Some inverse problems in biophysics

Carlo Barbieri

UNIVERSITE PIERRE ET MARIE CURIE

THÈSE

pour obtenir le grade de

Docteur de l'Université Pierre et Marie CURIE

Spécialité : **Physique**

préparée au **Laboratoire de Physique Statistique de l'École Normale Supérieure**

dans le cadre de l' **École Doctorale de Physique la Région Parisienne — ED 107**

présentée et soutenue publiquement
par

Carlo BARBIERI

le 01/09/2011

Titre:

Des problèmes inverses en Biophysique

Directeurs de thèse: **Simona Cocco**

Jury

M. Jean-François Joanny,	Président du jury
M. Felix Ritort,	Rapporteur
M. Massimo Vergassola,	Rapporteur
M. Christophe Deroulers,	Examinateur
M.lle. Simona Cocco,	Directeur de Thèse

Résumé

Ces dernières années ont vu le développement de techniques expérimentales permettant l'analyse quantitative de systèmes biologiques, dans des domaines qui vont de la neurobiologie à la biologie moléculaire. Notre travail a pour but la description quantitative de tels systèmes à travers des outils théoriques et numériques issus de la physique statistique et du calcul des probabilités.

Cette thèse s'articule en trois volets, ayant chacun pour but l'étude d'un système biophysique.

Premièrement, on se concentre sur l'infotaxie, un algorithme de recherche olfactive basé sur une approche de théorie de l'information proposé par Vergassola et collaborateurs en 2007: on en donne une formulation continue et on en caractérise les performances.

Dans une deuxième partie on étudie les expériences de micromanipulation à molécule unique, notamment celles de dégraffage mécanique de l'ADN, dont les traces expérimentales sont sensibles à la séquence de l'ADN: on développe un modèle détaillé de la dynamique de ce type d'expérience et ensuite on propose plusieurs algorithmes d'inférence ayant pour objectif de caractériser la séquence génétique.

Finalement, on donne une description d'un algorithme qui permet l'inférence des interactions entre neurones à partir d'enregistrements à électrodes multiples et on propose un logiciel intégré qui permettra à la communauté des biologistes d'interpréter ces expériences a partir de cet algorithme.

Abstract

During the past few years the development of experimental techniques has allowed the quantitative analysis of biological systems ranging from neurobiology and molecular biology. This work focuses on the quantitative description of these systems by means of theoretical and numerical tools ranging from statistical physics to probability theory.

This dissertation is divided in three parts, each of which has a different biological system as its focus.

The first such system is Infotaxis, an olfactory search algorithm proposed by Vergassola et al. in 2007: we give a continuous formulation and we characterize its performances.

Secondly we will focus on single-molecule experiments, especially unzipping of DNA molecules, whose experimental traces depend strongly on the DNA sequence: we develop a detailed model of the dynamics for this kind of experiments and then we propose several inference algorithm aiming at the characterization of the genetic sequence.

The last section is devoted to the description of an algorithm that allows the inference of interactions between neurons given the recording of neural activity from multi-electrode experiments; we propose an integrated software that will allow the analysis of these data.

Acknowledgments

First of all I would like to thank my advisor Simona Cocco for the time she has spent mentoring me, the patience she has shown and the countless things I learnt from her.

I am also obliged to the members of the committee for having agreed to participate to this occasion and for devoting the time needed to read my manuscript.

This dissertation would not have been possible without the interaction with many of the scientits at ENS in Paris and IAS in Princeton. In particular I wish to mention Rémi Monasson for countless hours of help and discussion. I'm also indebted to Francesco Zamponi, Marco Tarzia and Guilhem Semerjian for the scientific and human advice they have provided me with throughout my thesis. Stan Leibler for making the extremely enriching experience at Princeton possible and all the Members of the Simons' Center for System Biology at IAS with a special thought to Arvind Murugan.

I really have to thank all the staff at ENS: Annie, Marie and Nora for their professionality and warmth and Eric Perez for always taking the time of asking how things went.

I'm obliged to Jean-Pierre Nadal and Jean-François Allemand for making my teaching experience possible, to my teaching colleagues Fréderic Van Wijland, Christophe Mora, Gwendal Fève for their great advice and mentoring. I'm truly indebted to all my fellows grad students at ENS: first of all Florent Alzetto with whom I shared two offices and who is a true friend. The guys in DC21: Marc, Antoine, Félix and the two Laetitias. I also need to mention Vitor Sessak which has been of great help throughout my thesis. The geophysics lab: Rana, Penelope, Maya, Laureen, Amaya, Laure and Marianne for our meals and coffees together. The LPS cycling team: Arnaud, Ariel, Xavier, Florent, Clément and especially Sébastien Balibar.

I am really grateful to my friends in Princeton: Giulia, Joro, Francesco, Ali, Mathilde, the two Gabrieles, Julien and Daphne. They have made my nine months in Princeton a really pleasant surprise.

I wish to thank the various persons who have endured me as a roommate: Filippo, Laetitia, Vitor and Simone and everyone at the ENS college in Montrouge, especially Olivier who is always a good friend and a very stimulating mind.

I wouldn't be here without my family and their moral support, I have to thank them for who I am.

I would also like to show my gratitude to the countless Italian friends who have visited me during this happy exile in Paris, I hope they haven't forgotten me.

This thesis is dedicated to Marie for her loving presence throughout these years.

Contents

Introduction

Probabilistic models

Many systems encountered in quantitative biology are best described by probabilistic models. There are essentially three reasons why a probabilistic model would be preferred: either the process is thermally activated, either experimental conditions cannot be controlled in full detail or there are many possible realizations of annealed disorder in some of the involved variables.

Systems where the dynamics are thermally activated are widespread at the macromolecular scale (sizes ranging $1 - 100$ nm), because of this the dynamics of most systems from molecular biology will exhibit stochastic behavior. In this dissertation we will touch such systems in Part II while addressing DNA unzipping experiments.

Many biological experiments are performed in conditions where several variables cannot be controlled in detail: organisms which are genetically identical will exhibit different phenotypes, conditions of the medium will vary. In Part I we will observe turbulence can have such an effect in the description of olfactory searches.

Thirdly, many biological systems exhibit a characteristic which is similar to that of annealed disorder in condensed matter physics, that is, there are a number of variables which can be treated as random because they are drawn from an ensemble of possible realizations but do not change during experiments. Examples include DNA and RNA (where the variable is the genetic sequence), proteins (amino-acid content) and neural systems (interaction matrix). Such systems will be addressed in Parts II and III.

A probabilistic model will assign a proability to the outcome of an experiment. As it is possible to do this, the inverse problem can be of interest, that is we can assign a probability to a model or a set of parameters given the outcome of an experiment. This type of question is at the core of our thesis and of Bayesian inference.

Inference

Theory/Parameters   Experiment

Prediction
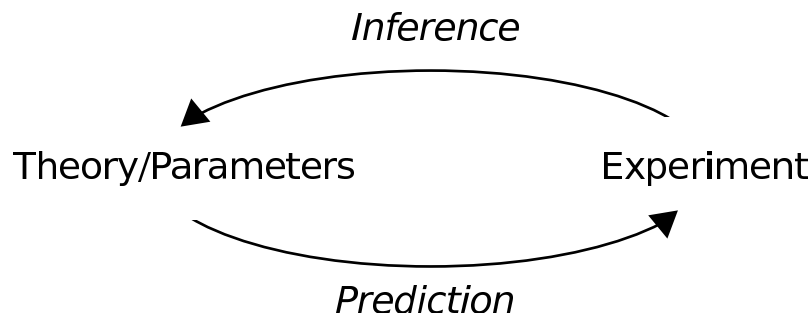
Bayes' theorem

Bayes' theorem was derived by Thomas Bayes and was only published posthumously in 1763 [Bayes 63, Bayes 58]. It is now regarded as one of the founding pillars of probability
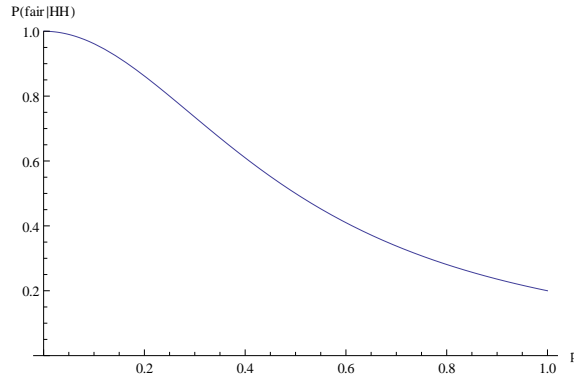
Figure 1: $P(\text{fair}|\text{HH})$ as a function of p. Note how the probability is maximum when p vanishes and it's minimum and equal to 1/5 when the unfair coin always returns heads, that is when $p = 1$.

theory.

By today's standards the name *theorem* is probably a misnomer since its derivation is a straightforward manipulation of the the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

where $P(A \cap B)$ is the probability of event A and B both happening.

If we now switch A and B and redefine combine the two definitions we obtain the classical expression of Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2}$$

where $P(A)$ is usually called the *prior*, $P(B|A)$ *likelihood function* and $P(A|B)$ *posterior*. The importance of this theorem in performing statistical inference can only be understated in fact, if one interprets A as the parameters of a model and B as the outcome of an experiment we can see how this theorem relates the predictive power of a model to the inference of the best model or set of parameters. By rewriting the model this way:

$$P(\text{model}_1|\text{data}) = \frac{P(\text{data}|\text{model}_1)P(\text{model}_1)}{\sum_i P(\text{data}|\text{model}_i)P(\text{model}_i)} \tag{3}$$

Let us give an example to further clarify this statement. Let us suppose we have two coins: one fair and one which is biased with probability p of heads turning up.

While it is straightforward to compute the outcome of an experiment knowing which coin we are handling: say two consecutive heads yield $P(\text{HH}|\text{fair} = 1/4$, we wish to know $P(\text{fair}|\text{HH})$.

Thanks to Bayes' theorem this can be done in a straightforward manner:

$$P(\text{fair}|\text{HH}) = \frac{1/4}{1/4 + p^2} \tag{4}$$

The attentive reader will have noticed we have placed ourselves in a very specific situation: we know we only have two coins, and we know the bias of one of them.

The problem of testing the hypotesis of whether a coin is biased or not in the most general conditions is a much more complicated one and is illuminating as to the limitations of

Bayesian inference.
Our toy example had the very compelling feature of defining naturally the *prior distribution*, that is $P(\text{model}_i)$ was $1/2$ for $i = 1, 2$: both coins were equiprobable. How do we define priors for more general cases?
Sometimes some general choices are available, for example one could the maximum entropy probability distribution with given characteristics such has a given support or a given expected value. However this is not always possible especially when the support of the distribution is unbounded.
However if we consider successive experiments and we refine the posterior every time we expect the choice of prior to be unimportant asymptotically.

