three dimensions. We find that when the polymer can move through the attachment point, it typically slides to one end; the apex exponents quantify the entropic barrier to threading the eye of the probe.

Chapter 4

Diffusion in a half–space via path integrals

4.1 Example: unconstrained Gaussian chain

In the field theoretic approach, a flexible chain is described by a function (path) $\mathbf{c}(\tau)$, where τ measures the position along the chain. The energy of a self-avoiding chain in an external potential is given by [81, 82]

$$H[\mathbf{c}] = \frac{1}{2}\int_0^N \dot{\mathbf{c}}^2(\tau)\, d\tau + \int_0^N U[\mathbf{c}(\tau)]d\tau + \frac{v}{2}\int_0^N \int_0^N \delta[\mathbf{c}(\tau) - \mathbf{c}(\tau')]d\tau d\tau'. \tag{4.1}$$

The first two terms can be viewed as a harmonic potential between neighboring segments of the chain and the external potential, respectively, whereas the last one accounts for excluded volume effects: each time the chain self-intersects, a penalty of v is paid. In what follows, we omit the self-avoidance constraint. The partition function for such a chain is a sum over all possible paths $\mathbf{c}(\tau)$ given by a path integral

$$\mathcal{Z}(N) = \int D[\mathbf{c}(\tau)]\, e^{-H[\mathbf{c}]}. \tag{4.2}$$

If the external potential is set to zero, Eq. (4.2) counts the number of configurations of a Gaussian chain of length N. The configurations are weighted by $e^{-H[\mathbf{c}]}$. If we want to count only paths starting at the origin and leading to some point $\mathbf{r}$, then after normalization by $\mathcal{Z}(N)$, we obtain the probability density

$$G(\mathbf{r}, N) = \frac{1}{\mathcal{Z}(N)} \int D[\mathbf{c}(\tau)] \delta[\mathbf{c}(N) - \mathbf{c}(0) - \mathbf{r}] e^{-H[\mathbf{c}]}. \tag{4.3}$$

To calculate this integral, we have to define a measure of integration. One way would be to discretize the chain and view $\mathcal{Z}(N)$ as a limit of a multidimensional integral. In this case, the problem is almost identical to calculating a quantum free particle propagator by path integration as was done in Refs. [83] and [84]. Another way is to integrate over the Fourier components of $\mathbf{c}(\tau)$; this way is somewhat easier because the degrees of freedom decouple in Fourier space for harmonic Hamiltonians. Thus, if we set

$$\int D[\mathbf{c}(\tau)] \to \int \prod_q \frac{d^3\tilde{\mathbf{c}}(q)}{(2\pi)^3}, \tag{4.4}$$

and Fourier transform Eq. (4.3), we obtain

$$G(\mathbf{k}, N) = \frac{\displaystyle\prod_q \int \frac{d^3\tilde{\mathbf{c}}(q)}{(2\pi)^3}\, e^{i\mathbf{k}\cdot\tilde{\mathbf{c}}(q)[e^{iqN}-1] - \frac{1}{2}q^2|\tilde{\mathbf{c}}(q)|^2}}{\displaystyle\prod_q \int \frac{d^3\tilde{\mathbf{c}}(q)}{(2\pi)^3}\, e^{-\frac{1}{2}q^2|\tilde{\mathbf{c}}(q)|^2}}. \tag{4.5}$$

Both the numerator and the denominator contain products of Gaussian integrals which can be calculated by completing the square. The result is

$$G(\mathbf{k}, N) = \exp\Big(- \mathbf{k}^2 \int_{-\infty}^{+\infty} \frac{dq}{2\pi} \frac{1 - \cos qN}{q^2} \Big) = e^{-\mathbf{k^2} N/2}, \tag{4.6}$$

which is the Fourier transform of

$$G(\mathbf{r}, N) = \frac{e^{-\mathbf{r}^2/(2N)}}{(2\pi N)^{3/2}}. \tag{4.7}$$

4.2 Anchored polymer: The partition function

Now that we are familiar with the methodology of path integrals, we study a more complicated problem – anchored Gaussian chain in a half-space $z > 0$. In terms of path integrals, we immediately see that our task will not be as simple as before because the possible values of $c_z(\tau)$ must be positive. If we stay in real space and calculate the limit of a multidimensional integral, we see that the resulting integrals cannot be calculated analytically with the constraint $z > 0$. If we were to move to Fourier space, it is not even clear how to define the measure of integration. The way out of this complication is as follows. We allow the polymer to cross the boundary and introduce a strong repulsive interaction between the plane at $z = 0$ and the chain. Each time the polymer crosses or touches the plane, it is penalized by a certain amount of energy. The modified Hamiltonian is then

$$H = H_0 + H_1, \tag{4.8}$$

where

$$H_0 = \frac{1}{2}\int_0^N \dot{\mathbf{c}}^2 \, d\tau, \tag{4.9}$$

and

$$H_1 = g \int_0^N \delta[c_z(\tau) - c_z(0)] \, d\tau. \tag{4.10}$$

Thus, we expect that when the coupling constant $g > 0$ becomes infinitely large, the polymer will be entirely on one (either positive or negative) side of the plane. The partition function is then

$$\mathcal{Z}(g, N) = \int D[\mathbf{c}(\tau)] e^{-H_0[\mathbf{c}] - H_1[g,\mathbf{c}]}. \tag{4.11}$$

To evaluate $\mathcal{Z}(g, N)$, we expand the integrand in Eq. (4.11) in powers of g. Such an expansion could be problematic when g is large, the limit of primary interest. However, if we are able to calculate all terms in the expansion and to perform the summation, then this expansion is not an issue. The nth ($n = 1, 2, \ldots$) term of the expansion reads

$$\frac{(-g)^n}{n!}\int D[\mathbf{c}(\tau)]e^{-H_0[\mathbf{c}]}\prod_{l=1}^{n}\int \delta[c_z(\tau_l)-c_z(0)]\,d\tau_l. \tag{4.12}$$

We order the set $\{\tau_l\}$, Fourier transform the δ-functions, and rewrite Eq. (4.12) as

$$\Big(\frac{-g}{2\pi}\Big)^n\int D[\mathbf{c}(\tau)]\,e^{-H_0[\mathbf{c}]}\int_0^N d\tau_n\int_0^{\tau_n}d\tau_{n-1}\ldots\int_0^{\tau_2}d\tau_1\prod_{l=1}^{n}\int_{-\infty}^{+\infty}dk_l\,e^{-ik_l[c_z(\tau_l)-c_z(0)]}. \tag{4.13}$$

Henceforth, we focus on $c_z(\tau)$ and denote it $c(\tau)$ for simplicity. If we integrate it out by completing the square, we are left with

$$\Big(\frac{-g}{2\pi}\Big)^n\int_0^N d\tau_n\int_0^{\tau_n}d\tau_{n-1}\ldots\int_0^{\tau_2}d\tau_1\int_{-\infty}^{+\infty}dk_1\ldots dk_l\,\exp\Big[-\frac{1}{2}\sum_{l,m}k_l\mathbb{T}^{(n)}_{lm}k_m\Big]. \tag{4.14}$$

Here,

$$\mathbb{T}^{(n)}_{lm}=\tau_{\min[l,m]}=\begin{pmatrix}\tau_1 & \tau_1 & \tau_1 & \cdot & \cdot & \tau_1\\ \tau_1 & \tau_2 & \tau_2 & \cdot & \cdot & \tau_2\\ \tau_1 & \tau_2 & \tau_3 & \cdot & \cdot & \tau_3\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\ \tau_1 & \tau_2 & \tau_3 & \cdot & \cdot & \tau_n\end{pmatrix} \tag{4.15}$$

To perform multiple integration over k_i, we use the well known formula

$$\int_{-\infty}^{+\infty}\prod_j dk_j\,\exp\Big(-\frac{1}{2}\sum_{l,m}k_l\mathbb{T}^{(n)}_{lm}k_m\Big)=\sqrt{\frac{(2\pi)^n}{\det\mathbb{T}^{(\mathrm{n})}}}. \tag{4.16}$$

It is straightforward to show that

$$\det\mathbb{T}^{(\mathrm{n})}=\tau_1(\tau_2-\tau_1)(\tau_3-\tau_2)\ldots(\tau_\mathrm{n}-\tau_{\mathrm{n}-1}), \tag{4.17}$$

so that after integration over k_i, Eq. (4.12) reduces to

$$\Big(\frac{-g}{\sqrt{2\pi}}\Big)^n \int_0^N d\tau_n \int_0^{\tau_n} \frac{d\tau_{n-1}}{\sqrt{\tau_n - \tau_{n-1}}} \cdots \int_0^{\tau_3} \frac{d\tau_2}{\sqrt{\tau_3 - \tau_2}} \int_0^{\tau_2} \frac{d\tau_1}{\sqrt{\tau_1(\tau_2 - \tau_1)}}$$

$$= \Big(\frac{-g}{\sqrt{2\pi}}\Big)^n \int_0^N d\tau_n \tau_n^{n/2-1} \prod_{m=1}^{n-1} \int_0^1 x^{m/2-1}(1-x)^{-1/2} dx = \frac{\Big(-g\sqrt{N/2}\Big)^n}{\Gamma\Big(\dfrac{n}{2}+1\Big)}$$

$$\equiv \frac{(-\hat{g})^n}{\Gamma\Big(\dfrac{n}{2}+1\Big)}, \quad (4.18)$$

where $\hat{g} \equiv g\sqrt{N/2}$. Thus,

$$\mathcal{Z}(g, N) = \mathcal{Z}(\hat{g}) = \sum_{n=0}^{\infty} \frac{(-\hat{g})^n}{\Gamma\Big(\dfrac{n}{2}+1\Big)} = e^{\hat{g}^2}[1 - \Phi(\hat{g})], \quad (4.19)$$

where $\Phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$ is the error function. When $N \to \infty$, so does $\hat{g}$. We expand $\mathcal{Z}(\hat{g})$ for large $\hat{g}$ and obtain

$$\mathcal{Z}(\hat{g}) = \frac{1}{\sqrt{\pi}} \Big[\frac{1}{\hat{g}} - \frac{1}{2\hat{g}^3} + O(\hat{g}^{-5})\Big]. \quad (4.20)$$

Hence,

$$\gamma = 1 + \lim_{N\to\infty} \frac{\partial \ln \mathcal{Z}}{\partial \ln N} = \frac{1}{2}. \quad (4.21)$$

Several remarks should be made at this point. First, note that while calculating the general term of the expansion, we omitted the common factor

$$\mathcal{Z}_0(N) = \int D[\mathbf{c}(\tau)] e^{-H_0[\mathbf{c}]}, \quad (4.22)$$

which has the form ζ^N. Second, the coupling constant g plays here the role of the inverse cutoff length, which often occurs in field theory. Finally, we note that for any value of $g > 0$, we can find N large enough to make the non-dimensional coupling $\hat{g} = g\sqrt{N/2} \gg 1$, so that the number of accessible configurations scales as $N^{-1/2}$ relative to the unconstrained case. This important observation is a signature of universality: no matter how small g is, for a long enough polymer, the overall repulsion is infinitely strong!

4.3 Probability distribution

The Fourier transform of the (unnormalized) probability distribution function for the end-to-end distance of the random walk is given by

$$G(\mathbf{q},N)=\int D[\mathbf{c}(\tau)],\, e^{-i\mathbf{q}\cdot[\mathbf{c}(N)-\mathbf{c}(0)]}\, e^{-H_0[\mathbf{c}]-H_1[g,\mathbf{c}]}. \tag{4.23}$$

As before, we focus only on the z-dependent part of the probability distribution function $G(z,N)$. If we expand the path integral in powers of g and integrate out $c(\tau)$, we observe that the nth term of the expansion reads

$$\begin{aligned}&\Big(\frac{-g}{2\pi}\Big)^n\int_0^N d\tau_n\int_0^{\tau_n}d\tau_{n-1}\dots\int_0^{\tau_2}d\tau_1\int\prod_l dk_l\exp\big[-\frac{1}{2}(q^2N+\sum_l k_l\tau_l+\sum_{l,m}k_l\mathbb{T}^{(n)}_{lm}k_m)\big]\\ &=\Big(\frac{-g}{\sqrt{2\pi}}\Big)^n\int_0^N d\tau_n\int_0^{\tau_n}d\tau_{n-1}\dots\int_0^{\tau_2}\frac{d\tau_1}{\sqrt{\mathrm{det}\mathbb{T}^{(\mathrm{n})}}}\exp\big[-\frac{q^2}{2}(N-\sum_{l,m}\tau_l[\mathbb{T}^{(n)}]^{-1}_{lm}\tau_m)\big],\end{aligned} \tag{4.24}$$

where $\mathbb{T}^{(n)}$ is given by Eq. (4.17). Now,

$$[\mathbb{T}^{(n)}]^{-1}_{lm}=\begin{cases}-\dfrac{\delta_{l+1,m}}{\tau_{l+1}-\tau_l}-\dfrac{\delta_{l-1,m}}{\tau_l-\tau_{l-1}}+\delta_{m,l}\big(\dfrac{1}{\tau_{l+1}-\tau_l}+\dfrac{1}{\tau_l-\tau_{l-1}}\big) & (l\neq n,0)\\ -\dfrac{\delta_{2,m}}{\tau_2-\tau_1}-\delta_{1,l}\big(\dfrac{1}{\tau_2-\tau_1}+\dfrac{1}{\tau_1}\big) & (l=1)\\ -\dfrac{\delta_{n-1,m}}{\tau_n-\tau_{n-1}}+\dfrac{\delta_{m,n}}{\tau_n-\tau_{n-1}} & (l=n)\end{cases} \tag{4.25}$$