Bayesian inference

Bayesian inference is the iterative application of Bayes' theorem to update one's knowledge about a random variable which might be a parameter of our model. It is not the only form of statistical inference, but it has several characteristics which make it more desirable than other techniques such as frequentist inference, where the frequency is interpreted as a probability.
First of all Bayesian inference will return a probability distribution, which in general contains a lot more information than an inferred value and a confidence interval.
On the other hand, as we have said before, Bayesian inference can depend strongly on the choice of a prior distribution of which there might not always be a natural choice.
Let us give an example where a Bayesian approach is much superior: a hunter is hunting with his dog, we can observe the position of the dog but we cannot observe the position of the hunter, we further know the that the dog to be located with a certain probability p in a radius r around the hunter.
The frequentist approach would lead to the following reasonment: since I have observed the dog in a given position: the hunter is in a radius r around this position with probability p.
However relies on several tacit assumptions: the isotropy of the distribution of the dog around the hunter, different directions need not be equiprobable, in fact the dog will prefer to be upwind from the hunter; secondly the uniformity of the distribution of positions of the hunter regardless of where the dog is.
To put it in a mathematical form the frequentist approach equates $P(D|H)$ to $P(H|D)$ ignoring $P(H)$, the prior or the distribution of the position of the hunter and ignoring that $P(D|H)$ might depend on more than just the distance between the dog and the hunter.
Another classical application of Bayesian inference is the computation of the number of false positive in a medical test: Let us suppose there is a very rare disease which occures only in a tiny fraction ϵ of the population. A test for this disease returns a false result with probability p.

$$P(\text{negative}|\text{sick}) = P(\text{positive}|\text{healthy}) = p$$
$$P(\text{positive}|\text{sick}) = P(\text{negative}|\text{healthy}) = 1 - p$$
$$P(\text{sick}) = \epsilon$$

Bayes theorem tells us that:

$$P(\text{false negative}) = P(\text{sick}|\text{negative}) = \frac{p\epsilon}{p\epsilon + (1-p)(1-\epsilon)}$$
$$P(\text{false positive}) = P(\text{healthy}|\text{positive}) = \frac{p(1-\epsilon)}{p(1-\epsilon) + (1-p)\epsilon} \,.$$

As you can see these probabilities look much different even if the accuracy of the test is the same for false positives and false negatives. What is happening? The rarity of the disease determines a very high rate of false positives, in fact it can be shown that more than half of the positives are false unless the probability p of having an inaccurate result is smaller than the prevalence of the disease ϵ.

Bayesian inference in quantitative biology

Bayesian inference has an increasingly important role in quantitative biology: the emergence of large data sets coming from molecular biology, neurosciences and molecular biology has increased the need for sophisticated mathematical techniques for their analysis.

Examples of biological systems are being successfully investigated through the use of Bayesian inference range from phylogenetics [Huelsenbeck 01], where one wants to reconstruct the most likely evolutionary tree from genetic data to gene regulatory networks where a stochastic approach has been recently shown to be very successful [Elowitz 02, Zou 05].

Moreover moving away from the molecular scale systems such as neural networks and bacterial motility have greatly benefited by such approaches.

In what follows we will concentrate on two main problems and give a brief outline of a third.

The first problem we tackled is that of spatial searches with dilute and stochastic information about the location of an object. More precisely we will turn to a strategy originally devised by Vergassola et al. [Vergassola 07b] that makes use of an informational theoretical approach for the location of an odor emitting source.

During our thesis we have developed a continuous version of the algorithm and an extensive analysis of its performances and trajectories.

The second problem we will turn to regards unzipping experiments of DNA molecules: the force-extension signal that can be measured in these experiments is strongly dependent on the DNA sequence.

At first we will describe the direct problem of reproducing experimental traces on a computer and we will describe a software package we have developed with F. Zamponi, R. Monasson and S. Cocco during our thesis, that can simulate the dynamics of such an experiment in a highly modular way.

Then we will propose several strategies for the inverse problem of reconstructing the sequence from the unzipping traces.

Lastly we have devoted a section (appendix A) to a brief technical description of an algorithm for the inference of the interaction matrix of integrate and fire neurons. This algorithm has been developed by Monasson and Cocco and our effort during our thesis has been a translation of the code to the C language, the development of an interface with Matlab and code optimization.

Part I

Infotaxis

Chapter 1

Introduction

1.1 Taxes and the biology of searching

A *taxis* is the innate directional response of the motility of an organism to a stimulus. On the other hand responses that imply a change in orientation or in the direction of growth are called *tropisms* and those which are not directional are called *kineses*.

The term taxis is most commonly found speaking of unicellular organisms, because of its automatic and innate nature, even thought it is sometimes applied to insects and crustaceans. Stereotyped responses in higher organisms are commonly thought to be less reflex-like, they are usually categorized as instincts and are the subject of study of ethology. Taxes can be distinguished according to the nature of the sensory organs implied:

Klinotaxis Different successive stimuli are measured by a single sensory organ.

Tropotaxis Well spaced sensory organs measure stimuli on different parts of the organism.

Telotaxis The perception is mediated by a single directional organ. When the motor response is at an angle to the direction of the source some sources distinguish **menotaxis**.

Taxes can also be divided according to the type of stimulus they respond to: chemotaxis (chemical gradients), phototaxis (light sources), geotaxis (gravitational fields), magnetotaxis (magnetic fields) and so on and so forth.

1.2 Chemotaxis

The type of taxis which has attracted the most interest in biology is probably chemotaxis, because of its ubiquity in unicellular organisms as inside multicellular organisms.

The first observation of bacterial motility date back to the beginnings of microscopy, but we have to wait for the end of the nineteenth century for the first observations of responses to chemical gradients.

It is important to distinguish, as we will do in the following, between bacterial and eukaryotic chemotaxis.

Bacteria are very small cells, whose size is of the order of the micrometer, below that of typical fluctuations of chemical fields: this forbids them to be directly sensitive to chemical gradients. Because of this chemosensation must happen through successive intensity

assays. According to the preceding section definitions it is a klinotaxis.

Eukaryotic cells can be much bigger than bacteria: some species can reach sizes of the order of a millimeter and typical sizes range in the tens and hundreds of micrometers. Because of this in eukaryotes chemosensation happens through the instantaneous differentiation of stimuli coming from different parts of the organisms. In this case chemotaxis can be defined as a tropotaxis.

In the light of this distinction and of the differences between motor organs in different organisms, bacterial and eukaryotic chemotaxis must be considered as different phenomena.

1.2.1 Chemotaxis in bacteria

Many reviews of bacterial chemotaxis exist in literature, for example the classic Adler's [Adler 66] or Berg's [Berg 88], which has an extensive bibliography. Here we will follow another Berg's review [Berg 75] which is more focused on theory than on bacterial physiology.

Microbiology's workhorse is certainly *Escherichia coli* (pictured in figure 1.1), partly for historical reasons, because of it's ubiquity in human guts and certainly for its simplicity. *E. coli* is endowed with about six flagella positioned on its surface. When those turn anti-clockwise they form a bundle and push the bacterium in a definite direction. Flagella can turn clockwise too: when this happens the bundle opens up and the bacteria tumbles on itself in a random fashion.

Those two modes of movement are the fundamental components o chemotactic response

Figure 1.1: A specimen of *Escherichia coli*. Notice the flagella that enable it to move, now unbundled.

in flagellates and are called *swims* in the first case and *tumbles* in the second.

Swims length is temporally limited by Brownian noise which, at room temperature for a body of size of a micrometer, decorrelates the heading of the bacteria in about ten seconds. Because of this reason bacteria tumble before losing their original heading completely.

Tumbles on the other hand are a random event which last about a tenth of a second. The new heading of after a tumble is completely independent of the one before.

Up to here the description of the motion of a flagellate does not differ significantly from a random walk; in the absence of chemical gradients the duration of swims is distributed as an exponentially random variable (that is to say that tumbles are a Poisson process).

Directional response in the motion of *E. coli* happens through the variation of the average duration of swims: if the bacteria is moving in a favorable direction swims become longer. This observation is compatible with what we have said about the klinotactic nature of

bacterial chemotaxis. Because of diffusive reasons, bacteria are not capable of discriminating between favorable and unfavorable directions during a tumble, but it is forced to sample the gradient during the swim. In other words the chemical gradient signal to noise ratio is big enough only on distances of the order of swims, not on the scale of the size of bacteria.

E. coli temporal response to gradients has been studied thanks to the response to short impulses. Bacteria effectuate time differentiation through an integral of concentration at different times multiplied to a function which has a positive weight for the first second immediately in the past and a negative weight for the three preceding seconds:

$$P(\text{tumble}) = l - k \int_{-\infty}^{0} dt\, c(t) w(t)\,, \tag{1.1}$$

where k and l are positive real constants that ensure normalization and $w(t)$ is a compact support weight function which has the characteristics we have just described and which were measured by Segall et al. in [Segall 86] (see Figure 1.2). This can be rewritten integrating by parts as:

$$P(\text{tumble}) = l + k \int_{-\infty}^{0} dt\, c'(t) W(t)\,, \tag{1.2}$$

where W is a compact support probability distribution which is zero outside the integration domain and $W'(t) = w(t)$.

The real world w has been measured by [Segall 86] and is shown in figure 1.2, the two lobes have equal area, which is consistent with our definition of W. The fact that the
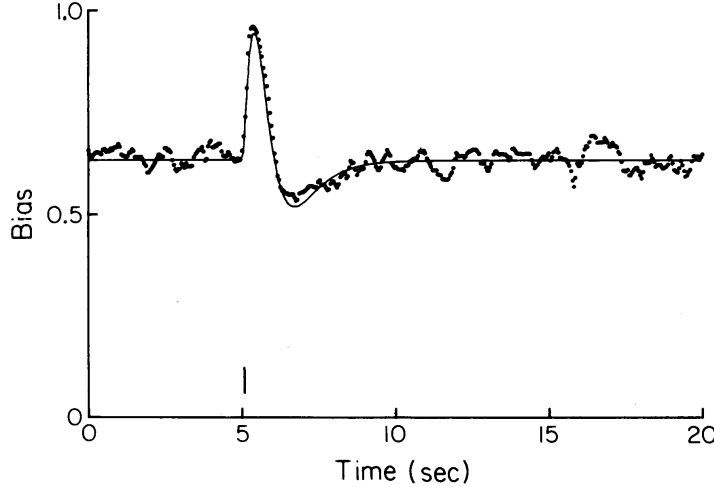
Figure 1.2: The response of bacteria to a chemoattractant in wild type *E. coli*. The dotted curve is the bias in the rate of tumbles after some attractant was pulsed at the vertical bar. From [Segall 86].

derivative is averaged over a finite period of time is a desirable property, in fact it allows bacteria to average out fluctuations in concentration fields. On the other hand run lengths never get longer than a few seconds, because bacteria aren't able to go in a straight line for long periods of time because of rotational diffusion.

1.2.2 Chemotaxis in eukaryotes

As we have previously mentioned, eukaryotes sense chemical gradients in a way which is much different from bacteria. This difference has an effect on typical trajectories of a

chemotactic eukaryote which, being able to sens gradients instantaneously and being much less affected by Brownian effects, is able to climb the chemoattractant gradient directly.

Motility in eukaryotic cells happens through ameboid movement (as in slime molds), cilia (as in *Tetrahymena*, or through the eukaryotic flagellum (as in *Chlamydomonas*), all these means of transportation are much more precise than the bacterial flagellum.

Eukaryotic chemotaxis is not confined to unicellular organisms: it plays a central role in embryogenesis, in the immune system and also the spread of metastases.

As is the case with many biological phenomena eukaryotic chemotaxis has its model organism: *Dictyostelium discoideum* (pictured in figure 1.3), a soil living amoeba which cycles through an unicellular and a multicellular state according to the environmental conditions.