A straightforward calculation yields

$$\sum_{l,m}\tau_l[\mathbb{T}^{(n)}]^{-1}_{lm}\tau_m=\tau_n. \tag{4.26}$$

Equation (4.24) therefore reduces to

$$\Big(\frac{-g}{\sqrt{2\pi}}\Big)^n\int_0^N d\tau_n\int_0^{\tau_n} d\tau_{n-1}\ldots\int_0^{\tau_2}\frac{d\tau_1}{\sqrt{\det\mathbb{T}^{(\mathrm{n})}}}\exp\Big[-\frac{q^2}{2}(N-\tau_n)\Big]$$
$$=\frac{(-g)^n\,e^{-q^2N/2}}{2^{n/2}\Gamma(n/2)}\int_0^N\frac{d\tau}{\tau}\tau^{n/2}e^{q^2\tau/2}=\frac{(-\hat{g})^n\,e^{-q^2N/2}}{\Gamma(n/2)}\int_0^1\frac{ds}{s}s^{n/2}e^{(q^2N/2)s}. \qquad (4.27)$$

If we sum over n, we obtain

$$\tilde{G}(q,N)=e^{-q^2N/2}\Big[1+2\int_0^1\frac{ds}{s}e^{(q^2N/2)s^2}\,f(\hat{g}s)\Big], \qquad (4.28)$$

where

$$f(x)=-\frac{x}{\sqrt{\pi}}+x^2e^{x^2}(1-\Phi(x)). \qquad (4.29)$$

Thus,

$$G(z,N)=\frac{e^{-z^2/2N}}{\sqrt{2\pi N}}F(\hat{g},\hat{z}), \qquad (4.30)$$

where $\hat{z}\equiv z/\sqrt{N}$, and

$$F\big[\hat{g},\hat{z}\big]=1+2\int_0^1\frac{ds}{s}\frac{e^{-\frac{1}{2}\hat{z}^2s^2/(1-s^2)}}{\sqrt{1-s^2}}\,f(\hat{g}s) \qquad (4.31)$$

is the scaling function. To calculate $F(\hat{g},\hat{z})$, we rewrite Eq. (4.31) as

$$F(\hat{g},\hat{z})=A(\hat{g})-2\int_0^1\frac{ds}{s}\frac{1-e^{-\frac{1}{2}\hat{z}^2s^2/(1-s^2)}}{\sqrt{1-s^2}}\,f(\hat{g}s), \qquad (4.32)$$

where

$$A(\hat{g})=1+2\int_0^1\frac{ds}{s\sqrt{1-s^2}}\,f(\hat{g}s). \qquad (4.33)$$

The integral in Eq. (4.32) cannot be calculated analytically in general, that is, for arbitrary $\hat{g}$ and $\hat{z}$. However, because we are interested in the limit $\hat{g}\gg 1$, we can easily calculate the leading term. We note that the integrand is essentially nonzero only for values of s larger than some value $s_0(\hat{z})$. However small $s_0(\hat{z})$ is, we can always take $\hat{g}$ large enough to make $\hat{g}s_0(\hat{z}) \gg 1$. Thus, we can take

$$f(\hat{g}s) \simeq -\frac{1}{2\sqrt{\pi}\hat{g}s}, \tag{4.34}$$

so that

$$F(\hat{g},\hat{z}) \simeq A(\hat{g}) + \frac{1}{\sqrt{\pi}\hat{g}} \int_0^1 \frac{ds}{s^2} \frac{1 - e^{-\frac{1}{2}\hat{z}^2 s^2/(1-s^2)}}{\sqrt{1-s^2}}. \tag{4.35}$$

By using the substitution

$$u = \frac{|\hat{z}|s}{\sqrt{2(1-s^2)}}, \tag{4.36}$$

we finally obtain

$$F(\hat{g},\hat{z}) \simeq A(\hat{g}) + \frac{|\hat{z}|}{\sqrt{2\pi}\hat{g}} \int_0^\infty \frac{du}{u^2}(1 - e^{-u^2}) = A(\hat{g}) + \frac{|\hat{z}|}{\sqrt{2}\hat{g}}. \tag{4.37}$$

For large values of $\hat{g}$, we have

$$A(\hat{g}) = \frac{1}{2\sqrt{\pi}\hat{g}^2} + O(\hat{g}^{-4}). \tag{4.38}$$

Thus, when $\hat{g} \to \infty$, the scaling function is linear in $\hat{z}$ and the normalized probability distribution function has the form

$$G(z,N) = \frac{|z|}{2N} e^{-z^2/2N}. \tag{4.39}$$

Apart from a factor of 2, this function is identical to the one obtained by solving the diffusion equation (see Appendix B). This factor appears because now the chain can be either in the $z < 0$ or in the $z > 0$ half-space. Figure 4-1 shows the numerically computed scaling function $F(\hat{g},\hat{z})$ for different values of $\hat{g}$. As expected, the larger $\hat{g}$, the closer is $F(\hat{g},\hat{z})$ to the linear dependence.

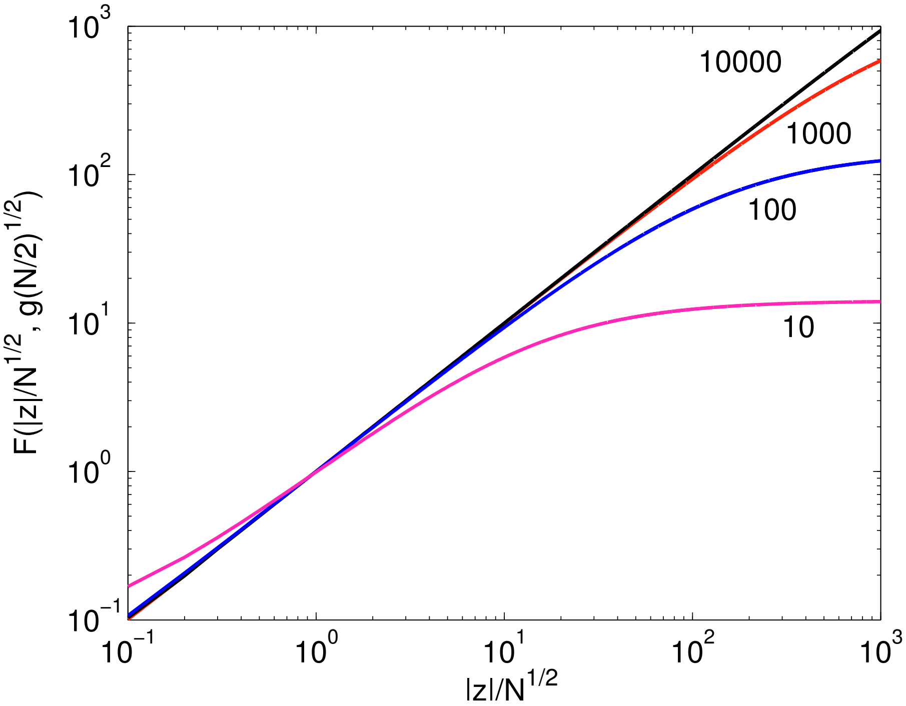


Figure 4-1: The (normalized) scaling function $F(\hat{g}, \hat{z})$. Different curves are labeled by corresponding values of $\hat{g}$. The larger $\hat{g}$, the closer is the scaling function to a linear dependence.

Chapter 5

Apex exponents for polymer–probe interactions

5.1 Introduction

There has been remarkable progress in recent years in nanoprobing and single–molecule techniques. These developments have had a direct impact on biopolymer research, producing a wealth of beautiful results on DNA dynamics [85], molecular motors [86], and protein/RNA folding [87, 88]. Today it is possible to measure statistical properties of a single macromolecule rather than deducing them from experiments with solutions of many polymers. This naturally leads to questions regarding the theoretical limitations of these techniques, such as the effects of microscopic probes on the measured properties of the polymer. Consider, for instance, a polymer attached to the apex of a cone–shaped probe (e.g. a micropipette or the tip of an atomic force microscope [89, 90]). What is the configurational entropy for this system? Suppose that this probe is a microscopic needle with a hole at the end. How hard is it to thread a polymer through the needle's eye?

As we already mentioned, the number of configurations $\mathfrak{N}$ of a polymer of length N or, equivalently, of an N–step self–avoiding walk (SAW), behaves as [91]

$$\mathfrak{N} = \text{const} \times z^N N^{\gamma-1}. \tag{5.1}$$

The "effective coordination number" z, depends on microscopic details, while the exponent γ is "universal." As we have seen in the previous chapter, γ does depend on geometric constraints which influence the polymer at all length scales. In particular, there are a number of results demonstrating the variations of γ for polymers confined by wedges in two and three dimensions [92, 93, 94, 95, 96]. SAW anchored at the origin and confined to a solid wedge (in 3D) or a planar wedge (in 2D) has an angle–dependent exponent γ that diverges as the wedge angle α vanishes. A limiting case which has been extensively studied, both analytically [97, 98] and numerically [96, 95], is a SAW anchored to an impenetrable surface, for which $\gamma \equiv \gamma_s = 0.70 \pm 0.02$ [96].

To model the polymer–probe system, we consider a SAW attached to the apex (tip) of an impenetrable obstacle (needle). To avoid introduction of an external length scale, we focus on obstacles of scale–invariant shape, such as a planar slice (sector) of angle α (Fig. 5-1a), or a conical needle of apex semi-angle β (Fig. 5-1b). While both geometries are natural extensions of the 2D wedge, they are clearly different in three dimensions (and also distinct from the 3D wedge, which consists of two planes intersecting at a line). The former excludes the polymer from the volume of a cone, while the latter prevents it from crossing the surface of a slice. Nonetheless, the resulting phenomenology is rather similar. Indeed, one of the technical innovations of this chapter is the demonstration that many such geometries can be treated in the same manner by an $\epsilon = 4-d$ expansion focusing on the interaction with a 2D surface. The ϵ–expansion, as well as numerical simulations in 3D, shows that the exponent $\gamma \equiv \gamma_1$ varies continuously with the apex opening angles in Fig. 5-1. Continuously varying exponents are rather uncommon in critical phenomena. In the present case they arise from the interaction of two self-similar entities, the polymer and the probe.

Another variant of this problem occurs when a polymer is attached to the apex at its *midpoint*. This case is described by Eq. (5.1) with exponent $\gamma \equiv \gamma_2$. More generally, let us denote by $\mathfrak{N}_2(N, N_1)$, the number of accessible configurations for a polymer attached to the apex at an arbitrary monomer, dividing it in two segments of lengths N_1 and $N_2 = N - N_1$. If we allow the two segments to exchange monomers

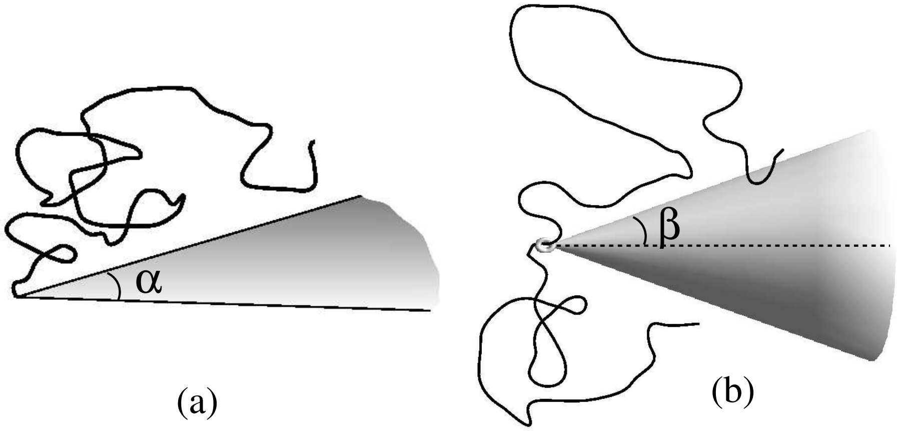


Figure 5-1: Configurations of a polymer near an obstacle: (a) attached to the apex of a planar sector of angle α; (b) threaded through the eye of a cone with apex semi–angle β.

with each other (which can be done by replacing a rigid attachment with a slip–ring as depicted in the Fig. 5-1b), then the equilibrium configurations will be distributed with a weight proportional to $\mathfrak{N}_2(N, N_1)$. A natural interpolation formula as a function of α (or β), supported by the ϵ-expansion at first order, is

$$\mathfrak{N}_2(N, N_1) \propto N^{c(\alpha)}[N_1(N - N_1)]^{c_1(\alpha)}. \tag{5.2}$$

To get a feeling for this scaling relation, let us look at some limits: When the probe is absent, we recover Eq. (5.1) and $c(0) = \gamma_0 - 1$, where $\gamma_0 \simeq 1.158$ describes the geometrically unconstrained SAW. If the obstacle is present but the two segments do not interact with each other, then $c = 0$ and $c_1 = \gamma_1 - 1$. By fitting to the limits of $N_1 \to 0$ and $N_1 \sim N_2$, we find $c_1 = \gamma_2 - \gamma_1$ and $c = 2\gamma_1 - \gamma_2 - 1$. Below, we estimate the exponents in Eq.(5.2) both analytically and numerically. For now, assuming Eq. (5.2) holds, we see that if $c_1 < 0$, the maximum number of configurations is realized when either N_1 or N_2 equals N. This brings us to one of our main findings: No matter how small the apex angle, we find $c_1 < 0$, i.e. the most likely states have $N_1 \simeq N$ or $N_2 \simeq N$, with an *entropy barrier* separating the two. Threading a needle is hard!

5.2 Analytic calculations

To treat the problem analytically, we start with the Edwards [99] model of a self-avoiding polymer, and add an interaction with the obstacle. In this formulation, configurations of the polymer are described by $\mathbf{r}(\tau) \in \Re^d$, where τ measures the position along the chain, and are weighted according to the energy [1]

$$\begin{aligned} \mathcal{H} &= \frac{1}{2}\int_0^N \dot{\mathbf{r}}^2 \, d\tau + \frac{v_0}{2}\int_0^N d\tau \int_0^N d\tau' \delta[\mathbf{r}(\tau) - \mathbf{r}(\tau')] \\ &+ g_0 \int_{\mathcal{M}} d^2\mathbf{R} \int_0^N d\tau \delta[\mathbf{r}(\tau) - \mathbf{R}]. \end{aligned} \tag{5.3}$$

The self-avoiding interaction is replaced by a "soft" repulsion of strength v_0. In the same spirit, the impenetrable obstacle is replaced with a soft repulsion of magnitude g_0. The key observation is that in 3D the polymer can only sense the exterior of an impenetrable obstacle, and will not care if its interior is hollow. In generalizing to d-dimensions, we keep the dimensions of the now softened exterior manifold (indicated by $\mathbf{R} \in \mathcal{M}$) as two. The advantage of this choice is that both g_0 and v_0 have the same bare dimensions, and in a perturbative scheme, they simultaneously become relevant in $d \leq 4$. We then analyze the model using a renormalization group (RG) scheme [97, 98], which is a modification of the conformation space RG [100, 82]. The scaling exponents are calculated using dimensional regularization in $d = 4 - \epsilon$ dimensions to order $O(\epsilon)$.

It is customary to define non-dimensionalized coupling constants $\tilde{v}_0 = v_0 L^\epsilon$, $\tilde{g}_0 = g_0 L^\epsilon$ where L is some length scale. Bare coupling constants are related to the renormalized ones by

$$\begin{aligned} \tilde{v}_0 &= Z_v(v,g)v = (1 + Av + Dg + ...)v, \\ \tilde{g}_0 &= Z_g(v,g)g = (1 + Cv + Bg + ...)g. \end{aligned} \tag{5.4}$$

Inverting these equations, we obtain series expansions $v(v_0, g_0)$ and $g(v_0, g_0)$ that can

[1] To simplify notation, the monomer size is absorbed into a redefinition of N, giving it dimensions of [length]^2.

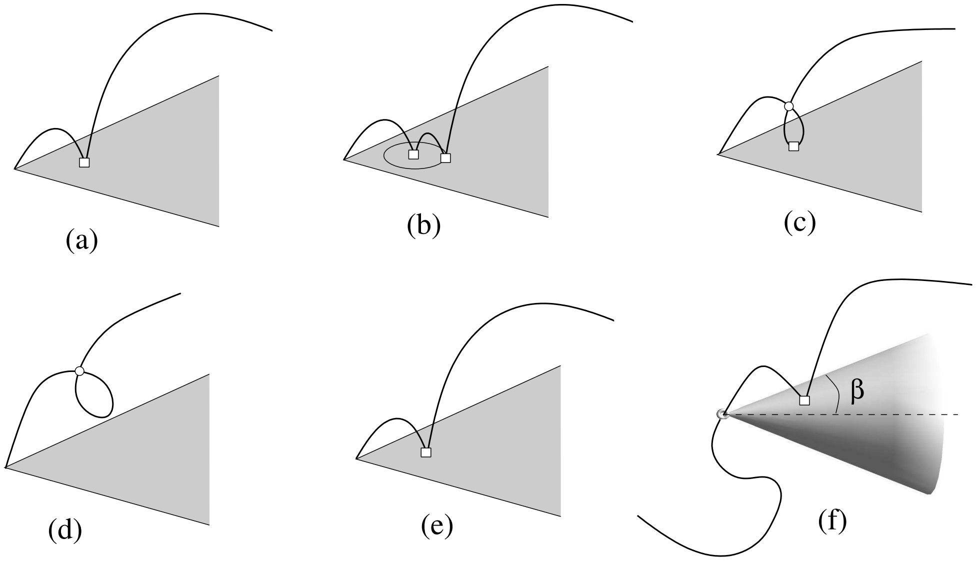


Figure 5-2: Diagrams contributing to renormalization of g to second order (a–c); to $\mathcal{Z}$ in first order (d,e) at the apex of a slice; and to $\mathcal{Z}_2$ in first order (f) at the eye of a conic needle.

be viewed as perturbative expansions in the bare coupling constants

$$\begin{aligned} v &= (1 - A\tilde{v}_0 - D\tilde{g}_0 + ...)\tilde{v}_0, \\ g &= (1 - C\tilde{v}_0 - B\tilde{g}_0 + ...)\tilde{g}_0. \end{aligned} \tag{5.5}$$

To obtain the leading corrections to critical indices, we perturbatively calculate the coefficients of the expansion in $d = 4 - \epsilon$ dimensions. First, we note that self-interaction of the polymer is not influenced by the presence of the sector. Therefore, in the expansion (5.5) we can put $D = 0$. Also, the value of A is well known [82]; as $\epsilon \to 0$, A has a pole

$$A = \frac{2}{\pi^2 \epsilon} + O(1). \tag{5.6}$$

Diagrams contributing to the renormalization of g_0 are shown in Fig. 5-2a–c and involve both interaction of polymer with the plane and polymer self-interaction. To estimate the contribution of the former, we first consider the limiting case $\alpha = 2\pi$, i.e. a complete plane [98, 101]. In this case, first-order correction to g_0 is

$$g_0 \int_0^N d\tau \int_{\mathcal{M}} d^2\mathbf{r}\ G_d(\mathbf{r}, \tau), \tag{5.7}$$

where

$$G_d(\mathbf{r},\tau) = \frac{1}{(2\pi\tau)^{d/2}} e^{-r^2/2\tau} \tag{5.8}$$

is a d-dimensional Gaussian. Note, that since $\mathbf{r}$ is a two-dimensional vector, $G_d(\mathbf{r},\tau)$ can be rewritten as

$$G_d(\mathbf{r},\tau) = \frac{1}{(2\pi\tau)^{1-\epsilon/2}} \cdot \frac{1}{2\pi\tau} e^{-r^2/2\tau}. \tag{5.9}$$

Plugging this expression into (5.7), we obtain

$$g_0 \int_0^N d\tau \int_{\mathcal{M}} d^2\mathbf{r}\ G_d(\mathbf{r},\tau) = \frac{g_0 L^\epsilon}{2\pi} \frac{2}{\epsilon} \left(\frac{2\pi N}{L^2}\right)^{\epsilon/2} = \frac{\tilde{g}_0}{\pi\epsilon} + O(1). \tag{5.10}$$

A correction to g_0 due to the polymer self-interaction is shown in Fig. 5-2c. Its value is

$$\begin{aligned} v_0 \int_0^N d\tau \int_0^N d\tau' \int d^d\mathbf{r}\ G_d(\mathbf{r},\tau) G_d(-\mathbf{r},\tau') &= v_0 \int_0^N d\tau \int_0^N d\tau'\ G_d(0,\tau+\tau') \\ &= \frac{\tilde{v}_0}{2\pi^2\epsilon} + O(1). \end{aligned} \tag{5.11}$$

The β-functions describing the RG flow for this problem are calculated straightforwardly yielding

$$\beta_1(v,g) \equiv L\frac{dv}{dL} = \epsilon(v - Av^2) + O(\epsilon^2) \tag{5.12}$$

$$\beta_2(v,g) \equiv L\frac{dg}{dL} = \epsilon(g - Bg^2 - Cgv) + O(\epsilon^2). \tag{5.13}$$

The nontrivial RG fixed point is thus

$$(v^*, g^*) = \left(\frac{1}{A}, \frac{1}{B}\left[1 - \frac{C}{A}\right]\right) = \left(\frac{\pi^2\epsilon}{2}, \frac{3\pi\epsilon}{4}\right). \tag{5.14}$$

For arbitrary sector angle α, the only coefficient that can possibly change is B, i.e. the second-order interaction with the plane. However, the leading singularity in ϵ comes from small distances. Therefore, as long as it is possible to draw a circle of a finite radius around any point of polymer-sector contact (see Fig. 5-2b), the leading singularity remains unchanged. This also shows why the interaction with the sector in the degenerate $\alpha = 0$ case is irrelevant.

The fixed point of the RG depends only on the dimension of the manifold and not on its shape; this fact was mentioned before [101, 102]. However, the number of accessible configurations does depend on certain details of the manifold, as described below.

Consider first-order corrections to the partition function $\mathcal{Z}$ due to the self–interaction (Fig. 5-2d) and the interaction with the slice (Fig. 5-2e). Combining them and adding relevant counterterms to eliminate poles in ϵ, we obtain

$$\mathcal{Z} = 1 + \frac{1}{4\pi^2}(v^* - \alpha g^*) \ln\left(\frac{2\pi N}{L^2}\right) + O(\epsilon^2). \tag{5.15}$$

Comparing this with $N^{\gamma_1(\alpha)-1} = 1 + (\gamma_1(\alpha) - 1) \ln N + \cdots$, and substituting the fixed point values (v^*, g^*), we find

$$\gamma_1(\alpha) = 1 + \frac{\epsilon}{8}\left(1 - \frac{3\alpha}{2\pi}\right) + O(\epsilon^2). \tag{5.16}$$

The above treatment is easily generalized to a polymer attached by its midpoint. For the number of configurations, we observe that the contribution from the interaction with the obstacle is doubled. Interaction between the two halves of the polymer, however, makes no separate correction and is already included. (Note that if we ignore the obstacle and consider self-interactions only, we get a "degenerate" star polymer with two branches that is equivalent to a linear polymer.) Thus, for the calculation of γ_2 at order of ϵ, we can add the separate contributions from self-avoidance and avoidance of the obstacle; cross-terms can only occur at higher orders. This enables us to identify the scaling exponent

$$\gamma_2(\alpha) = 1 + \frac{\epsilon}{8}\left(1 - \frac{3\alpha}{\pi}\right) + O(\epsilon^2). \tag{5.17}$$

Repeating this argument for the slip–ring geometry, we find

$$\mathfrak{N}_2 \propto N^{\epsilon/8}[N_1(N - N_1)]^{-3\alpha\epsilon/(16\pi)}, \tag{5.18}$$

which confirms the Ansatz in Eq. (5.2) to first order in ϵ.

It is straightforward to extend the above formalism to obstacles of different shapes, such as the conical manifold with apex angle β (Fig. 5-1b). In counting the number of configurations we obtain a result similar to Eq. (5.15), with α replaced by $2\pi \sin\beta$. Since the fixed point location is the same as before, this substitution in Eqs. (5.16-5.17) gives

$$\begin{aligned} \gamma_1^{\text{cone}}(\beta) &= 1 + \frac{\epsilon}{8}(1 - 3\sin\beta) + O(\epsilon^2), && (5.19) \\ \mathfrak{N}_2^{\text{cone}}(\beta) &\propto N^{\epsilon/8}[N_1(N - N_1)]^{-(3/8)\epsilon\sin\beta}. && (5.20) \end{aligned}$$

The difference between the two geometries is thus merely quantitative.

Thus, our approach provides a simple way of calculating critical exponents for geometries intractable by other methods that explicitly exclude entire d-dimensional regions [92, 93]. However, it does break down in certain limits. For instance, in the case of a cone, the polymer is free to occupy either side of the hollow cone– the partition sum is dominated by the arrangement with the largest number of configurations. Hence the result for $\gamma_1(\beta)$ is valid only for $\beta \leq \pi/2$. The restriction for $\gamma_2(\beta)$ is even more severe. For values of β larger than some critical angle $\beta_c < \pi/2$, the self-avoidance will cause the two halves of the polymer to be on the opposite sides of the conical surface thereby invalidating the calculation. Certain limitations exist for the planar sector geometry as well. For example, in 3D we must have $\gamma_2(2\pi) = 2\gamma_s - 1$. This equality does not hold in the ϵ–expansion. The reason is that in 3D, a complete plane prevents two polymers on its opposite sides from interacting with each other, whereas in 4D it does not. In short, the method described above, despite its appealing simplicity, is not omnipotent and must be used with some caution.