When *D. discoideum* undergoes starvation, it starts secreting cyclic AMP which is a
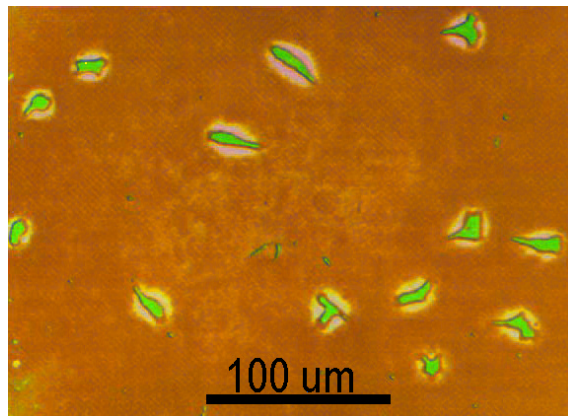
Figure 1.3: A few specimens of *Dictyostelium discoideum*.

chemoattractant, this way cells move towards one another until they stick to each other. When the cells are lumped together they form what is referred to as a pseudoplasmodium, or more colloquially a slug which measures a few millimeters. Some other slime molds can form pseudoplasmodia of sizes of square meters which are commonly found on forest floors.

D. discoideum we observed for the first time in 1933 [Raper 35], in the following years its life cycle was described in detail [Raper 40] and in the fifties cyclic AMP was identified as playing a central role in aggregation [Shaffer 53]. Nevertheless it wasn't until the beginning of the seventies that a model for aggregation was proposed [Keller 70], and despite some resistance in the microbiology community later accepted.

What was novel about this model was that aggregation was described as a truly collective phenomenon, like those found in the statistical physics of phase transitions.

1.3 Discrete infotaxis

1.3.1 Historical models

The description we have given for chemotactic cells relies heavily on the size of cells and on the nature of chemical gradients at their scale. If one wishes to model olfactory search, one has to deal with turbulence, intermittent signals and dilution of fields.

First of all most chemoattractants degrade over times and scales which are relevant over the size of a typical search, we will see that this leads to exponentially decaying concen-

trations and that this has to be taken into account.

Moreover the nature of olfactory system is such that it is impossible to instantaneously
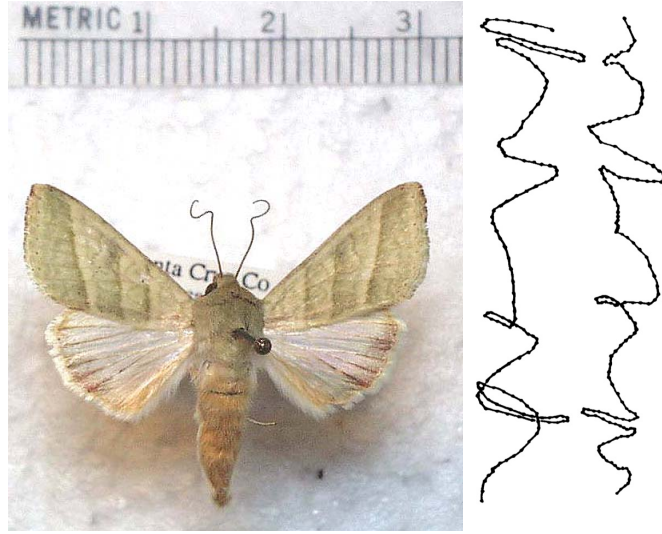
Figure 1.4: Left: a specimen of tobacco budworm (*Heliothis virescens*), a species of moth. Right: a few recorded trajectories of *H. virescens* from [Vickers 94].

perceive the spatial derivatives as in a tropotaxis: nostrils are usually very close and even if they were to be as far apart as ears or eyes the spatial information they would get would not be reliable. This is because of the effect of turbulence, local concentrations do not necessarily reflect the distance or direction of the source.

In the past there have been a few attempts to define search strategies when information is scarce: one classic reference is Gal's book on search games [Gal 80], but the amount of information in classical search games is simply too scarce for our purposes: there is no equivalent of the odor field, that is the source is found when the searcher is close enough and the searcher has no clue whether the source is close by or not unless it has been found. One further development of search strategies was given by Balkovsky and Shraiman [Balkovsky 02] who proposed a model for olfactory searches where both the searcher and the odor particles are bounded to move on the sites of a bidimensional discrete lattice. The model supposes an average wind direction, that we can take without loss of generality to be up to down. Odor particles then are made to move down at every time-step and can either move left, right or not move at all on the horizontal axis with equal probabilities. Odor particles don't decay as in more refined models, thus the odor field is never dilute when the searcher is downwind with respect to the source and close to the wind axis.

The authors observed that the stationary probability of finding an odor particle in x, y when the wind blows in y direction and one particle per time step is emitted is given by:

$$P(x,y) = \frac{1}{\sqrt{4\pi Dy}} e^{-\frac{x^2}{4\pi Dy}} \,, \tag{1.3}$$

where $D = (p_r + p_l)/2 = 1/3$ and $p_r = p_l = 1/3$ are the probabilities of moving left and right. That is, being the variance proportional to y, most odor particles will be confined to the area $x^2 < (4\pi Dy)$.

If an encounter has just been made and the searcher has no prior information on the position of the source, it follows from the Bayes' theorem that the source is most probably located in the area defined by a parabola having for vertex the position of the odor encounter. From this observation stems the strategy devised by the authors: once an odor

particle has been encountered the searcher explores exhaustively zigzagging the area where the source is most probably located until either the source is found on another particle encountered. For a clearer pictures of what a typical trajectory looks like see figure 1.5.

The main drawback of this strategy is that it is guaranteed to work only in the case
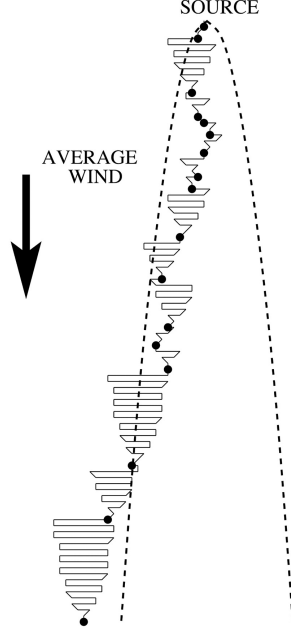
Figure 1.5: A sample trajectory of the algorithm proposed by Balkovsky and Shraiman. The continuous line is the trajectory, the dashed line the parabola that is the boundary to the area where the probability of encountering odor particles is significantly different from zero and the circles are the odor hits. From [Balkovsky 02]

of non-decaying odor particles, that is when the odor concentration does not decrease exponentially with the distance.

1.3.2 Definition of the odor detection model

Recently Vergassola et al. have proposed an algorithm for olfactory searches: here we will describe what is the odor model that underlies their search strategy using the formalism used in the Supplementary informations of their paper [Vergassola 07b].
The stationary concentration of odor particles $c(y)$ in the absence of an average wind is given by:

$$D\nabla c(y) - \frac{1}{\tau}c(y) + R\delta(y - y^*) = 0\,, \tag{1.4}$$

where D is the diffusion coefficient, that stems from molecular and turbulent diffusion, τ is the mean decay time, R is the rate of emission of odor particles and $y*$ is the position of the source.
This equation has analytic solutions and in two dimensions yields:

$$c_2(y) = \frac{R}{2\pi D}K_0\left(\frac{|y - y^*|}{\lambda}\right)\,, \tag{1.5}$$

where K_0 is the zero-order modified Bessel function of the second kind, λ is a characteristic length given by $\lambda = \sqrt{D\tau}$ and can be interpreted as the mean length traveled by an odor

particle before decaying. It will be used in the following as the natural unit of lengths. In three dimensions the solution is:

$$c_3(y) = \frac{R}{4\pi D} \frac{e^{-\frac{|y-y^*|}{\lambda}}}{|y-y^*|} \, . \tag{1.6}$$

The rate of encounter of odor particles per unit for a spherical searcher of radius a is given by relation due to Smoluchowski [Smoluchowski 17]:

$$R_3(y) = 4\pi Dac_3(y) = R\frac{a}{\lambda} \frac{\lambda e^{-\frac{|y-y^*|}{\lambda}}}{|y-y^*|} \, . \tag{1.7}$$

While in two dimensions the relation is:

$$R_2(y) = \frac{2\pi D}{\ln\left(\frac{\lambda}{a}\right)} c_2(y) = R\frac{K_0\left(\frac{|y-y^*|}{\lambda}\right)}{\ln\left(\frac{\lambda}{a}\right)} \, , \tag{1.8}$$

where R is the number of emitted particles per second.

These two equations define the natural unit of time that we will use throughout this work: in three dimensions the unit of time is $\frac{\lambda}{aR}$, while in two dimensions it is $\log\left(\frac{\lambda}{a}\right)/R$. Once the unit of time and length are defined through the actual physical constants of the system we need not worry about those details anymore: the description we will give of the system will be completely independent of them.

Once this relation is known, the idea is to model the erratic nature of odor detection in a
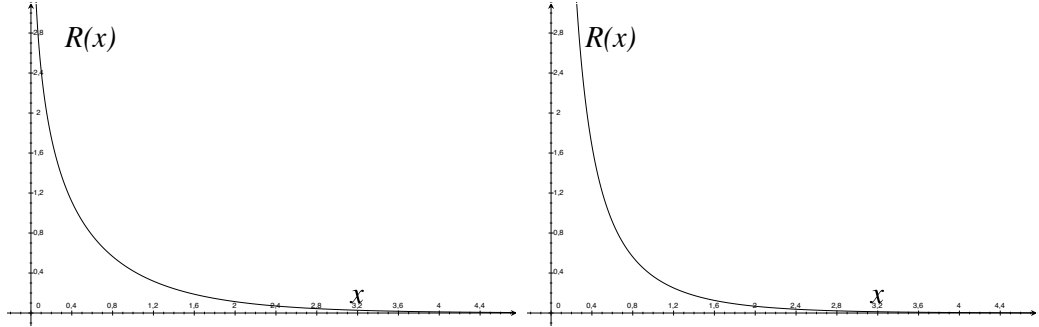
Figure 1.6: The rate of encounter of odor particles, in two (left) and three (right) dimensions. The divergence in the origin is much more abrupt in the three-dimensional case, than in the two-dimensional one, but the asymptotic behavior for large arguments is the same.

turbulent flow as a Poisson process with a rate proportional to this rate of detection. This way odor is perceived through discrete *hits* which vary in frequency as we move closer to the source. Hits contain no information pertaining the direction of the source and are all equal in intensity. The probability of getting n hits during time Δt while standing still at coordinates y is:

$$P_y(n) = \frac{(\Delta t R(y))^n}{n!} e^{-\Delta t R(y)} \, . \tag{1.9}$$

This equation allows us to write the probability of receiving a number of hits n along a trajectory at times t_i given the knowledge of the position of the source, that is:

$$P(x(t), t_i|y) = \exp\left(-\int_0^t dt' R(|y - x(t')|)\right) \prod_{i=1}^H R(|y - x(t_i)|) \, , \tag{1.10}$$

where we have supposed no two hits happen at the same time. While this is reasonable for a continuous time description, in a discrete time framework one has to divide by $n!$ whenever n hits happen during the same time-step, but we will see later this is of no importance.