5.3 The numerics

To check the validity of the analytic approach, we present here the values of scaling exponents $\gamma_{1,2}$ calculated from numerical simulations, performed by Dr. Roya Zandi

(UCLA) and Prof. Yacov Kantor (Tel-Aviv University).

5.3.1 Entropic competition

The earlier discussion of "threading a needle" illustrates the essence of the method of "entropic competition" [103, 104], which we employ to numerically estimate the exponents $\gamma_i(\alpha)$ in 3D. We sample the ensemble of different configurations of *two polymer segments* which can exchange monomers and thus "compete entropically." To calculate $\gamma_1(\alpha)$, we prevent the two segments from interacting with each other. The number of configurations is then

$$\mathfrak{N}_1 \propto [N_1(N-N_1)]^{\gamma_1(\alpha)-1}, \tag{5.21}$$

so that the resulting histogram for N_1 allows us to calculate the exponent $\gamma_1(\alpha)$. Possible Monte Carlo (MC) moves include attempts to remove one monomer from the *free* end of a randomly chosen polymer segment and add it to the *free* end of the other segment; both segments also undergo random configuration changes via pivoting [105]. Figure 5-3 illustrates the dramatic effect of the angle α on $p(N_1)$, the probability distribution function (PDF) for the segment length N_1. For small α, the distribution is peaked at the center while for α bigger than a critical value α_c, the maximum of the PDF moves to the sides. The numerical data from entropic competition suggest $\alpha_c \approx 5\pi/8$, which is not too far from the first order ϵ–expansion result of $\alpha_c = 2\pi/3$ in Eq. (5.16).

For the purpose of calculating $\gamma_2(\alpha)$, we include interactions between the segments. Open symbols in Fig. 5-4 show variations of the exponents $\gamma_{1,2}(\alpha)$ fitting histograms from entropic competition, such as in Fig. 5-3, to power-laws as in Eqs. (5.21) and (5.2).

5.3.2 Dimerization

It is instructive to compare the results of entropic competition with those of a more established procedure, such as dimerization [106, 107]. The latter is quite an efficient

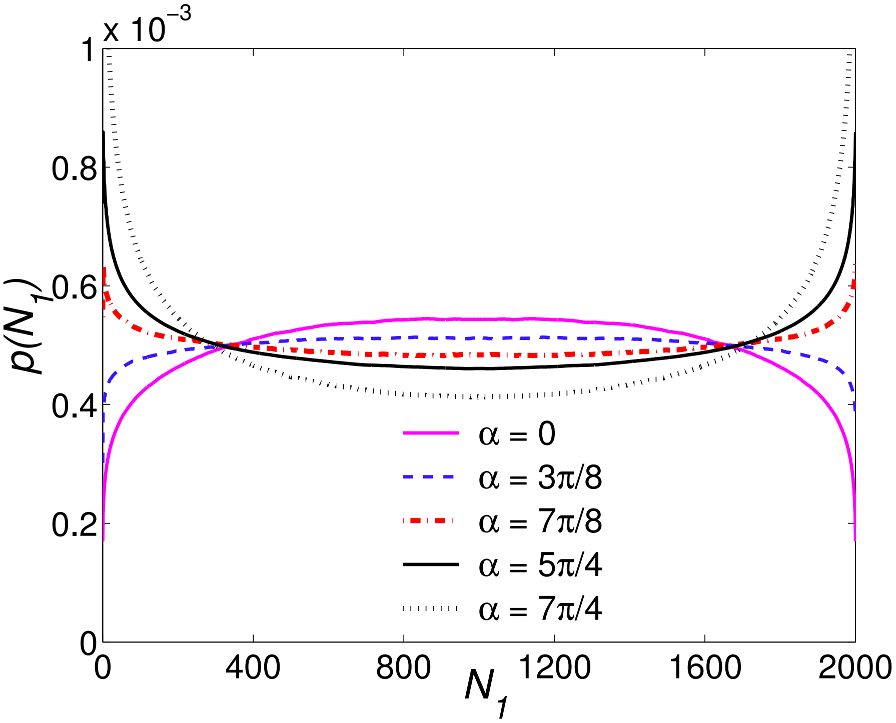


Figure 5-3: The probability distributions $p(N_1)$ for two non-interacting segments of lengths N_1 and $N - N_1$ attached to the apex of a planar slice for diffrent values of angle α. The curves are the result of 10^9 MC steps for $N = 2000$.

method [105], in which an N-step SAW is created by generating two $(N/2)$-step SAWs and attempting to concatenate them. We generated SAWs for $N = 16,\ 32, \cdots, 2048$, and by attempting to attach them to the end point of an appropriate sector, measured a success probability p_N. Let us indicate the number of SAWs not attached to the sector by $A_0 z^N N^{\gamma_0 - 1}$, and those attached to the sector either (1) by their ends, or (2) by their mid-point as $A_i z^N N^{\gamma_i(\alpha)-1}$ ($i = 1, 2$ corresponds to the notation introduced earlier). Then, the ratio between the number of configurations, $p_N \equiv (A_i/A_0) N^{\gamma_i(\alpha)-\gamma_0}$, represents the probability to attach an N–step polymer to a sector with angle α. Fitting a power law to this ratio thus provides a means of estimating the exponent difference

$$\Delta\gamma_i \equiv \gamma_0 - \gamma_i = \ln(p_N/p_{2N})/\ln 2. \tag{5.22}$$

The functional form represented by Eq. 5.1 is valid only in the limit of $N \to \infty$. For large finite N, we expect the expressions in Eqs. 5.1 and 5.22 to be modified by a multiplicative factor of the form $1 + c/\sqrt{N}$, where we assume that the correction to scaling exponent [108] is $\frac{1}{2}$. Consequently, the value of $\Delta\gamma_i$ will also depend on N. By examining the successive estimates as a function of $1/\sqrt{N}$ and extrapolation of the results to $1/\sqrt{N} = 0$, one can estimate the asymptotic value of the exponent. Using the dimerization method we generated $M = 10^6$ SAWs. We were able to obtain reasonable estimates of the exponent for all values of α, as shown in Fig. 5-4 (full symbols).

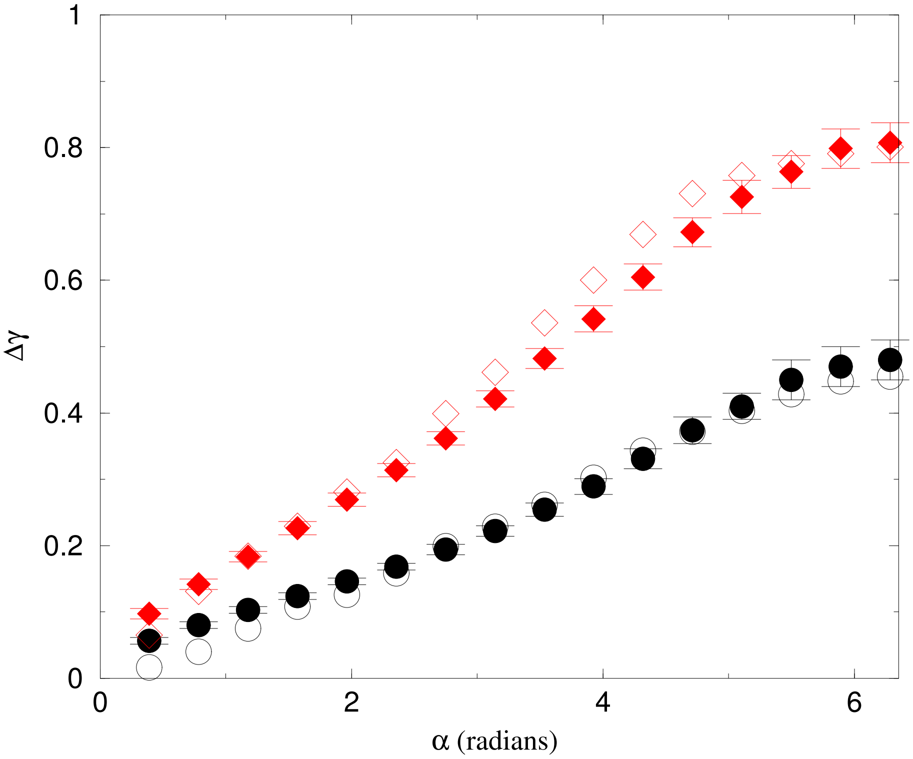


Figure 5-4: Extrapolated values of the exponents $\Delta\gamma_1 = \gamma_0 - \gamma_1$ (circles) and $\Delta\gamma_2 = \gamma_0 - \gamma_2$ (diamonds) as a function of sector angle α from “entropic competition” (open symbols), and dimerization (full symbols). Error bars represent statistical uncertainties of individual estimates of the exponents, as well as the uncertainty in the extrapolation $N \to \infty$.

The two numerical approaches are in very good agreement; error bars for “entropic competition” results are even smaller than those for dimerization. For $\alpha = 0$, our results deviate from zero beyond the statistical error range. We believe this deviation to be a finite size effect, due to discreteness of the lattice. As a check, we estimated $\Delta\gamma_{1,2}$ when the obstacle consists of the positive x-axis. While asymptotically such a situation corresponds to $\alpha = 0$ and should lead to $\Delta\gamma_i = 0$, we obtained $\Delta\gamma_1 = 0.02$ and $\Delta\gamma_2 = 0.05$. For $\alpha = 2\pi$, we expect to have $\Delta\gamma_1 = \gamma_0 - \gamma_s \approx 0.46$, and $\Delta\gamma_2 = \gamma_0 - 2\gamma_s + 1 \approx 0.76$; our results are quite close to these estimates.

5.4 Conclusions

We consider configurations of a polymer attached to the apex of a self-similar probe (at least on the scale of the polymer size). The geometric constraints imposed by the impenetrable probe lead to exponents γ which vary continuously with the apex angle. Two such exponents are associated with attachment of the polymer by one end or by a mid-point. Together, they determine if a mobile attachment point is likely to be in the middle or slide to one side. These apex exponents are obtained analytically by an $\epsilon = 4 - d$ expansion and through independent numerical schemes in $d = 3$. The ϵ-expansion takes advantage of the marginality of interactions of a polymer with a two-dimensional manifold in four dimensions, and can be applied to a variety of shapes. The numerical method of "entropic competition" is shown to be a powerful tool in this context, comparable to or better than the more standard dimerization approach. The numerical and analytical results agree up to 10-15% and indicate the presence of an entropic barrier that favors attachment of the polymer to the apex at its end. It would be interesting to see if these predictions can be probed by single molecule experiments.

Chapter 6

Concluding remarks and future directions

6.1 Protein-DNA interactions

6.1.1 Relevance to a bacterial cell

Needless to say, the model described in Chapter 1 (as any quantitative model) is a gross simplification of protein-DNA recognition *in vivo*. Despite this simplification, the proposed mechanism can be generalized to describe *in vivo* binding.

Simultaneous search by several proteins

If several TFs are searching for their sites on the DNA, the total search time is given by Eq. (1.36) and is obviously shorter than the time for a single TF. For example, if 100 copies of a TF are searching in parallel for the cognate site, then assuming $k_{on}^{\text{cytoplasm}} \approx 10^8\text{M}^{-1}\text{s}^{-1}$ and a cell of 1 μm^3 volume, we obtain the search time of $t_s \approx 0.1\text{sec}$. Increasing the number of TF molecules can further decrease the search time, but can have harmful effects by causing molecular crowding in the cell. Note, however, that increasing the number of TF molecules to $100 - 1000$ per cell cannot resolve the speed-stability paradox.

“Funnels,” local organization of sites

In both bacterial and eukaryotic genomes, sites that tend to cluster together have been observed. One may suggest that such clustering or other local arrangement of sites can create a “funnel” in the binding energy landscape, leading to a more rapid binding of cognate sites. Our model suggests that even if such “funnels” do exist, they would not significantly speed up the search process. The proposed search mechanism involves $\sim M/\bar{n}_{\text{opt}} \sim 10^4$ rounds of 1D/3D diffusion. So a TF spends nearly all the search time far from the cognate site. Only the last round (out of 10^4) will be sped up by the “funnel,” which will not lead to a significant decrease of the search time.

Local organization of sites and other sequence-dependent properties of the DNA structure (flexibility of AT-rich regions, DNA curvature on poly-A tracks, etc.) may influence the preferred localization of TFs and lead to faster asociation and dissociation rates and fast equilibration on neighboring sites (see [109] for details).

Protein hopping: intersegment transfer

Our model assumes that rounds of 1D diffusion are separated by periods of 3D diffusion. Intersegment transfer is another mechanism that can separates rounds of 1D diffusion. If two segments of DNA come close to each other, a TF sliding along one segment can “hop” to another. The benefit of this mechanism is that it significantly shortens the transfer time τ_{3d}. Several pieces of experimental evidence suggest that tetrameric *LacI*, which has two DNA-binding sites, travels along DNA through 1D diffusion and intersegment transfer.

We did not consider this mechanism because of the following two considerations. First, it is unclear whether TFs that have only one binding site can perform intersegment transfer. Second, for this mechanism to work, distant segments of DNA need to come close to each other. While DNA packed into a cell or nuclear volume crosses itself every $\sim$ 500bp, DNA in solution (at *in vitro* concentrations) is unlikely to have any such self-crossings. Hence intersegment transfer cannot explain “faster than diffusion” binding rates observed *in vitro*. However, this mechanism may play a role *in* *vivo*, especially for proteins that have multiple DNA-binding sites.

Nonspecific binding energy

As we have shown above, the nonspecific binding energy E_{ns} controls the balance between sliding and 3D diffusion. By checking the optimality condition $\tau_{1d} = \tau_{3d}$, one can see whether a given TF was optimized for fast target location. It is known that bacterial transcription factors exhibit quite a wide range of nonspecific DNA affinities [110, 111], whereas Eq. (1.39) has a very general character. Therefore, for each specific transcription factor or other DNA-binding protein, there must be a set of evolutionary driving forces and factors that determine E_{ns}, possibly to optimize a certain function.

It would be interesting (and relatively simple) to check, for instance, if there is a correlation between the nonspecific binding energy and the number of binding sites for a given protein. In bacteria, some proteins (like *LacI*) have only one or just a few binding sites, whereas others, highly *pleiotropic* TFs (like $PurR$ or Crp) possess tens or even hundreds of cognate sequences. The exact number of copies per cell for each TF generally unknown, but it typically ranges from about 10 for the former to 100–1000 for the latter. Highly pleioptropic TFs are usually versatile regulators, providing universal repression or activation "services" throughout the bacterial DNA [1]. The large number of protein copies may compensate for slower 1D or 3D diffusion or for the imbalance between the two.

On the other hand, if nonspecific binding is anomalously strong, the TF will dissociate from the DNA only infrequently, and thus it will be present in the vicinity of the specific site for minutes. This may be the case for *LacI*; the fact that the gene that encodes for *LacI* is situated (on the DNA) close to *Lac* operon may also be of functional significance.

Cytoplasm inhomogeneity

In developing our model, we tacitly assumed that the cell cytoplasm is homogeneous, so that protein motion inside the cytoplasm is adequately described by normal 3D diffusion. However, in reality, the cell cytoplasm is crowded with enzymes, structural proteins, nucleic acids, ribosomes, etc. For example, the total density of protein and RNA inside a bacterial cell is about 300-400 g/L [112], whereas a typical biochemical experiment *in vitro* deals with total macromolecular densities of 1-10 g/L. Macromolecules occupy about 30% of the cell volume, and the mean distance between enzymes is of the order of the diameter of a typical tetrameric protein [113, 114]. Under these conditions, it is only natural to question the relevance of the adopted picture.

Recently, there were several attempts to assess the effects of macromolecular crowding in the cytoplasm on diffusion-limited reactions. The most obvious consequence is a considerable reduction of the diffusion coefficient, as reported by Lipkow *et al.* [115]. Furthermore, recent observations by Golding and Cox [116, 117] suggest a possibility of a change in the 3D diffusion law. Namely, in a crowded cytoplasm, a protein moves subdiffusively, i.e. its RMS displacement R scales with the diffusion time t as

$$R \sim t^{\alpha}, \qquad \alpha < 1/2. \tag{6.1}$$

This change in the diffusion law is usually a signature of a the change in the fractal dimension or connectivity of the underlying matrix [30]. If this is a real effect, it would be useful to study its implications for protein-DNA interaction.

It might appear counterintuitive, but macromolecular crowding may actually expedite protein-DNA association kinetics by sequestering the irrelevant part of the cell volume. Also, effective microcompartmentation of the cell may help in keeping DNA-binding proteins close at hand so that, upon receiving a signal, they can find their sites without exploring the entire cell. Similar mechanisms are now widely believed to be responsible for “channeling” and kinetics expedition in metabolic pathways [113].

6.1.2 Implications for eukaryotes

The basic assumptions of the developed picture (naked DNA, regulation by single TF molecules, etc.) make it directly applicable to bacteria only. However, the developed framework can provide useful guidelines for analyzing transcription regulation in higher organisms as well.

The effect of chromatin

Above we assumed that a TF is free to slide along the DNA. *In vivo* the picture is complicated by other proteins and protein complexes (nucleosomes, polymerases, etc.) bound to DNA, preventing a TF from sliding freely along DNA. What are the effects of such molecular crowding on the search time?

Our model suggests that molecular crowding on DNA can have little effect on the search time if certain conditions are satisfied. Obviously, the the cognate site should not be blocked by other DNA-bound molecules or nucleosomes. DNA-bound molecules can interfere with the search process by shortening regions of DNA scanned on each round of 1D diffusion. If, however, the distance between DNA-bound molecules or nucleosomes in the vicinity of the cognate site is greater than $\bar{n}_{\text{opt}} \sim 300 - 500$ bp (Eq. (1.32) and [23]), then obstacles on the DNA do not shorten the rounds of 1D diffusion and, hence, do not slow down the search process. Our analysis also suggests that sequestering of part of genomic DNA by nucleosomes can even speed up the search process by decreasing the effective genome size.

Formation of regulatory complexes

In prokaryotes, the regulation of a gene or operon is usually accomplished by a single repressor and/or a single activator. Also, bacterial TFs form a large number of specific contacts and thereby achieve high sequence specificity. In eukaryotes, the situation is much more complicated, even in the most simple cases.

Gene activation or repression in eukaryotes is performed by large protein complexes. Some members of these complexes bind DNA directly; others mediate interactions between different DNA-binding proteins, each forming only a small number of specific contacts on DNA. This makes the regulating signal quite extended (sometimes up to thousands of base-pairs long) and very vaguely defined. Also, whereas prokaryotic genes are switched on or off by one signal, eukaryotic genes are often regulated by quite a few different signals, interacting with different TFs and integrated when all these TFs bind the DNA – an example of the so-called *combinatorial control* [2, 118].

The small number of significant contacts formed on the DNA by eukaryotic transcription factors makes them very nonspecific. In addition to functional regulatory sequences, there may be thousands of pseudosites at random places on the DNA. However, when the sites for all relevant TFs are in close proximity, the resulting regulatory complex is highly specific and very stable.

It is clear that the framework presented in the first part of this thesis is not directly applicable to such complex situations. However, it can provide a quantitative insight for further development. For example, consider the kinetics of the regulatory complex formation. The traditional "building-block" picture of molecular biologists is based on *recruitment* of some members of the regulatory complex by other ones. Physically, this may mean that the high final specificity of the complex is achieved gradually, on several timescales. The first TF to find its site on the DNA will be bound rather loosely and will stay there for a short period of time, say, 10 msec. However, this waiting time may be sufficient to bind another TF. Together, the two members of the complex bind the DNA much stronger. Since the lifetime of the complex grows exponentially with the interaction energy, the recruitment of the next member can occur on the scale of a few seconds, etc. This hierarchy of lifetimes may also reflect itself in the hierarchy of concentrations of each TF.

6.1.3 Sequence, energy and folding

One of the main novel ideas introduced in this thesis is the notion of sequence dependence of the interaction with non-cognate DNA and the role it may play in the target location. While there is no reason to ignore this possible dependence, and there is no firm experimental evidence disproving it, most biochemists stick to the original von Hippel picture where $\sigma_{\text{search}} = 0$. Most biochemical and structural studies concentrate on specific rather than non-specific complexes. A possible (and probably the real) explanation of this fact is that, for a given TF, there usually just a few cognate sequences and $10^7 - 10^9$ non-cognate ones. Also, most non-cognate complexes are too unstable for proper structural studies.

However, in the light of our theory, it would be instructive to study the range of binding energies corresponding to the *search* mode of protein-DNA interaction. This may be not too difficult, though possibly quite time-consuming. For example, standard biochemical binding assays used on a large random set of binding sequences may provide us with at least some answers, such as the approximate value of σ_{search}.

A much more difficult point to establish is the correlation between *search* and *recognition* energy profiles. The main conceptual problem is that at equilibrium either only cognate or only non-cognate complexes are observed, depending on the sequence. Modern protein engineering techniques may provide a solution by stabilizing the protein in the *search* mode and measuring its binding affinity to both cognate sequences and sequences with mutations at random positions. Recently, Kalodimos *et al.* [59] used a DNA sequence that was mutated virtually at all positions compared to the consensus sequence. The non-cognate structure they reported had most of its protein-DNA contacts in the sugar-phosphate backbone, thus allegedly proving the widely accepted view of sequence independence of non-specific interactions. However, using a variety of binding sequences, this all-or-none picture could be significantly refined, which would be an important contribution to our current understanding of protein-DNA interaction.

6.1.4 DNA conformation effects

One of the central parameters of our model is τ_{3d}, the mean interval of time between a dissociation of the protein from DNA till the next binding to DNA. Exact calculation of τ_{3d} is a very difficult task, considering the nontrivial packing of the DNA molecule inside a bacterial cell, electrostatic effects and the inhomogeneity of the cytoplasm.

Considering the microscopic picture, one can easily obtain a reasonable estimate for the upper limit of τ_{3d} as a characteristic time of 3D diffusion across the nucleoid (the region of a bacterial cell to which the DNA is confined). The corresponding diffusion length depends on the conformation of the DNA molecule. If the DNA molecule was a single homogeneous globule, there would be a single relevant length scale, which is the molecule characteristic size l_{m} (the radius of gyration). On the other hand, as Fig. 6-1 shows, diffusion of a protein molecule inside a more realistic non-homogeneous multi-domain molecule involves at least one additional length scale l_{d}, which is a characteristic size of a domain. These two lengths may differ by a factor of ~ 10 [119], making the ratio of the resulting diffusion times $\tau_{3d}^{m}/\tau_{3d}^{d} \sim 10^2$. In the original problem (a single protein molecule searching for a single site on the DNA), the search process is dominated by the larger time-scale, since at least few domains must be explored before the target site is located. However, there are often about 10^2 TF molecules present in a cell, so it is reasonable to assume that the domains are scanned in parallel, making the inter-domain transfer processes irrelevant.

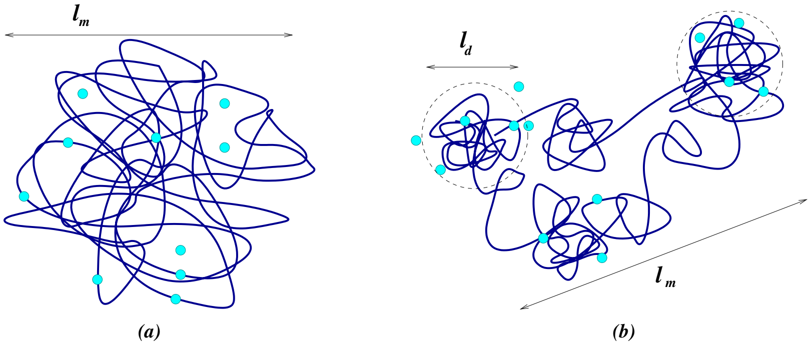


Figure 6-1: Effect of DNA conformation on the effective diffusion distance: (a) Single globule; (b) Multi-domain conformation.

6.2 Random walks and polymers

The above discussion about a protein diffusing inside a DNA molecule is in fact a special case of a rather old problem. In its most general form, it can be formulated as a random walk interacting with absorbing manifold. The walk can be Gaussian or self-avoiding, normal or anomalous [30, 120]; the manifold can be smooth or fractal. In the second part of this thesis, we study a couple of examples demonstrating that most posible complications of the problem beyond normal diffusion in fairly simple geometries makes the problem only approximately tractable (if at all). If the diffusion is anomalous, very unusual features emerge even in the simplest geometries. For instance, the method of images has recently shown to be inapplicable for a superdiffusing particle in a half-space [121].

There is no formal way to describe the conformation of DNA in a real cell. It certainly has many elements we expect to meet in randomly conformed polymers, it has compact and swollen regions; in addition, DNA is irregularly looped and supercoiled [122]. Solving the diffusion equation with absorbing boundary conditions on such a structure is clearly impossible. Nevertheless, it is instructive to study the influence of possible regular conformation elements, such as swollen self-avoiding coil or compact globule on the 3D diffusion. Standard procedures exist for simulating these structures, so that this problem is accessible at least numerically.

To complete the picture, we mention several attempts to treat the problem analytically. Oshanin *et al*. [123, 124] have analyzed chemical reaction kinetics in polymer systems and recognized the importance of correlations in reacting particle positions. These correlations are very strong in polymer-trap systems. It was also pointed out that the tail of the diffusion time distribution is governed by the distribution of trap-free volumes. In a uniform solution of absorbers, the diffusion times are distributed exponentially, as dictated by a simple diffusion equation with traps. However, in a dense polymer solution, the distribution of cavities is quite nontrivial, which produces a stretched exponent in the tail of difusion time distribution. In a dilute solution or in the presence of a single (infinite) absorbing swollen polymer, there is no cutoff on the maximal trap-free volume. In this case, the picture is much more complicated. Cates and Witten [125] have studied this limit using an RG scheme which revealed the multifractal nature of the probability density for the diffusing particle. It is possible that diffusion times inside a random coil have a power-law distribution, and the association rates in this limit have a strong stochastic component. However, as far as protein-DNA association kinetics are concerned, the "average" DNA state is probably more appropriately described by some kind of a random globule, with a finite average density. In this limit, there is nothing anomalous about 3D diffusion times and one can more or less safely assume that their distribution is exponential (see e.g. [126]), so that τ_{3d} is well defined. As a possible extension, it would be interesting to consider the effect of correlations between 3D diffusion and the distance along DNA between dissociation and reassociation sites. Such correlations could introduce important corrections to the presented framework or even reformulate it anew.