1.3.3 The Bayesian posterior

Using Bayes' theorem we can write the probability of the source being at position y given the trajectory and the hits' times:

$$P_t(y|x(t), t_i) = P_0(y) \frac{\exp\left(-\int_0^t dt' R(|y - x(t')|)\right) \prod_{i=1}^H R(|y - x(t_i)|)}{\int dy' \exp\left(-\int_0^t dt' R(|y' - x(t')|)\right) \prod_{i=1}^H R(|y' - x(t_i)|)}, \qquad (1.11)$$

where P_0 is the prior distribution for the position of the source, we will see later how this plays a central role. On the other hand the attentive reader will have noticed how the previously mentioned $n!$ is cancelled out in this expression.

This expression has a few interesting features: the exponential term accounts for the vanishing probability of finding the source along the trajectory, that is: if the source was along the trajectory it would be found; it is also responsible for the low probability of points close to the trajectory. On the other hand the terms in the product are diverging and concentrate the probability around the points where most hits have occurred.

1.3.4 The expected value of the variation of entropy

The main idea behind Vergassola et al. algorithm is to exploit the Bayesian posterior as defined in the previous section to define the best movement at the next step.

This is done by defining the entropy of the posterior at a given time and by choosing the direction that maximizes its decrease: that is the direction where we expect to gain the most information on the source.

We will now compute this quantity in order to analyze the different contributions that make it up.

Even if the description given up to now is completely independent of the nature of the space where the searcher moves, be it a discrete lattice or an Euclidean space, and whether the time is discretized or continuous, we will from now on follow the description of the discrete version of the algorithm given by Vergassola et al..

Let $P_t(y)$ be the posterior probability distribution at time t. It's entropy is defined by:

$$S(P_t) = -\sum_y P_t(y) \log(P_t(y)), \qquad (1.12)$$

where the sum runs on all the lattice sites y.

If our searcher is on one of the site of the lattice, it is now possible to compute the expected variation of entropy of the posterior distribution described above, resulting from a move on one of the adjacent lattice sites x:

$$\langle S(P_{t+\Delta t}) - S(P_t) \rangle = -P_t(x)S(P_t) + (1 - P_t(x)) \left(\Delta S_{\text{norm}} + \sum_i \rho_i \Delta S_i\right), \qquad (1.13)$$

where the expected value has been taken with respect to the posterior probability distribution at time t.

Let us analyze the terms one by one:

$-P_t(x)S(P_t)$ The source is found to be in x and the entropy vanishes. The probability for this event to happen is given by the posterior $P_t(x)$ and the new value of the entropy is zero, that is the variation is $-S(P_t)$.

$(1 - P_t(x))\Delta S_{\mathbf{norm}}$ The source is not found, the probability of it being at site x is now zero and the whole probability distribution has to be normalized. It can be easily computed as:

$$
\begin{aligned}
\Delta S_{\mathrm{norm}} &= -\sum_{y \neq x} \frac{P_t(y)}{1 - P_t(x)} \log\left(\frac{P_t(y)}{1 - P_t(x)}\right) + S(P_t) \\
&= \frac{1}{1 - P_t(x)} \left(P_t(x)S(P_t) - S_{\mathrm{b}}(P_t(x))\right),
\end{aligned}
\tag{1.14}
$$

where $S_{\mathrm{b}}(p) = -p\log(p) - (1 - p)\log(1 - p)$ is the binary entropy function.

$(1 - P_t(x))\sum_i \rho_i \Delta S_i$ The source is not found, but at site x the searcher receives i hits. ρ_i is the probability of receiving i hits and ΔS_i is the corresponding entropy variation, that can be calculated remembering that:

$$
P_{t+\Delta t}^{(i)}(y) = \frac{R(y - x)^i e^{-\Delta t R(y-x)}}{\left\langle R(y - x)^i e^{-\Delta t R(y-x)} \right\rangle} P_t(y).
\tag{1.15}
$$

The main idea behind Infotaxis is to use this variation of entropy as an instantaneous potential and to move in the direction where the entropy decreases faster. With this in mind different terms play the contrasting roles of exploration and exploitation in the search. The first term is more negative when the probability of finding the source at site x is larger and can be thought as an exploitation term, where the searcher tries to move greedily where the source is more likely to be found. This term only dominates at the end of the search when the probability is well concentrated.

The last two terms favor the collection of new information, through, on one hand, the elimination of possible candidates for the source position, and, on the other, the collection of hits.

One of the most compelling features of this algorithm is that the balance between exploration and exploitation seems to be automatic, we will see in the following that this statement needs to be refined, and that one can see the algorithm as greedy on the entropy potential and that a class of more powerful algorithms can be imagined on the basis of this.

Chapter 2

Continuous infotaxis

2.1 Derivation of continuous infotaxis

We now turn to the problem of the derivation of a continuous form for infotaxis which we have done during our PhD. There are a few reasons for doing so: first of all, real organisms experience the world as continuous and a lattice based description of the world seems very artificial.

Secondly, the original algorithm poses a very realistic odor propagation model, while retaining a discrete description of the searcher and of its vision of the world. This makes the model anisotropic, in fact if we suppose the source is at a certain euclidean distance, the searcher will experience the same number of hits (on average) regardless of the direction of the source with respect to the axes of the lattice, but the direction of the source might decrease the number of steps needed to reach it of a factor of up to $\sqrt{d}$, where d is the dimension of the space.

Another inconvenient of a discrete model is that the lattice is finite and the time needed to sum over all of its sites limits what can be practically done, especially in three dimensions where only a few trajectories on a small lattice were generated [Masson 09].

A continuous description on the contrary allows the description of unbounded domains and the use of adaptive techniques to improve precision if needed.

One important thing must be stated before we begin: there is not one possible translation of infotaxis in the continuous limit, what we will do is only one of the many options.

In the following we will derive our version of continuous infotaxis in two different ways: the first is somewhat lengthy and cumbersome, but it follows closely from the discrete definition, while the second is much more compact but we think showing both might shine different lights on the problem.

The first difference between a discrete and a continuous model is the nature of the probabilistic description: from now on we have to distinguish between probabilities and probability densities which we will denote with $p_t(x)$. In order to complete the discussion of the continuous limit we have to identify three independent scales which are relevant in the spatial part of the limit which are identical in the discrete version of the algorithm. These are: the lattice spacing, the size of the source σ_{s} and the area (or volume) perceived by the searcher in a time-step σ_{p}.

To rephrase this: in the discrete version of the algorithm during one time step the searcher is able to rule out the presence of the search on one lattice site. The source size is one lattice site. Performing the continuous limit we could, in principle, leave the size of the source and of the searcher perceptions finite for a vanishing lattice spacing.

If we analyze one by one the terms of equation (1.13) we obtain:

$-P_t(x)S(P_t)$ While dealing with discrete probabilities the entropy of a sure event is zero,

on the other hand for continuous distributions the entropy of a Dirac distribution is negative and divergent. In this case if the source is found the entropy does not diverge because the source has a finite size σ_s. Therefore this term is $\sigma_\mathrm{p} p_t(x)(\log(\sigma_\mathrm{s}) - S(p_t))$

$(1 - P_t(x))\Delta S_{\mathbf{norm}}$ When the searcher moves the probability in the area σ_p around its position $x(t)$ becomes zero, thus the expected value for the variation of entropy due to the new normalization reads $p_t(x)\sigma_\mathrm{p} S(p_t) - S_\mathrm{b}(p_t(x)\sigma_\mathrm{p})$, where we have considered p_t constant in the area σ_p and where $S_\mathrm{b}(p)$ is the binary entropy, as function defined in the previous chapter.

$(1 - P_t(x))\sum_i \rho_i \Delta S_i$ For what concerns the terms depending on the expected number of hits, we will focus on none or a single hit in a time Δt, because the probability of having more is negligible when Δt is small. That is: $\rho_1 = \Delta t \langle R(y - x(t))\rangle + O(\Delta t^2)$ and $\rho_0 = 1 - \rho_1$.

Thanks to the definition of the posterior we can write down the probability density at time $t + \Delta t$ if an hit has occurred in the interval Δt as:

$$p^{(1)}_{t+\Delta t}(y) = p_t(y)\frac{R(y - x(t))}{\langle R(z - x(t))\rangle} + O(\Delta t), \tag{2.1}$$

or if it hasn't occurred:

$$\begin{aligned} p^{(0)}_{t+\Delta t}(y) &= p_t(y)\frac{1 - \Delta t R(y - x(t))}{1 - \Delta t \langle R(z - x(t))\rangle} \\ &= p_t(y)[1 + \Delta t(\langle R(z - x(t))\rangle - R(y - x(t)))] + O(\Delta t^2), \end{aligned} \tag{2.2}$$

where we have omitted the vector norms in the argument of the R and where the average is performed over the variable z. Notice that we only need the zeroth order in Δt for the term for one hit.

Omitting all dependencies, the entropy variation for no hits reads:

$$\begin{aligned} \rho_0 \Delta S_0 &= (1 - \Delta t \langle R\rangle)\left(S\left(p^{(0)}_{t+\Delta t}\right) - S(p_t)\right) \\ &= (1 - \Delta t \langle R\rangle)\left(\Delta t \langle (\langle R\rangle - R)\log(p_t)\rangle\right) \\ &= -\Delta t \langle(\langle R\rangle - R)\log(p_t)\rangle, \end{aligned} \tag{2.3}$$

while that for one hit is:

$$\begin{aligned} \rho_1 \Delta S_1 &= \Delta t \langle R\rangle\left(S\left(p^{(1)}_{t+\Delta t}\right) - S(p_t)\right) \\ &= \Delta t \langle R\rangle\left(\left\langle \frac{R}{\langle R\rangle}\log\left(p_t \frac{R}{\langle R\rangle}\right)\right\rangle + \langle \log p_t\rangle\right) \\ &= \Delta t \left\langle R\log\left(p_t \frac{R}{\langle R\rangle}\right) + \langle R\rangle \log(p_t)\right\rangle. \end{aligned} \tag{2.4}$$

Putting all the terms together one obtains:

$$\begin{aligned} \langle S(p_{t+\Delta t}) - S(p_t)\rangle &= \sigma_\mathrm{p} p_t(x)\log(\sigma_\mathrm{s}) + S_\mathrm{b}(p_t(x)\sigma_\mathrm{p}) \\ &\quad + \Delta t \left\langle R(y - x)\log\left(\frac{R(y - x)}{\langle R(z - x)\rangle}\right)\right\rangle, \end{aligned} \tag{2.5}$$

at first order in Δt.

When the size of the area observed by the searcher vanishes all the terms on the first line vanish (even if the area of the source is zero). One must also observe that in the

continuous limit the area σ_{p} must be written as $sv\Delta t$ where s is the cross section of the searcher's perception and v its speed.

On the other hand we can regard the number of received hits as a message on the position of the source. We can compute the mutual information between the random variable Y, the position of the source and the random variable N, number of hits at first order in Δt. If one remember the meaning of the rate function R, $P(N = 1|Y = y) = \Delta t R(y-x) + o(\Delta t)$ and conversely $P(N = 0|Y = y) = 1 - \Delta t R(y - x) + o(\Delta t)$, it follows that:

$$
\begin{aligned}
I(N, Y) &= \int dy\, P(y) \sum_n P(n|y) \log\left(\frac{P(n|y)}{P(n)}\right) \\
&= \left\langle \sum_n P(n|y) \log\left(\frac{P(n|y)}{\langle P(n|y)\rangle}\right)\right\rangle \\
&= \Delta t \left\langle R(y-x) \log\left(\frac{R(y-x)}{\langle R(z-x)\rangle}\right)\right\rangle + o(\Delta t),
\end{aligned}
\tag{2.6}
$$

where all the terms except for $N = 1$ are of higher order in Δt.