Appendixes

Appendix A

Protein-DNA interaction energetics

A.1 Equilibrium Model for Interaction Energy

Protein–DNA interaction plays a central role in many critical cellular functions [127]. A broad class of DNA–binding proteins, such as transcription factors, restriction and DNA repair enzymes, exhibit a vast ($\sim 3-8$ orders of magnitude) range of affinities to DNA depending on the actual underlying sequence. Each such protein has one or several *cognate sites* on the DNA molecule where the binding is the strongest. The exact location of these sites on the chromosome usually has a clear functional (e.g. regulatory) meaning, therefore, knowing the sequence→energy mapping function would be of great assistance to biologists and bioengineers. Unfortunately, an exact calculation of such a function is extremely difficult and requires the knowledge of many interaction parameters that can be neither derived nor measured with the desired degree of precision. Instead, a number of heuristic knowledge–based models have been developed in the recent years [16, 25, 40, 45]. These models differ a lot in complexity and site prediction reliability; however, they usually produce binding energy spectra sharing many common features. For instance, cognate sites always reside at the lower edge of binding energy spectrum, which should be broad enough to ensure cognate complex stability with respect to the rest of the genome. Also, most spectra can be approximated by a Gaussian over a wide range of energies [24, 45]. These properties have been verified in a number of equilibrium measurements [128, 129, 9].

In this section, we discuss the first attempts to formulate the evolutionary framework for protein-DNA interaction. This theory has been applied quite successfully to a variety of experimental situations [130, 131] and remains a default starting point for any theoretical work in the field.

A.1.1 Berg - Von Hippel theory

Exact (*ab-initio*) calculation of a protein-DNA complex energy is generally a very difficult problem. The energy constituents are, to name a few: the direct electrostatic interaction between charged elements of the protein and the DNA (e.g. phosphate backbone), hydrogen bonds between binding domain amino acids and DNA bases, effective hydrophobic interactions, water-mediated interactions, etc. Though much effort is invested presently in this direction, a coherent picture is still missing.

However, a heuristic approach to this seemingly intractable problem originating in the seminal papers by von Hippel and Berg has proved to be very successful. The complete theory is described in detail elsewhere [16, 132]; here, we provide only the necessary background.

Suppose there are n_s specific sites of length L for a given regulatory protein. In thermal equilibrium with proteins in solution, the probability of a certain site i to be occupied (or a site binding constant) is proportional to a Boltzmann factor $e^{-\beta E_i}$. Then, by measuring site affinities, it is possible to estimate site binding energies. Futhermore, if we assume that each base contributes independently to the binding energy, it is possible to measure individual contributions of the bases by mutating the binding sequence[1].

The argument of Berg and von Hippel is based on the analogy they draw between thermodynamic picture and an evolutionary selection process. This analogy appears reasonable if we assume that during evolution, only sequences with binding energies

[1] The independence conjecture has been verified experimentally for a very wide class of transcription factors. In this paper, we ignore correlations between bases inside the binding sequences.

in a certain interval $E_s \pm \Delta/2$ are selected. Suppose that the binding domain of the regulatory protein is conserved throughout the evolution process and that there exists some strongest (*consensus*) binding sequence. Then every base-pair mismatch in the sequence will weaken the binding by a certain *discrimination energy*, the value of which depends both on the position and the identity of the mutated base-pair. If all positions are equally important and any mutation contributes the same discrimination energy, then specifiyng the required sequence energy (for selection) is equivalent to specifying the number of base-pair mismatches.

In Fig. A-1, the results of numerical simulation for 20 bp sequence binding energy for a random "genome" of size 10^7 bp are shown. The logarithm of the density of states $\Omega(E)$ can be quite adequately fitted by a parabola, which is merely a consequence of the Central Limit Theorem (CLT) applied to a sum of 20 random variables. Thus, the genome binding energy spectrum can be described by the Random Energy Model (REM) [24, 61], so that we can define the *evolutionary temperature* T^* as

$$T^* \equiv \left[\frac{d}{dE}\ln\Omega(E)\right]^{-1}_{E=E_s} = \frac{\Sigma^2}{|E_s - \langle E\rangle|}, \tag{A.1}$$

where Σ^2 is the variance and $\langle E\rangle$ is the average binding energy. This equation establishes the transition to the canonical description, which is more appropriate in the general case, when different positions and mutations contribute nonequally. Then, if the entire genome is at "thermal equilibrium" at temperature T^*, the partition function for a set of all possible sequences of length L is

$$Z^* = \prod_{i=1}^{L}\sum_{\alpha=1}^{4} e^{-\beta^*\epsilon_{i,\alpha}}, \tag{A.2}$$

where α counts the possible mutations and $\epsilon_{i,\alpha}$ is the corresponding discrimination energy. Under these conditions, the probability of α-th base to be observed in the selected sequence at the i-th position is

$$p_\alpha(i) = \frac{e^{-\beta^*\epsilon_{i,\alpha}}}{\sum_{\lambda=1}^{4} e^{-\beta^*\epsilon_{i,\lambda}}}. \tag{A.3}$$

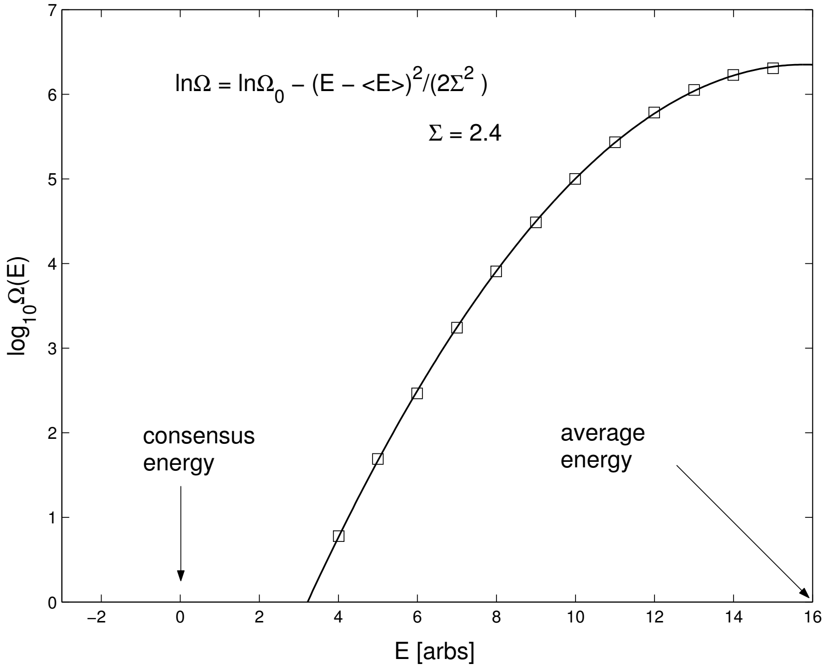


Figure A-1: Energy spectrum of 20 bp sequence with unit discrimination energy. The squares are the results of computer simulation; the solid line is a quadratic fit.

Thus, if a collection of binding sites for a certain protein is known, it is possible to estimate the binding energies (up to a certain constant factor[2]) by observing the base frequencies at various positions in the sites and taking a logarithm, thus constructing the *weight matrix* [130]. The weight matrix is a characteristic of the binding domain of the protein; applying it to any arbitrary DNA sequence produces this sequence binding energy (see Fig. A-2).

A.1.2 Energy Gap

A large energy gap between the cognate site $\vec{s_c}$ and the bulk of genomic sites would solve the paradox of rapid search and stability. One may seek parameters $\epsilon(j, s)$ of the energy function

$$U(\vec{s} = s_i, ..s_{i+l-1}) = \sum_{j=1}^{l} \epsilon(j, s_j), \tag{A.4}$$

[2] Interestingly enough, most experiments[130, 131] suggest values of $T^*/T \sim 1$.

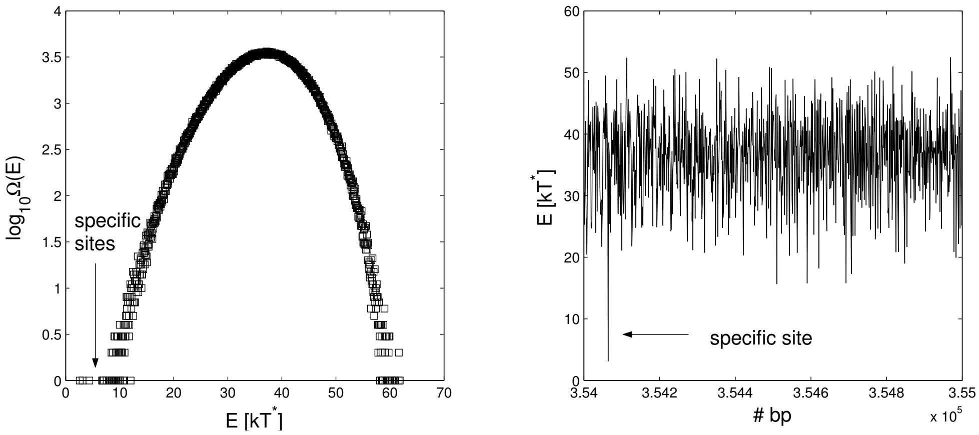


Figure A-2: Energy spectrum and energy profile for E. coli purine repressor ($PurR$). The weight matrix was built by analyzing 35 known binding sites for $PurR$.

to maximize the energy gap by minimizing the Z-score

$$Z(\vec{s}_c) = \frac{U(\vec{s}_c) - \langle U \rangle}{\sigma}, \tag{A.5}$$

where both the mean and the variance are taken over all possible sequences of length l (or over genomic words of length l). It's easy to see that $Z(\vec{s}_c)$ is minimal if

$$\epsilon^{\text{opt}}(j, s) = -\delta(s, s_{cj}) \tag{A.6}$$

where $\delta(x, y)$ is Kronecker delta. For K types of nucleotides, assuming their equal frequency in genome, we obtain the maximal reachable energy gap of

$$Z^{\min} = -\sqrt{lK}. \tag{A.7}$$

For $K = 4$ and $l \approx 8$ we get $Z^{\min} \approx -5$. For the genome of 10^6-10^7bp the energy spectrum of the genomic DNA ends at $Z \approx -5$. While sufficient to provide stability of the bound complex (see main text), such an energy gap is unable to resolve the search-stability paradox.

Appendix B

Diffusion in a Half-Space - Classical results

B.1 Diffusion and random walks

B.1.1 Some history

The theory of diffusion was first developed in the beginning of the 19th century by Joseph Fourier; his work was summarized in the famous *Théorie analytique de la chaleur* [133], first published in 1822. It contains an extensive treatment of *homogenous* heat diffusion problems for a variety of geometries, mostly by the variable separation method.

The first generalized approach to solving non-homogenous diffusion probems was formulated by Sir William Thomson [134], more widely known as Lord Kelvin, in 1850. He realized that particular solutions can be obtained by superposition of solutions for "instantaneous simple point sources" (which are now called by physicists "Dirac's delta–functions"). In short, what he did was to invent the Green's function method for the diffusion equation; it was later used by E. W. Hobson to treat heat-conduction problems with a variety of sources and boundary conditions [135].

Kelvin was also the first to apply the method of images to account for boundary conditions for electricity conduction in a semi–infinite telegraph line [134].

B.1.2 Boundary conditions

Consider a N–step random walk starting at $\mathbf{r}_0 = 0$ in the three–dimensional (3D) space. Let $G(\mathbf{r}, N)$ be the probability density for the walk to end at $\mathbf{r}$. For large N and in the absence of obstacles and boundaries, $G(\mathbf{r}, N)$ is a solution of the *diffusion equation* [30]

$$\left(\frac{\partial}{\partial N} - \frac{1}{2}\nabla^2\right) G(\mathbf{r}, N) = 0 \tag{B.1}$$

with the initial condition

$$G(\mathbf{r}, 0) = \delta(\mathbf{r}). \tag{B.2}$$

Here, we took the diffusion coefficient $D = 1/2$. The solution has a well–known form

$$G(\mathbf{r}, N) = \frac{e^{-\mathbf{r}^2/(2N)}}{(2\pi N)^{3/2}}. \tag{B.3}$$

It is reasonable therefore to assume that solutions of the same kind can as well be found for any bounded region $\mathbb{R}$. Naturally, one has to specify the boundary conditions. This is not as trivial as it appears and, in fact, depends on the physical context of the problem. If, for example, the random walker is allowed to touch the boundary and then step back with probability 1, the *reflecting* boundary conditions are appropriate. Formally, it means that the flux across the boundary vanishes

$$\nabla G(\mathbf{r}, N) \cdot \mathbf{n}|_{\mathbf{r}\in\partial\mathbb{R}} = 0. \tag{B.4}$$

Here $\partial\mathbb{R}$ denotes the boundary of the region $\mathbb{R}$ and $\mathbf{n}$ is a unit vector locally normal to $\partial\mathbb{R}$. Another possible choice of boundary conditions corresponds to the case when the walker sticks to the boundary upon reaching it – the *absorbing* boundary conditions

$$G(\mathbf{r}, N)|_{\mathbf{r}\in\partial\mathbb{R}} = 0. \tag{B.5}$$

B.1.3 Method of images

Consider a random walk starting at $\mathbf{r}_0 = \hat{\mathbf{z}}a$ away from the plane $z = 0$ and confined to the $z > 0$ half–space. For the absorbing boundary, we expect the probability distribution for the end point to satisfy the following boundary value problem

$$\begin{cases} \left(\dfrac{\partial}{\partial N} - \dfrac{1}{2}\nabla^2\right) G(\mathbf{r}, N; a) = 0 \\ G(z = 0) = 0, \quad G(\mathbf{r}, 0) = \delta(\mathbf{r} - \hat{\mathbf{z}}a) \end{cases} \tag{B.6}$$

Any introductory textbook on PDEs contains a straightforward solution of this problem, which consists of introducing a sink, or *negative image*, at $(0, 0, -a)$ and extending the problem to the entire space[1]. The boundary condition at $z = 0$ is then automatically satisfied and the solution is

$$G(\mathbf{r}, N; a) = \frac{1}{(2\pi N)^{3/2}} \left[e^{-(\mathbf{r}-\hat{\mathbf{z}}a)^2/(2N)} - e^{-(\mathbf{r}+\hat{\mathbf{z}}a)^2/(2N)}\right]. \tag{B.7}$$

For $a \ll \sqrt{N}$, we can expand the expression in parentheses to obtain

$$G(\mathbf{r}, N; a) \simeq \left(\frac{2az}{N}\right) \frac{e^{-\mathbf{r}^2/(2N)}}{(2\pi N)^{3/2}}. \tag{B.8}$$

We see that the probability distribution can be factorized into z–dependent and z–independent parts. The latter, which includes degrees of freedom parallel to the boundary, is not affected by the presence of the boundary. Thus, in what follows we will be predominantly occupied with the z–dependent part of $G(\mathbf{r}, N; a)$.

B.1.4 Counting walks on a lattice

Chandrasekhar [136] suggested a direct way of counting the paths on a lattice when a reflecting or an absorbing boundary is present. We will briefly describe the derivation for an absorbing boundary. The reader is encouraged to read the original paper which, despite being written more than half a century ago, remains one of the best

[1] It is straightforward to verify that reflecting boundary conditions correspond to a positive image.

introductions into random walks and stochastic processes in general.

Consider a one–dimensional random walker on a lattice (discrete z–axis) with absorbing boundary at $z = 0$. Suppose, the walk starts some distance n from the origin; our task is to calculate the number of paths leading from n to some other point m, *without touching* the boundary. It turns out that it is easier to calculate the number of paths that do touch the boundary and then to subtract it from the total number of paths leading from n to m. To do so, we make use of a very elegant theorem – *the reflection principle.*

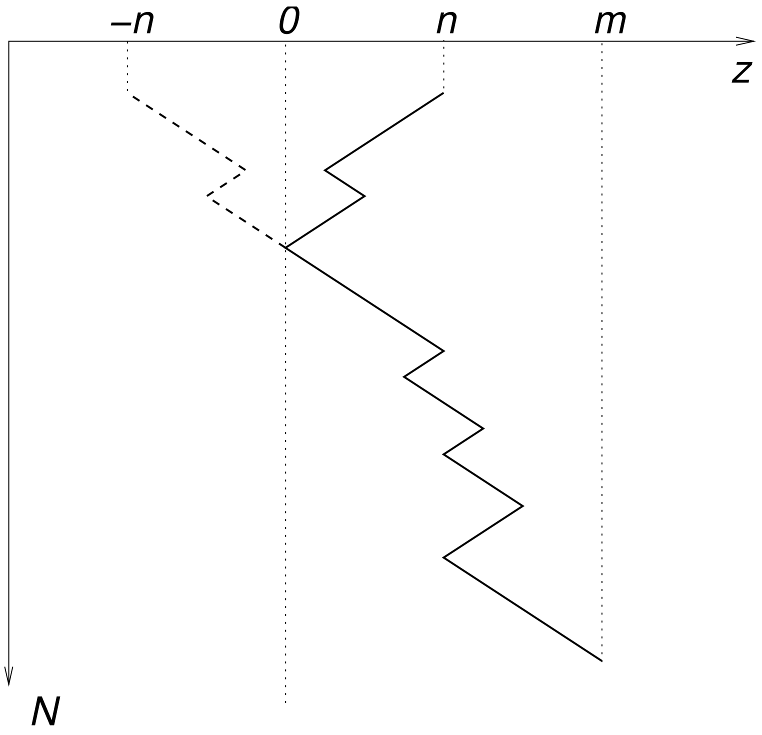


Figure B-1: The reflection principle.

Let us extend our lattice to include the negative part of the z–axis as well. Then, the reflection principle states that the number of N–step paths originating at n, ending at m and touching or crossing the boundary $z = 0$ is equal to the number of N–step paths that originate at $-n$ and end at m. Figure B-1 illustrates the reflection principle by presenting a way to build a one–to–one mapping between the two sets of paths. Thus, the number of paths not touching the boundary is

$$\mathfrak{N} = \begin{pmatrix} N \\ \frac{1}{2}[N+m-n] \end{pmatrix} - \begin{pmatrix} N \\ \frac{1}{2}[N+m+n] \end{pmatrix}. \tag{B.9}$$

For the starting point near the boundary and $m \ll N$, we can expand the binomial coefficients using Stirling's formula to obtain

$$\mathfrak{N} \simeq 2^N \left(\frac{2}{\pi N}\right)^{1/2} \frac{m\ e^{-m^2/(2N)}}{N}. \tag{B.10}$$

Dividing by the total number of paths of length N (which is 2^N), we obtain the probability density for a path to start near the boundary and to end at some point m without returning to the boundary

$$G(m,N) \simeq \frac{2m}{N} \frac{e^{-m^2/(2N)}}{(2\pi N)^{1/2}}. \tag{B.11}$$

Bibliography

[1] R. Wagner. *Transcription regulation in prokaryotes.* Oxford University Press, 2000.

[2] J.D. Watson, T.A. Baker, S.P. Bell, A. Gann, M. Levine, and R. Losick. *Molecular biology of the gene.* Benjamin Cummings, 2004.

[3] M. Ptashne. *Genetic Switch: Phage λ Revisited.* CSHL Press, 2004.

[4] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol.*, 1:1–37, 2000.

[5] C. E. Bell and M. Lewis. The Lac repressor: a second generation of structural and functional studies. *Curr. Opin. Struct. Biol.*, 11:19–25, 2001.

[6] C. E. Bell and M. Lewis. A closer view of the conformation of the Lac repressor bound to operator. *Nat. Struct. Biol.*, 7:209–214, 2000.

[7] M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271:1247–1254, 1996.

[8] M. A. Schumacher, K. Y. Choi, H. Zalkin, and R. G. Brennan. Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. *Science*, 266:763–770, 1994.

[9] Y. Takeda, A. Sarai, and V. M. Rivera. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl. Acad. Sci. USA*, 86:439–443, 1989.

[10] A. O. Grillo, M. P. Brown, and C. A. Royer. Probing the physical basis for trp repressor-operator recognition. *J. Mol. Biol.*, 287:539–554, 1999.

[11] R. S. Spolar and M. T. Record. Coupling of local folding to site-specific binding of proteins to DNA. *Science*, 263:777–784, 1994.

[12] N. Shimamoto. One-dimensional diffusion of proteins along DNA. its biological and chemical significance revealed by single-molecule measurements. *J. Biol. Chem.*, 274:15293–15296, 1999.

[13] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20:6929–6948, 1981.

[14] R. B. Winter, O. G. Berg, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor–operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20:6961–6977, 1989.

[15] P. H. von Hippel and O. G. Berg. Facilitated target location in biological systems. *J. Biol. Chem.*, 264:675–678, 1989.

[16] O. G Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750, 1987.

[17] A. D. Riggs, H. Suzuki, and S. Bourgeois. Lac repressor-operator interaction. 1. Equilibrium studies. *J. Mol. Biol.*, 48:67–83, 1970.

[18] A. D. Riggs, S. Bourgeois, and M. Cohn. The Lac repressor-operator interaction. 3. Kinetic studies. *J. Mol. Biol.*, 53:401–417, 1970.

[19] P. H. Richter and M. Eigen. Diffusion controlled reaction rates in spheroidal geometry. Application to repressor–operator association and membrane bound enzymes. *Biophys. Chem.*, 2:255–263, 1974.

[20] H. Flyvbjerg, F. Jülicher, P. Ormos, and F. David, editors. *Physics of biomolecules and cells*, volume 75 of *Les Houches*, chapter 1. Springer-Verlag Heidelberg, 2002.

[21] R. F. Bruinsma. Physics of protein-DNA interaction. *Physica A*, 313:211–237, 2002.

[22] M. B. Elowitz, M. G. Surette, P. E. Wolf, J. B. Stock, and S. Leibler. Protein mobility in the cytoplasm of Escherichia Coli. *J. Bacteriol.*, 181:197–203, 1999.

[23] J. G. Kim, Y. Takeda, B. W. Matthews, and W. F. Anderson. Kinetic studies on Cro repressor-operator DNA interaction. *J. Mol. Biol.*, 196:149–158, 1987.

[24] U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proc. Natl. Acad. Sci. USA*, 99:12015–12020, 2002.

[25] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 23:109–113, 1998.

[26] D.A. Erie, G. Yang, H.C. Schultz, and C. Bustamante. DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity. *Science*, 266:1562–6, 1994.

[27] J. P. Bouchaud and A. Georges. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. *Phys. Rep.*, 195:127–293, 1990.

[28] K. P. N. Murthy and K. W. Kehr. Mean first-passage time of random walks on a random lattice. *Phys. Rev. A*, 40:2082–2087, 1989.

[29] I. Goldhirsh and Y. Gefen. Analytic method for calculating properties of random walks on networks. *Phys. Rev. A*, 33:2583–2594, 1986.

[30] B. D. Hughes. *Random Walks and Random Environments.* Oxford University Press, New York, 1995.

[31] L.D. Landau and E.M. Lifshitz. *Fluid Mechanics.* Butterworth-Heinemann, 1987.

[32] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, 1981.

[33] A. Gutin, A. Sali, V. Abkevich, M. Karplus, and E.I. Shakhnovich. Temperature dependence of the folding rate in a simple protein model: Search for a glass transition. *J. Chem. Phys.*, 108:6466–6483, 1998.

[34] A.V. Finkelstein and O.B. Ptitsyn. *Protein Physics.* Academic Press, 2002.

[35] V. Pande, A. Grosberg, and T. Tanaka. Heteropolymer freezing and design: Towards physical models of protein folding. *Rev. Mod. Phys.*, 72:259–314, 2000.

[36] C. G. Kalodimos, N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, 305:386–9, 2004.

[37] M. Akke. NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis. *Curr. Opin. Struct. Biol.*, 12:642–647, 2002.

[38] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[39] D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22:403–434, 1976.

[40] K. Robison, A. M. McGuire, and G. M. Church. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete escherichia coli K-12 genome. *J. Mol. Biol.*, 284:241–254, 1998.

[41] P. Hänggi, P. Talkner, and M. Borkovec. Reaction-rate theory: fifty years after kramers. *Rev. Mod. Phys.*, 62:251, 1990.

[42] M. Vendruscolo and C. M. Dobson. Towards complete descriptions of the free-energy landscapes of proteins. *Phil. Trans. R. Soc. A*, 363:433, 2005.

[43] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. Diffusive dynamics of the reaction coordinate for protein folding funnels. *The Journal of Chemical Physics*, 104(15):5860–5868, 1996.

[44] R. Du, V. S. Pande, A. Yu. Grosberg, T. Tanaka, and E. S. Shakhnovich. On the transition coordinate for protein folding. *The Journal of Chemical Physics*, 108(1):334–350, 1998.

[45] M. Djordjevic, A.M. Sengupta, and B.I. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res.*, 13:2381–90, 2003.

[46] C.G. Kalodimos, N. Biris, A.M. Bonvin, M.M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, 305:386–389, 2004.

[47] M. Slutsky and L.A. Mirny. Kinetics of protein–DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.*, 87:4021–4035, 2004.

[48] M. S. Gelfand. 2003. private communication.

[49] J. J. Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA*, 71:4135–4139, 1974.

[50] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, 293:321–331, 1999.

[51] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, 12:54–60, 2002.

[52] V. N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, 11:739–756, 2002.

[53] P. D. Williams, D. D. Pollock, and R. A. Goldstein. Evolution of functionality in lattice proteins. *J. Mol. Graph. Model.*, 19:150–6, 2001.

[54] B.A. Shoemaker, J.J. Portman, and P.G. Wolynes. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U S A*, 97:8868–73, 2000.

[55] Y. Levy, P.G. Wolynes, and J.N. Onuchic. Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. U S A*, 101:511–6, 2004.