The main idea behind discrete infotaxis, that is: to move in the direction that minimizes the entropy of the posterior distribution, here translates into moving in the direction that maximizes the mutual information between the two variables.

One of the possible strategies to move in the direction that maximizes the gain in information, and arguably the simplest is that of forcing the searcher to obey Brownian dynamics, where the opposite of the information gain is viewed as a potential to be minimized, that is:

$$
V_t(x) = -\left\langle R(y-x) \log\left(\frac{R(y-x)}{\langle R(z-x)\rangle}\right)\right\rangle,
\tag{2.7}
$$

And for the searcher:

$$
\gamma\dot{x} = -\nabla_x V_t(x).
\tag{2.8}
$$

where γ is a friction coefficient that will be considered constant.

It can be argued that this equation cannot be considered equivalent to infotaxis, because the velocity is not constant. We have discussed this in detail in [Barbieri 11], and we will not dwell upon the details here.

It suffices to say that there is no way to impose a fixed velocity in a continuous framework: suppose for example that we choose γ as a function of the right hand side so that the velocity is equal to a constant V, we have observed that if we choose too big a V we observe long steps and a lot of backtracking. This is clearly an effect of the finite integration time-step and it is an effect that disappears in the small time-step limit, but we believe it is symptomatic of a system that chooses it's own velocity by changing the direction continuously.

2.2 Search strategy before the first hit

2.2.1 Choice of the prior

Bayesian techniques are usually very powerful, but the choice of a suitable prior can often be difficult. One can hope for the existence of a obvious choice, or that the results do not depend too much on the specifics of the prior.

The situation at hand is less clear cut: while all of our quantities have a clear probabilistic interpretation, what we ultimately want is for the algorithm to be performing well.

Vergassola et al. chose a prior proportional to the odor propagation function R which has a few desirable properties: it is normalizable, it has a possible interpretation in the

framework of our model and it does not define a new, arbitrary length scale while still concentrating most of the probability over a finite area.

Another possible choice in the discrete version of the algorithm is the uniform distribution, where every lattice site is given equal weight, even though this was not included in the original infotaxis paper, we have toyed with this prior only to obtain trajectories that go straight until they reach a distance of approximately λ from the boundary of the lattice. Unfortunately, this lattice choice does not have any equivalent in unbounded continuous space, because of this we cannot translate our results in this case.

We will now concentrate on two priors:

One-hit prior Proportional to the right R in the appropriate dimension. It has an integrable divergence at the origin, but for every τ, no matter how small $R(y)\exp(-\tau R(y))$ is finite for $y = 0$.

Exponential prior Proportional to $\exp(-y/d_0)$. Choosing $d_0 = \lambda$ we have the same asymptotic behavior for large y. This can be used to investigate how important the small scale behavior of the prior is.

In his original paper [Vergassola 07b, Vergassola 07a] Vergassola et al. proposed the first prior as a natural choice.

As suggested by the name we have chosen, we could consider the one-hit prior as the result of a search process that has started just after the searcher has received the first hit.

This of interpretation, however, poses some problems: how can we justify search trajectories that start very far from the source? If we stick to this interpretation they should be considered as very rare events.

This can be salvaged by considering only trajectories that start close enough to the source. As we will see in the following, it doesn't make much sense to employ such a sophisticated algorithm when there's effectively no information to gain.

2.2.2 Spirals

In [Vergassola 07b] Vergassola and collaborators described logarithmic spirals in discrete infotaxis, before the first hit. After observing several trajectories where the source of odor had been turned off, we have concluded [Barbieri 07] that spirals do appear in discrete infotaxis, but they are not logarithmic, but Archimedean in nature. That is the spacing between subsequent arms is constant.

In what follows we wish to characterize spirals in two dimensions and their equivalent in three dimensions for continuous infotaxis, the debate over discrete infotaxis having since been settled [Masson 09] with further simulations in hexagonal lattices.

One-hit prior

In two dimensions the searcher moves in spirals for a wide range of values of γ, as is shown in figure 2.1. When γ is too big spiral behavior breaks down.

This behavior can be explained by a very simple argument: for a large range of values of γ the searcher effectively visits a region of area proportional to the elapsed time. In a way the probability of finding the source in a given area is discounted in a given time thanks to the negative exponential term in the posterior. Once the source is not found the searcher moves elsewhere. This effect on the prior can be directly observed in figure 2.2.

This area does not depend on γ, while the linear velocity of the searcher does. For this

reason this only has an effect on the spacing of the arms. More quantitatively if b is the spacing between successive arms then what we observe is consistent with $b \sim \sqrt{\gamma}$ and $|\dot{x}| \sim 1/b$.

Spiral behavior is not observed for large γ (> 0.08), we think that this is due to the fact that the $R \log R$ kernel has a range which is proportional to λ and for large γ's we would expect arm spacing which are larger than this range. In other words the algorithm cannot be sensitive to the probability distribution at large distances.

To validate this hypothesis we have run a few simulation with a modified kernel with larger and shorter range, and we have indeed observed that this moves the spiral-breaking-down threshold in the expected direction.

In three dimensions there is no exact equivalent of a spiral: the searcher will try to stay
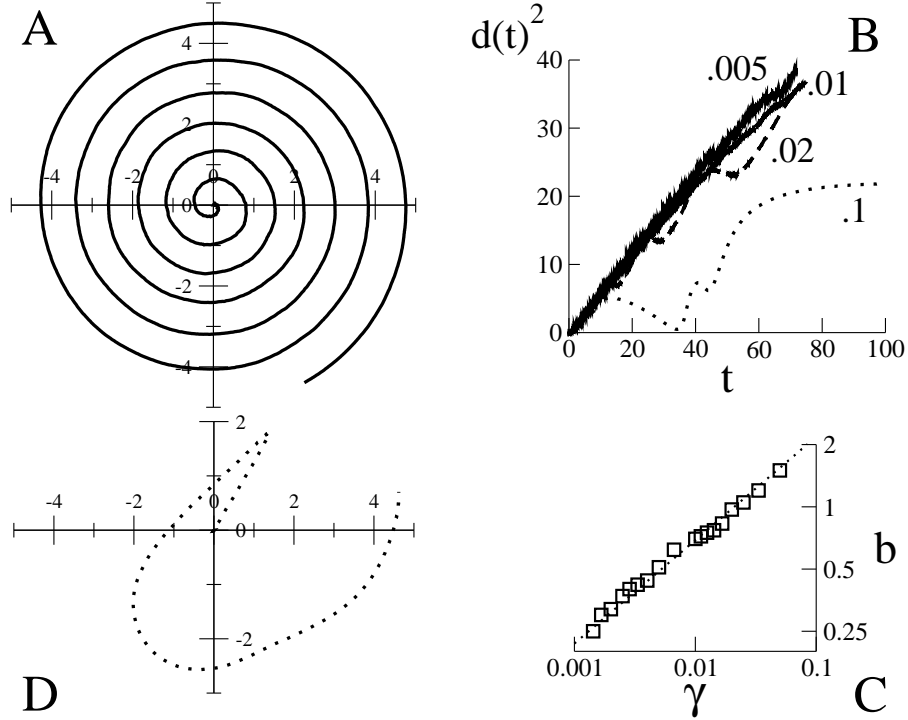
Figure 2.1: **A.** A spiral obtained for $\gamma = 0.01$. **D.** A spiral obtained for $\gamma = 0.1$. **B.** $d(t)^2$ as a function of time. As this quantity is proportional to the area explored in a given time, we show here that this is proportional to the elapsed time for several values of γ. For large γ this behavior breaks down and the searcher eventually halts. The trajectories for $\gamma = 0.01, 0.1$ correspond to panels **A** and **D**. **C.** Many values of the spacing b between spiral arms, as a function of γ. The dotted line corresponds to a slope of $-1/2$.

as close as possible to where it started as a result of the exponentially decreasing prior, but will move in a self avoiding trajectory, because of the term $\exp\left(-\int^t dt' R(y - x(t'))\right)$ in the posterior probability.

We have observed the first part of the trajectory to be quasi-planar and then to break off and start occupying all available space, this is shown in figure 2.3 where the dependence of the distance from the origin is plotted as a function of time and compared with the curve $t^{1/3}$ which corresponds to the prediction of space filling trajectories.

Three dimensional trajectories look like balls of yarn, compact coiled structures. We think that parallels can be drawn with the solutions of the Thomson problem for polyelectrolytes [Angelescu 08, Cerdà 05, Slosar 06], which has received a lot of attention recently because
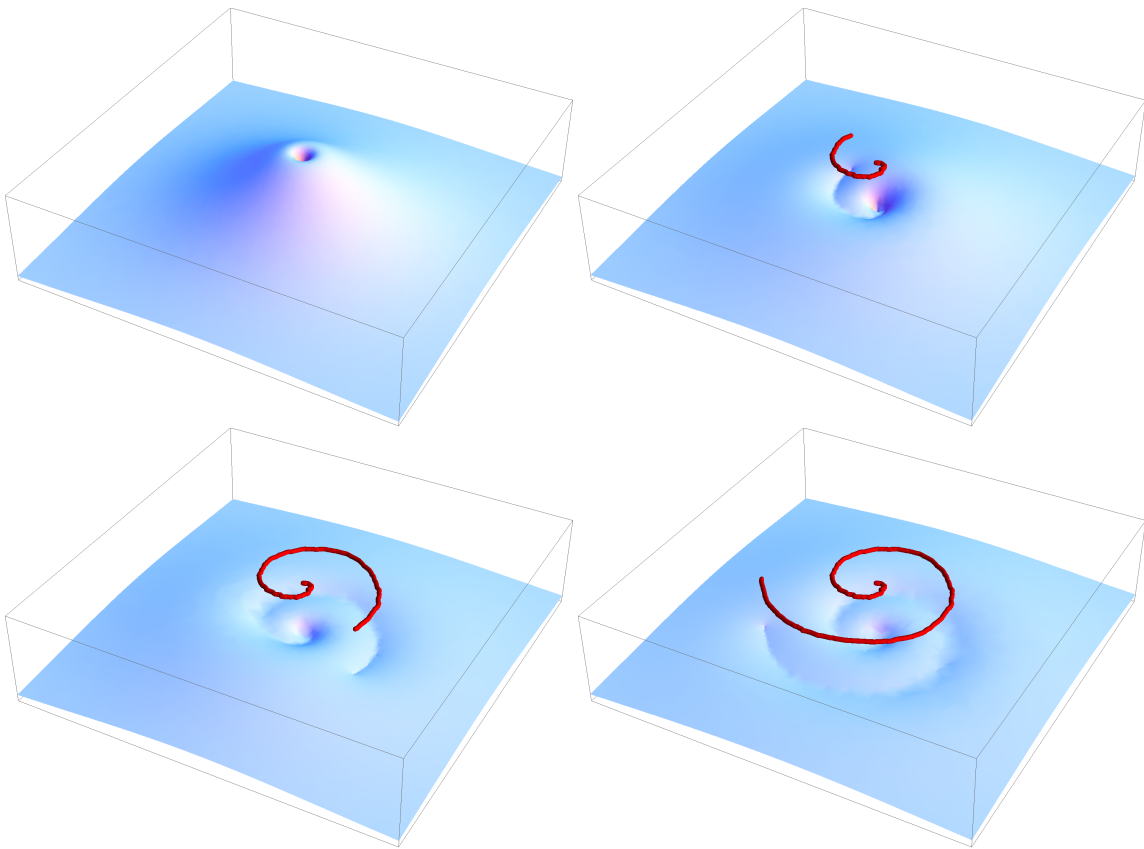
Figure 2.2: The effect of a spiraling trajectory on the probability distributiion at different times.