[56] F. K. Winkler, D.W. Banner, C. Oefner, D. Tsernoglou, R.S. Brown, S.P. Heathman, R.K. Bryan, P.D. Martin, K. Petratos, and K.S. Wilson. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.*, 12:1781–95, 1993.

[57] N.D. Clarke, L.J. Beamer, H.R. Goldberg, C. Berkower, and C.O. Pabo. The DNA binding arm of lambda repressor: critical contacts from a flexible region. *Science*, 254:267–70, 1991.

[58] K.T. O'Neil, R.H. Hoess, and W.F. DeGrado. Design of DNA-binding peptides based on the leucine zipper motif. *Science*, 249:774–8, 1990.

[59] C. G. Kalodimos, G. E. Folkers, R. Boelens, and R. Kaptein. Strong DNA binding by covalently linked dimeric Lac headpiece: evidence for the crucial role of the hinge helices. *Proc. Natl. Acad. Sci. USA*, 98:6039–6044, 2001.

[60] C. G. Kalodimos, R. Boelens, and R. Kaptein. A residue-specific view of the association and dissociation pathway in protein–DNA recognition. *Nat. Struct. Biol.*, 9:193–197, 2002.

[61] B. Derrida. Random-energy model - an exactly solvable model of disordered-systems. *Phys. Rev. B*, 24:2613–2626, 1981.

[62] J. Bryngelson and P. Wolynes. Spin-glasses and the statistical-mechanics of protein folding. *Proc. Natl. Acad. Sci. USA*, 84:7524, 1987.

[63] J. P. Bouchaud, A. Comtet, A. Georges, and P. Le Doussal. Classical diffusion of a particle in a one-dimensional random force-field. *Ann. Phys.*, 201:285, 1990.

[64] P. G. De Gennes. Brownian-motion of a classical particle through potential barriers - application to helix-coil transitions of heteropolymers. *J. Stat. Phys*, 12:463, 1975.

[65] D. Cule and T. Hwa. Polymer reptation in disordered media. *Phys. Rev. Lett.*, 80:3145, 1998.

[66] P. Le Doussal and V. M. Vinokur. Creep in one-dimension and phenomenological theory of glass dynamics. *Physica C*, 254:63, 1995.

[67] D. S. Fisher, P. Le Doussal, and C. Monthus. Random walks, reaction-diffusion, and nonequilibrium dynamics of spin chains in one-dimensional random environments. *Phys. Rev. Lett.*, 80:3539, 1998.

[68] T. Hwa, E. Marinari, K. Sneppen, and L. Tang. Localization of denaturation bubbles in random DNA sequences. *Proc. Natl. Acad. Sci. USA*, 100:4411, 2003.

[69] Ya. G. Sinai. The limit behavior of random walks in a one-dimensional random environment. *Theory Probab. Appl.*, 27:247, 1982.

[70] S. H. Noskowicz and I. Goldhirsh. Average versus typical mean 1st-passage time in a random random-walk. *Phys. Rev. Lett.*, 61:500, 1988.

[71] M. G. Munteanua, K. Vlahoviček, S. Parthasarathya, I. Simon, and S. Pongor. Rod models of DNA: sequence-dependent anisotropic elastic modelling of local bending phenomena. *Trends Biochem. Sci.*, 23:341, 1998.

[72] K. Vlahoviček, L. Kaján, and S. Pongor. DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Res.*, 13:3686, 2003. http://www.icgeb.trieste.it/dna/.

[73] I. Brukner, R. Sanchez, D. Suck, and S. Pongor. Sequence-dependent bending propensity of DNA as revealed by DNAse-I - parameters for trinucleotides. *EMBO J.*, 14:1812, 1995.

[74] A. Gabrielian and S. Pongor. Correlation of intrinsic DNA curvature with DNA property periodicity. *FEBS Lett.*, 393:65, 1996.

[75] H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart. Polymer reptation and nucleosome repositioning. *Phys. Rev. Lett.*, 86:4414, 2001.

[76] J. Widom. Structure, dynamics, and function of chromatin in vitro. *Annu. Rev. Biophys. Biomol. Struct.*, 27:285, 1998.

[77] J. Widom. Role of DNA sequence in nucleosome stability and dynamics. *Quart. Rev. Biophys*, 34:269, 2001.

[78] R. Kiyama and E. N. Trifonov. What positions nucleosomes? - a model. *FEBS Lett.*, 523:7, 2002.

[79] A. Meller. Dynamics of polynucleotide transport through nanometre-scale pores. *J. Phys.: Condens. Matter*, 15:R581, 2003.

[80] P. G. de Gennes. *Scaling Concepts in Polymers Physics.* Cornell University Press, Ithaca, New York, 1979.

[81] M. Doi and S. F. Edwards. *The Theory of Polymer Dynamics.* Clarendon Press, New York, 1986.

[82] K. F. Freed. *Renormalization Group Theory of Macromolecules.* John Wiley, 1987.

[83] R. P. Feynman and A. R. Hibbs. *Quantum Mechanics and Path Integrals.* McGraw-Hill Higher Education, 1965.

[84] F. W. Wiegel. *Introduction to path-integral methods in physics and polymer science.* World Scientific, 1986.

[85] H. Salman, D. Zbaida, Y. Rabin, D. Chatenay, and M. Elbaum. Kinetics and mechanism of DNA uptake into the cell nucleus. *Proc. Natl. Acad. Sci. USA*, 98:7247, 2001.

[86] R. J. Davenport, G. J. Wuite, R. Landick, and C. Bustamante. Single-molecule study of transcriptional pausing and arrest by E. coli RNA polymerase. *Science*, 287(5462):2497–2500, 2000.

[87] Julio M. Fernandez and Hongbin Li. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science*, 303(5664):1674–1678, 2004.

[88] J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco, Jr., and C. Bustamante. Reversible unfolding of single RNA molecules by mechanical force. *Science*, 292(5517):733–737, 2001.

[89] H. Clausen-Schaumann, M. Seitz, R. Krautbauer, and H. E. Gaub. Force spectroscopy with single bio-molecules. *Curr. Opin. Chem. Biol.*, 4:524–530, 2000.

[90] M.C. Williams and I. Rouzina. Force spectroscopy of single DNA and RNA molecules. *Curr. Opin. Struct. Biol.*, 12:330–336, 2002.

[91] P. G. de Gennes. *Scaling Concepts in Polymer Physics.* Cornell University Press, Ithaca, New York, 1979.

[92] J. L. Cardy. Critical behaviour at an edge. *J. Phys. A*, 16(15):3617–3628, 1983.

[93] J. L. Cardy and S. Redner. Conformal invariance and self-avoiding walks in restricted geometries. *J. Phys. A*, 17:L933, 1984.

[94] A. J. Guttmann and G. M. Torrie. Critical behaviour at an edge for the SAW and Ising model. *J. Phys. A*, 17:3539, 1984.

[95] K. De'Bell and T. Lookman. Surface phase-transitions in polymer systems. *Rev. Mod. Phys.*, 65:87–113, 1993.

[96] M. N. Barber, A. J. Guttmann, K. M. Middlemiss, G. M. Torrie, and S. G. Whittington. Some tests of scaling theory for a self-avoiding walk attached to a surface. *J. Phys. A*, 11:1833, 1978.

[97] M. K. Kosmas. A non-ideal polymer chain interacting with a penetrable surface. *J. Phys. A*, 18:539, 1985.

[98] J. F. Douglas and M. K. Kosmas. Flexible polymer with excluded volume at an interacting penetrable surface of variable dimension: a multiple-ϵ perturbation theory. *Macromolecules*, 22:2412, 1989.

[99] S. F. Edwards. The statistical mechanics of polymers with excluded volume. *Proc. Phys. Soc. London*, 85:613, 1965.

[100] Y. Oono, T. Ohta, and K. F. Freed. Application of dimensional regularization to single chain polymer static properties. III. Conformational space renormalization of polymers. *J. Chem. Phys.*, 74:6458, 1981.

[101] K. F. Freed. Excluded volume effects in polymers attached to surfaces: Chain conformational renormalization group. *J. Chem. Phys.*, 79:3121, 1983.

[102] B. Duplantier. Interaction of crumpled manifolds with euclidean elements. *Phys. Rev. Lett.*, 62:2337, 1989.

[103] R. Zandi, Y. Kantor, and M. Kardar. Entropic competition between knots and slip-links. *ARI*, 53:6–15, 2003. cond-mat/0306587.

[104] B. Marcone, E. Orlandini, A. Stella, and F. Zonta. What is the length of a knot in a polymer? *J. Phys. A: Math. Gen.*, 38:L15–L21, 2005.

[105] N. Madras and A. D. Sokal. The pivot algorithm - a highly efficient monte-carlo method for the self-avoiding walk. *J. Stat. Phys*, 50:109, 1988.

[106] K. Suzuki. *Bull. Chem. Soc. Japan*, 41:538, 1968.

[107] Z. Alexandrowicz. Monte carlo of chains with excluded volume: a way to evade sample attrition. *J. Chem. Phys.*, 51:561, 1969.

[108] N. Madras and G. Slade. *The Self–Avoiding Walk*. Birkhäuser, Boston, 1993.

[109] M. Slutsky, M. Kardar, and L.A. Mirny. Diffusion in correlated random potentials, with applications to DNA. *Phys. Rev. E*, 69:061903–061915, 2004.

[110] C. Kleinschmidt, K. Tovar, W. Hillen, and D. Porschke. Dynamics of repressor-operator recognition: the Tn10-encoded tetracycline resistance control. *Biochemistry*, 27:1094–1104, 1988.

[111] R. B. Winter and P. H. Von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli lac repressor-operator interaction: equilibrium measurements. *Biochemistry*, pages 6948–6960, 1981.

[112] S. B. Zimmerman and S. O. Trach. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J. Mol. Biol.*, 222:599–620, 1991.

[113] J. Ovádi and V. Saks. On the origin of intracellular compartmentation and organized metabolic systems. *Mol. Cell. Biochem.*, 256, 2004.

[114] R. J. Ellis. Macromolecular crowding: obvious but underappreciated. *Trends Biochem. Sci.*, 26:597–604, 2001.

[115] K. Lipkow, S. S. Andrews, and D. Bray. Simulated Diffusion of Phosphorylated CheY through the Cytoplasm of Escherichia coli. *J. Bacteriol.*, 187(1):45–53, 2005.

[116] I. Golding and E. C. Cox. RNA dynamics in live Escherichia coli cells. *PNAS*, 101(31):11310–11315, 2004.

[117] I. Golding. 2004. private communication.

[118] M. Ptashne and A. Gann. *Genes and signals*. CSHL Press, 2002.

[119] F. C. Neidhardt, R. Curtiss, and E. C. Lin, editors. *Escherichia Coli and Salmonella*, chapter 12. ASM Press, 1996.

[120] R. Metzler and J. Klafter. The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. *J. Phys. A: Math. Gen.*, 37:R161–R208, 2004.

[121] I.M. Sokolov and R. Metzler. Non-uniqueness of the first passage time density of Lévy random processes. *J. Phys. A: Math. Gen.*, 37:L609–L615, 2004.

[122] M. D. Frank-Kamenetskii. *Unraveling DNA: The Most Important Molecule of Life.* Addison Wesley Publishing Company, 1997.

[123] G.S. Oshanin and S.F. Burlatsky. Reaction kinetics in polymer systems. *J. Stat. Phys.*, 65, 1991.

[124] G. Oshanin, M. Moreau, and S. Burlatsky. Models of chemical reactions with participation of polymers. *Adv. Colloid Interface Sci.*, 49:1–46, 1994.

[125] M. E. Cates and T. A. Witten. Diffusion near absorbing fractals: Harmonic measure exponents for polymers. *Phys. Rev. A*, 35, 1987.

[126] M. Coppey, O. Benichou, R. Voituriez, and M. Moreau. Kinetics of Target Site Localization of a Protein on DNA: A Stochastic Approach. *Biophys. J.*, 87(3):1640–1649, 2004.

[127] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science, 2002.

[128] G. D. Stormo and M. Yoshioka. Specificity of the Mnt protein determined by binding to randomized operators. *Proc. Natl. Acad. Sci. USA*, 88:5699–5703, 1991.

[129] D. S. Fields, Y. He, A. Y. Al-Uzri, and G. D. Stormo. Quantitative specificity of the Mnt repressor. *J. Mol. Biol.*, 271:178–194, 1997.

[130] Q. K. Chen, G. Z. Hertz, and G. D. Stormo. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, 11:563–566, 1995.

[131] H. Kono and A. Sarai. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35:114–131, 1999.

[132] P. H. von Hippel and O. G. Berg. On the specificity of DNA–protein interactions. *Proc. Natl. Acad. Sci. USA*, 83:1608–1612, 1986.

[133] J. B. J. Fourier. *The analytical theory of heat.* Dover Publishers, New York, 1955.

[134] W. Thomson (Lord Kelvin). *Mathematical and Physical Papers, Vol. 2.* University Press, Cambridge, 1884.

[135] E. W. Hobson. Synthetical solutions in the conduction of heat. *Proc. Lond. Math. Soc*, 19:279–294, 1887.

[136] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:1–89, 1943.