2.2. SEARCH STRATEGY BEFORE THE FIRST HIT
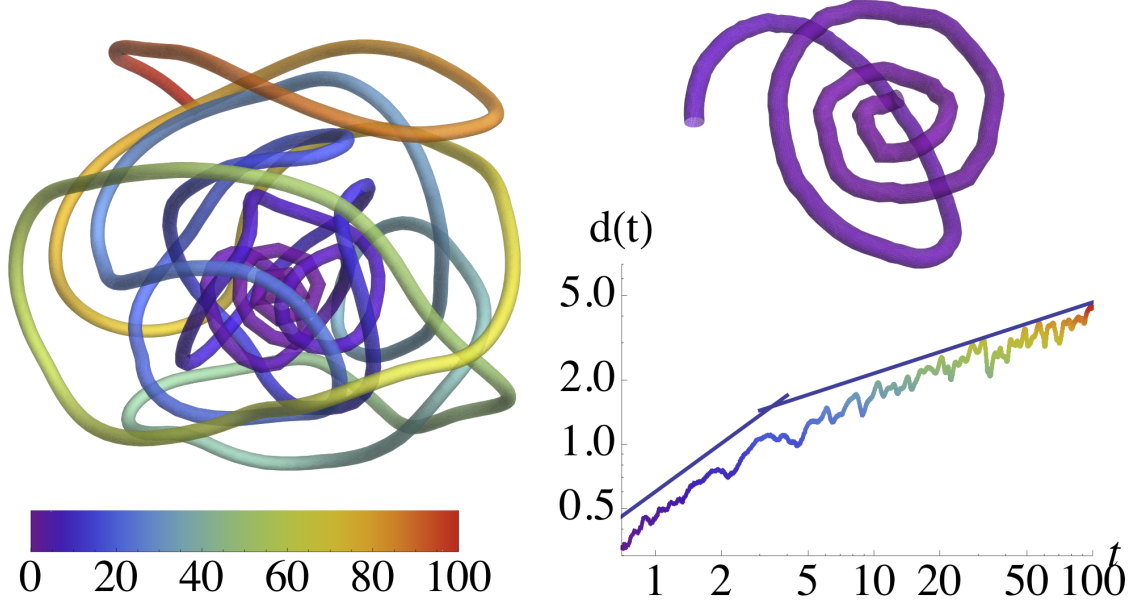
Figure 2.3: A three-dimensional trajectory in the absence of hit for $\gamma = .01$ (left), with its quasi two-dimensional initial portion (top); the time axis is color coded. Bottom: distance to the origin, $d(t)$, compared to the power laws $t^{.75}$, then $t^{1/3}$.

of its connections to the problem of DNA packing in virus capsides.

Exponential prior

Another way of interpreting the choice of the one-hit prior is to consider the details of the prior at short range from the starting point of the searcher as mostly irrelevant and to concentrate on the asymptotic behavior.

Ignoring small scale behavior makes a lot of sense in the case of discrete infotaxis, where the scales smaller than the lattice spacing are not accessible, and the probability at the starting point of the searcher is exactly zero regardless of the prior.

The exponential prior can be also justified because of its memorylessness property that is: $P(Y > y + d | Y > y) = P(Y > d)$ and furthermore because it is maximum entropy distribution with a fixed mean.

This two mathematical properties could be used to justify the Archimedean nature of the spirals, which can be checked in figure 2.4. The spirals however break down, as discussed before for the case of variable γ, when the arm spacing b would exceed the range of the kernel $R \log R$.

The fact that the behavior of the searcher for both the one-hit prior and the exponential prior produces spirals, suggests that the spirals are a consequence of the asymptotic behavior of the prior at large distances. We will try to verify this with a Taylor series expansion of the right hand side of the equation for the movement of the searcher.
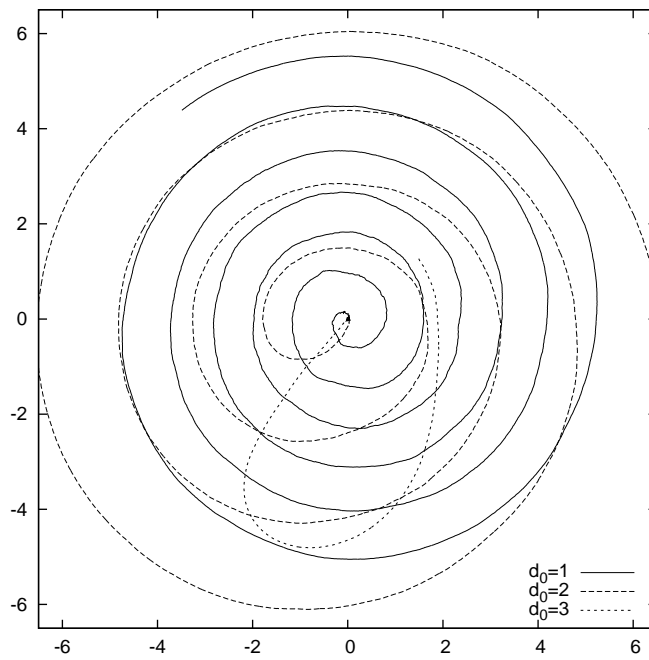
Figure 2.4: Three trajectories without hits for the exponential prior with varying d_0. As highlighted in the text, for low enough d_0 spirals are observed with spacing $b \propto d_0$. When d_0 is too large, as we have observed for varying γ, spirals break down. Such is the case for $d_0 = 3$.

2.2.3 Small x expansion

It is possible to characterize the spirals as an instability by performing an expansion for small x of equation 2.8:

$$
\begin{aligned}
\gamma \dot{\vec{x}}(t) = & \alpha_1(t)\vec{x}(t) + \alpha_2(t)\int_0^t dt'\,\vec{x}(t') \\
& + \int_0^t dt'\,\vec{x}(t')\left[\beta_1(t)|\vec{x}(t')|^2 + \beta_2(t)\vec{x}(t')\cdot\int_0^t dt''\,\vec{x}(t'')\right] \\
& + \int_0^t dt'\,\vec{x}(t')\left[\beta_3(t)\int_0^t dt''|\vec{x}(t'')|^2 + \beta_4(t)\left(\int_0^t dt''\vec{x}(t'')\right)^2\right] \\
& + o(x(t)^3)\,,
\end{aligned}
\tag{2.9}
$$

where the $\alpha_i(t)$ and $\beta_j(t)$ are time dependent coefficients, respectively for the first and third degree. All other terms vanish for symmetry reasons. We need to stress that this expansion is only valid for three dimensions.
Defining:

$$
\langle f(y)\rangle_t = \frac{\int d\vec{y}\,\exp(-tR(y))\,f(y)}{\int d\vec{y}\,\exp(-tR(y))}\,,
\tag{2.10}
$$

we can express the terms of the development as:

$$
\alpha_1(t) = \frac{1}{6}\left\langle \frac{(R'(y))^2}{R}(y) - \left(R''(y) + 2\frac{R'(y)}{y}\right)\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t
\tag{2.11}
$$

$$
\alpha_2(t) = \frac{1}{3}\left\langle \frac{(R'(y))^2}{R}(y)\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t
\tag{2.12}
$$

$$
\beta_1(t) = \frac{1}{3!5}\left\langle \left(R'''(y) + 2\frac{R''(y)}{y} + \frac{R'(y)}{y^2}\right)\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t
\tag{2.13}
$$

$$
\beta_2(t) = \frac{1}{15}\left\langle (R'(y))^2\left(R''(y) + \frac{2}{3}\frac{R'(y)}{y}\right)\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t
\tag{2.14}
$$

$$
\begin{aligned}
\beta_3 = & \frac{1}{30}\left\langle (R'(y))^2\left(R''(y) + \frac{2}{3}\frac{R'(y)}{y}\right)\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t \\
& - \frac{1}{18}\left\langle (R'(y))^2\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t\left\langle R''(y) + 2\frac{R'(y)}{y}\right\rangle_t \\
& - \frac{1}{18}\left\langle (R'(y))^2\right\rangle_t\left\langle R(y)\left(R''(y) + 2\frac{R'(y)}{y}\right)\right\rangle_t \\
& + \frac{1}{18}\frac{\left\langle R(y)\left(R''(y) + 2\frac{R'(y)}{y}\right)\right\rangle_t^2}{\langle R(y)\rangle_t}
\end{aligned}
\tag{2.15}
$$

$$
\begin{aligned}
\beta_4 = & \frac{1}{18}\frac{\left\langle R(y)(R'(y))^2\right\rangle_t^2}{\langle R(y)\rangle_t} - \frac{1}{18}\left\langle (R'(y))^2\right\rangle_t\left\langle R(y)(R'(y))^2\right\rangle_t \\
& - \frac{1}{18}\left\langle (R'(y))^2\right\rangle_t\left\langle (R'(y))^2\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t + \frac{1}{30}\left\langle (R'(y))^4\log\left(\frac{R(y)}{\langle R(y)\rangle_t}\right)\right\rangle_t
\end{aligned}
\tag{2.16}
$$

If one looks at the equation up to the first order, neglecting the β terms, one can already explain the instability that leads to spirals.
Since $\alpha_1 \simeq \frac{\sqrt{2}}{3e}\frac{\log t}{t} > 0$ and $\alpha_2 \simeq -\frac{3\sqrt{3}}{e^2}\frac{\log t}{t^2} < 0$ for large t. α_1 is positive so the trajectory starts as a straight line out of the origin, but then the term α_2 which is unstable makes it

unstable against local bending explaining planar spirals.

An analytic solution of this simplified equation is possible if one approximates the coefficients neglecting the logarithmic terms.

β_3 and β_4 are coefficients to terms that lie in the same plane as the first order ones. Because of this we will only concentrate on β_1 and β_2. Those are both positive and lead to the instability of the planar trajectory eventually leading to a full fledged three dimensional structure.

2.2.4 Waiting time

One interesting feature of the spirals is that they do not start immediately as in the discrete algorithm. This seems to be at odds with the results obtained in previous section: α_1 is always positive, this means that staying in the origin without moving should be unstable. How to reconcile this apparent paradox?

Let us define:

$$\tilde{V}_\tau(x) = -\left\langle R(y-x)\log\left(\frac{R(y-x)}{\langle R(z-x)\rangle}\right)\right\rangle_\tau, \tag{2.17}$$

where we have sticked to the definition of the brackets of equation (2.10) in the previous section.

If we plot $\tilde{V}_\tau(x)$ along a direction for different values of τ and we compute its minimum, as in figure 2.5 we find indeed that there always a maximum in $x = 0$, but there is also a non-trivial minimum for every $\tau > 0$, albeit this minimum can be very close to the origin for small τ.

The curve of the minumum $x_m(\tau) = \arg\min_x \tilde{V}_\tau(x)$ is well fit by as $x_m(\tau) \simeq 6.62\exp(-2.32/\tau)$ in two dimensions.

These results can be further substantiated by convolving the R with a Gaussian distribution of width σ, this way a σ-dependent crossover in τ can be shown to exist between waiting and moving. The interpretation of the Gaussian convolution is that, because of the numerical integration, when the searcher is waiting it effectively fluctuates around the origin.

All this can be summarized by saying that, if the noise is zero or stays within an acceptable range, the time it takes the searcher to move perceptibly out of the origin is $\simeq 0.4$.

This is a very important feature of the continuous version of infotaxis which is not present in its discrete counterpart. This is due to the fact that setting a whole lattice site probability to zero creates a very strong repulsive effect, and since the area that is set to zero in the continuous version is infinitesimal there is no inhibition of this effect.

The striking feature of this effect is that it reproduces itself whenever there is a new hit: the searcher stops, waits about 0.4 and then starts moving again. We can think of it as if it were trying to exclude that the source was in its immediate vicinity.

There exists a distance from the source when the expected arrival time of two successive hits is smaller than the waiting time, when this happens the searcher will be effectively stuck at this position. We will call this distance d_{halt}: it is dimension dependent. It is $\simeq 0.1$ for D=2 and $\simeq 0.3$ for D=3.

d_{halt} is more rigorously defined as $0.4R(d_{\text{halt}}) = 1$. The reason for different values for different dimensions is the different form of R.
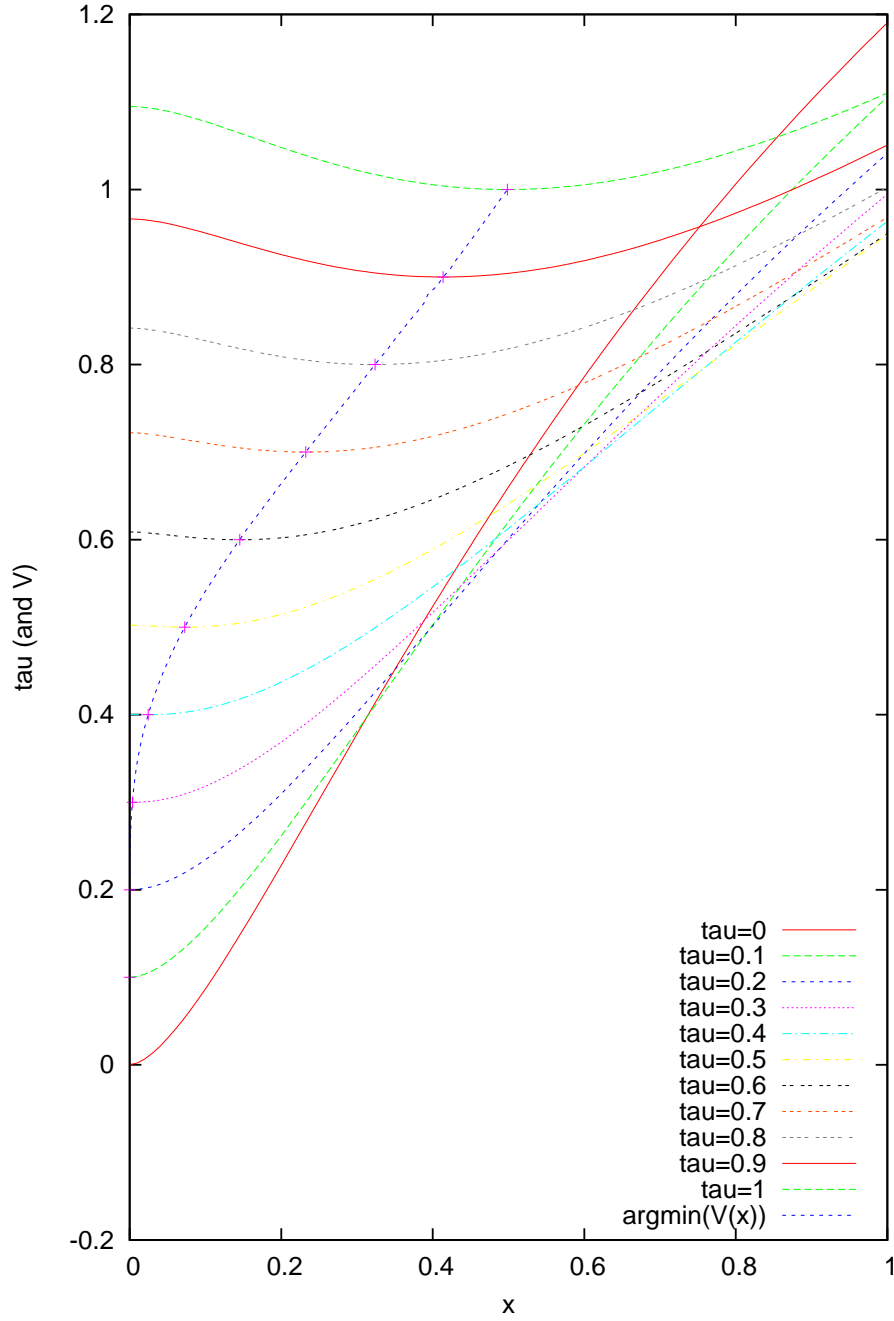
Figure 2.5: Profile of $V(x)$ at varying τ and position of its minimum. We have subtracted arbitrary constants to the various $V(x)$ in order for the curve of the minima to pass through the minima.

2.3 Numerical integration

In this chapter we will illustrate the techniques we have employed for numerically integrating the continuous infotaxis equations. We will devote some time to justifying the choice of a technique that increases the complexity of the algorithm in favor of precision. At every time-step we have to compute the integral of the kernel over the probability measure in order to know the velocity of the searcher. The position of the searcher is then updated with a simple Euler integration step, that is:

$$x(t + \Delta t) = x(t) + \Delta t v(t) \,, \tag{2.18}$$

where $v(t)$ is the velocity at time t defined as $-\nabla_x V_t(x)/\gamma$.

We have found empirically that a good choice for the integration time-step $\Delta t = \gamma$, this choice ensures precision when γ is small and then the searcher is fast and economy when γ is big and the searcher is slow.

At each time step a Poisson pseudo-random variable is generated for the number of hits, this is recorded in a vector as is the whole trajectory.

The whole procedure can be summarized in pseudo-code as:

```
searcher=origin
source=d_0/sqrt(dimension)
i=0
while(d_success<distance(source,searcher)<d_fail){
   old_n_hits=n_hits;
   n_hits+=poissonrandom(dt*R(distance(source,searcher)))
   for(j=old_n_hits;j<n_hits;j++)
       hits[j]=searcher
   force=average(force,R,R_prime,x,trajectory,history,hits)
   x+=force*dt/gamma
   trajectory[i]=searcher
   i++
}
```

An important detail that can't be omitted is the calculation of the averages over the probability distribution. The original discrete infotaxis implementation performed this by storing and updating the complete probability distribution over the lattice. This is clearly impossible in the absence of a lattice. Especially since the search is performed in unbounded Euclidean space.

We have, however, tried memorizing the probability distribution at points either on a non-square lattice or randomly picked in order to emulate the behavior of the original algorithm.

This approach is plagued by various serious shortcomings: first of all we need to choose the points at the beginning of the search, and it is natural to choose them concentrated around the starting position. After a certain time, however, the searcher will have moved farther away where the points are rarer and numerical precision will start suffering.

Another big problem is that the computation of integrals as sums over a set of point that does not change will effectively recreate a lattice, albeit not a regular one. The trajectories will stick to those *lattice* points because visiting them directly is optimal for the information gain.

In order to avoid these artifacts, that crippled the simulation even for relatively short run times, we have decided not to store and update the whole probability distribution, but to store the trajectory and the hits and to calculate the probability distribution dynamically

at each time-step. It is now possible to perform the integrals by Montecarlo importance sampling around the position of the searcher, and choose a different set of points at each time-step.

The procedure is as follows: one performs a change of variable for the argument $u = |\vec{x}(t) - \vec{y}|$ of the functions to integrate. $u = \phi(v) = u_0(1-v)/v$, where $v \in (0, 1]$, then angles are sampled uniformly in two or three dimensions.

N_{MC} points are sampled this way (typically 10^4) for each time step, and summed taking care of the Jacobian of the change of variables $v \mapsto u$.

Again in pseudo-code:

```
function average(functional,R,R_prime,x,trajectory,history,hits){
    sum=0
    for (i=0; i<MC_steps; i++) {
        y.angle=randomangle()
        y.radius=phi(randomreal())
        jacobian=phi_primep(inverse_phi(point(radius))
        hitscontrib=1
        for(j=0;j<size(hits);j++)
            hitscontrib*=R(distance(y,hits[j]))
         if(dimension==3) jacobian*=rs*rs
         else jacobian*=rs
         sum+=jacobian*priorprob(y)*exp(-history(R,y,trajectory))
            *functional(y,x,R,Rp)
    }
    return sum/MC_steps
}
```

The only bit left is the computation of the integral over the trajectory at the exponential:

$$\int_0^t dt' R(y - x(t')) \, . \tag{2.19}$$

To compute this we have used the classic composite Simpson's rule:

$$\int_o^t f(t')dt' \approx \frac{\Delta t}{3}\left[f(0) + 2\sum_{j=1}^{n/2-1} f(2j\Delta t) + 4\sum_{j=1}^{n/2} f((2j-1)\Delta t)) + f(t)\right] \, , \tag{2.20}$$

where $n = t/\Delta t$ needs to be even.

Taking extra care to ensure n is even, we get in pseudo-code:

```
function history(R,x,trajectory){
    sum=0
    if (size(trajectory)==1)
        return dt*R(distance(trajectory[0],x))
    if (size(trajectory)==2)
        return dt*(R(distance(trajectory[0],x))+R(distance(trajectory[1],x)))
    flag=size(trajectory)%2
    sum=R(distance(trajectory[flag],x))+R(distance(trajectory[size-1],x))
    for (i=1; i<=size/2-1; i++)
        sum+=2*R(distance(trajectory[2*i+flag],x))
    for (i=1; i<=size/2; i++)
        sum+=4*R(distance(trajectory[2*i-1+flag],x))
```

```
    sum*=dt/3;
    return sum;
}
```

2.4 Results and performances

2.4.1 Typical trajectories

In this section we wish to show what the typical trajectories of continuous infotaxis look like in two and three dimensions once we have introduced a source of odor, as the goal for the searcher.

In two dimensions one can superimpose the trajectory to the probability and gain some good insights as to how the posterior probability is affected by odor hits.

In figure 2.6 one can see the searcher starts its trajectory spiraling around its starting position, and how the probability distribution is affected by this: the maximum of the probability is always in front of the searcher, and a valley of minima is dug where it has passed.

In the third panel (bottom left) the first hit is received and the probability has a new maximum. If the searcher didn't receive further hits in the last panel (bottom right) it would start spiraling around the position of the new maximum.

In the last panel the probability distribution is very peaked around the real position of the source, which is about to be found.

In figure 2.7 two trajectories are shown for two-dimensional infotaxis: the one on the left is successful in finding the source while the second is not.

Notice how the unsuccessful searcher has received a very misleading hit, actually farther away from the source than when it started. We can imagine the probability distribution to be peaked somewhere closer to the position of the hit. This maximum becomes the center of its new spiraling, albeit these new spirals are not as regular as the ones we have observed without hits.

Trajectories with hits are much harder to visualize, we try to do so in figure 2.8, but the trajectory covers itself. What can be gleaned from these two trajectories is that the searcher seems to be using less information than in two-dimensional searches. In fact the unsuccessful searcher receives no hit at all while the successful one received only two.

2.4.2 Average signal

We now wish to define what we think will be a very useful tool for the evaluation of performances: as we will see in the following, a large number of runs are needed in order to sample the probability of success and the time of success. This is due to the fact that the arrival times and positions of hits can vary wildly, and have a very strong influence on the searcher trajectory.

If one observes the posterior probability density, one notices that the hits are encoded as the product of R functions centered at the position of each hit. As it is customary with multiplicative processes it is natural to look at the logarithm of the probability distribution.

$$\log P_t(y) = -\int_0^t dt' R(y - x(t')) + \sum_{i=1}^{H} \log R(y - x(t_i)) + \text{const}\,, \qquad (2.21)$$

where t_i are the times at which the hits occur.

If the searcher is at time t in position t the probability it will get a hit in the next Δt is
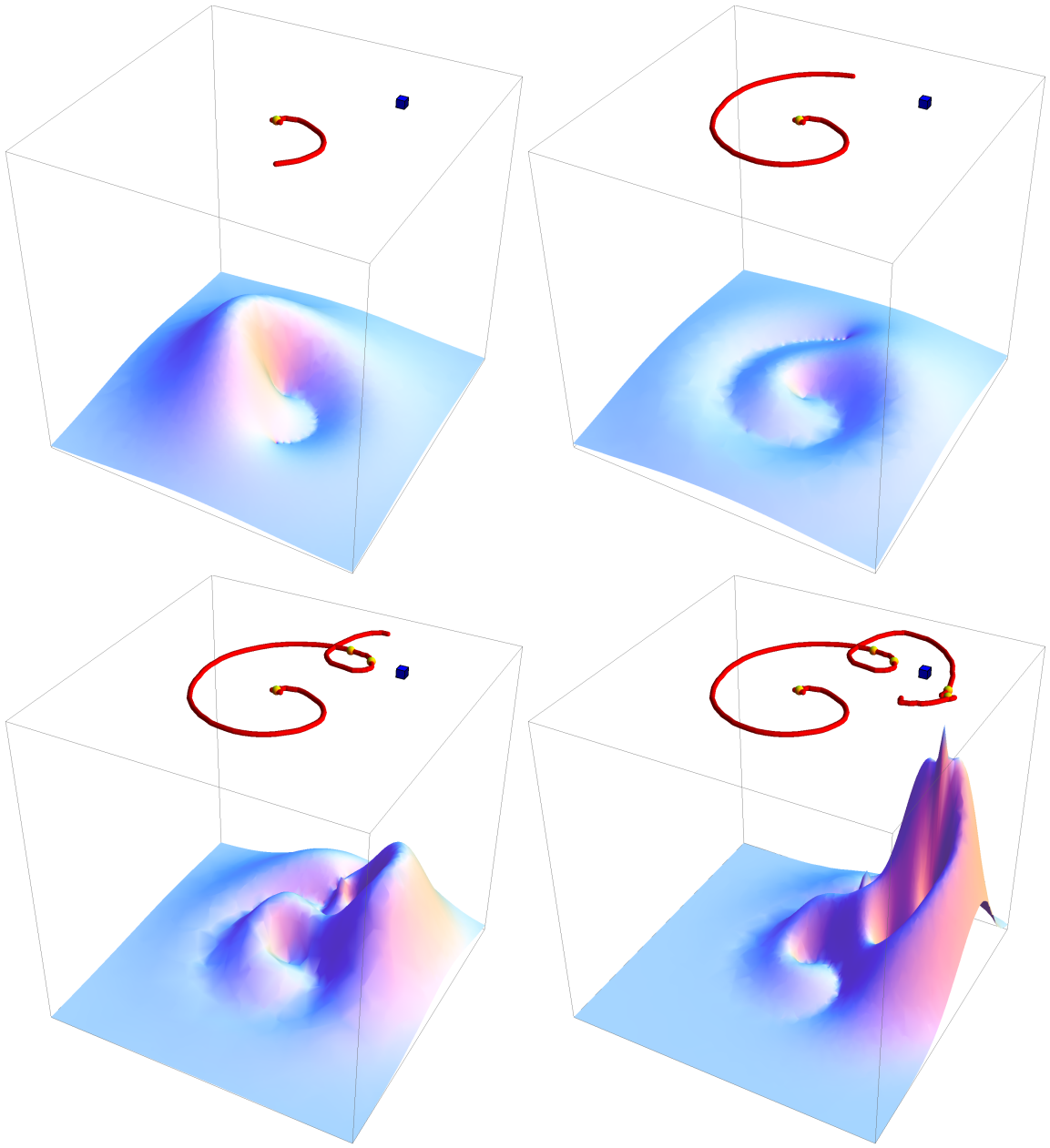
Figure 2.6: The trajectory is plotted in red, superposed to the posterior probability distribution. Along the trajectory hits are displayed as yellow spheres. The source is the blue cube.
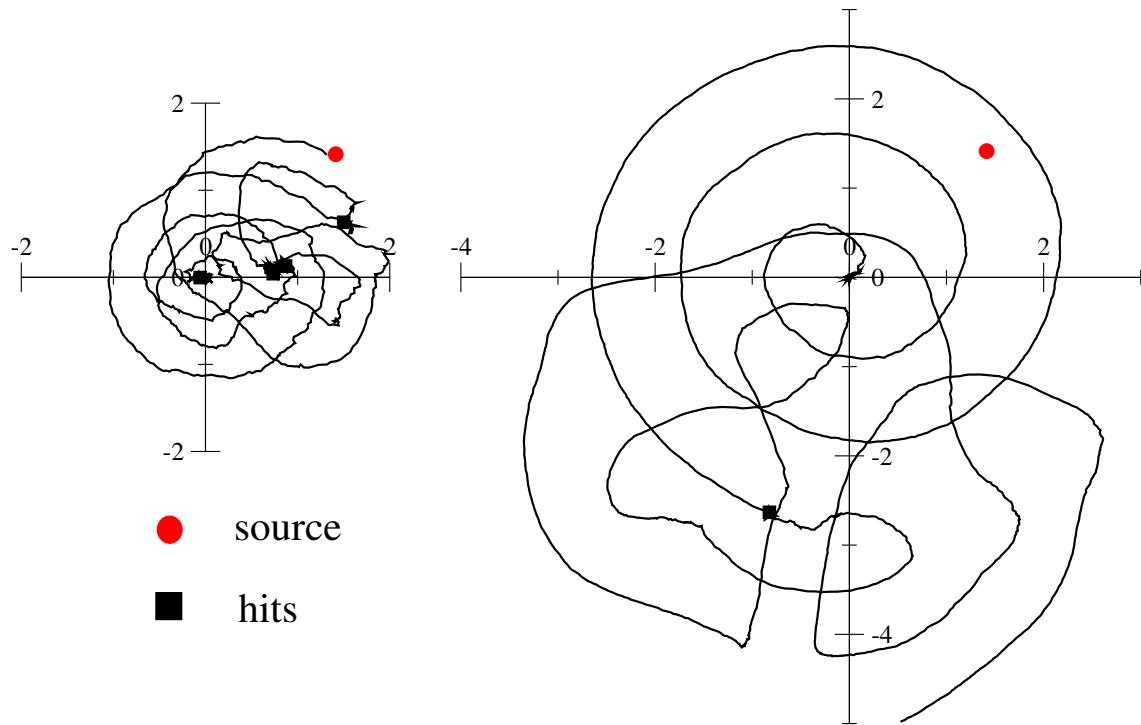
Figure 2.7: Examples of search trajectories with hits two dimensions ($\gamma = .02$). The trajectory on the left finds the source, while the one on the right is not successful. The initial distance to the source is $d_0 = 2$. The red disk represents points at distance $< d_{\mathrm{halt}}$ to the source. Black squares locate the hits.
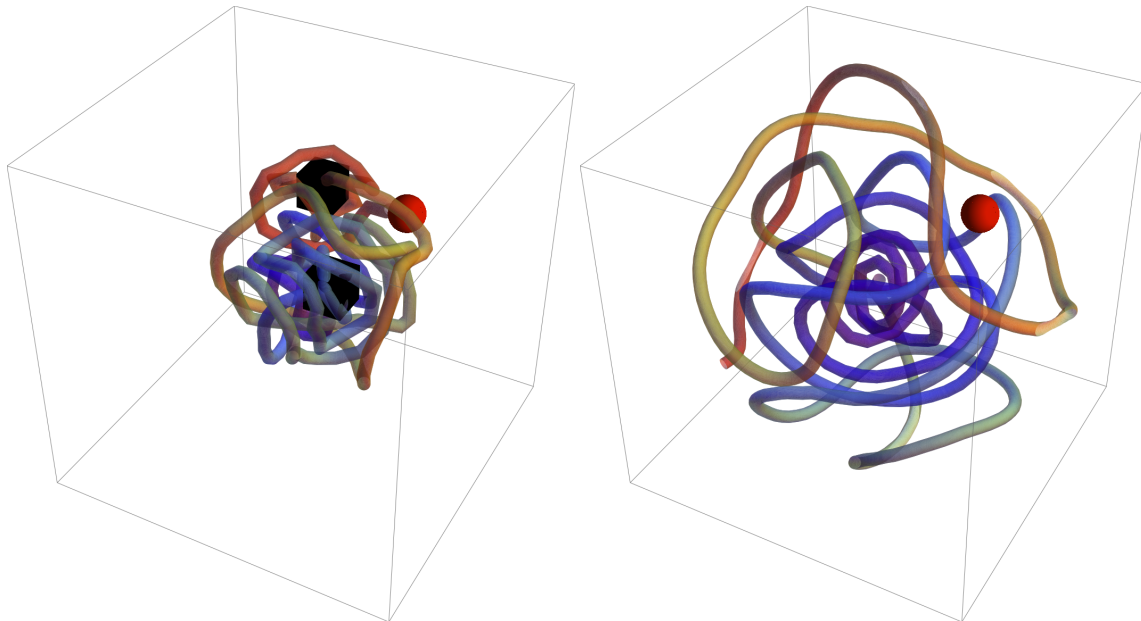
Figure 2.8: Examples of search trajectories with hits three dimensions ($\gamma = .01$). The trajectory on the left finds the source, while the one on the right is not successful. The initial distance to the source is $d_0 = 2$. The red sphere represents points at distance $< d_{\mathrm{halt}}$ to the source. Black cubes locate the hits.

given by $\Delta t R(y^* - x(t))$ where y^* is the actual position of the source. Having observed this, we can take the expected value of equation (2.21) with respect to the probability of receiving a hit at each time-step.
This yields:

$$\overline{\log P_t(y)} = - \int_0^t dt' \left[R(y - x(t')) + R(y^* - x(t')) \log R(y - x(t')) \right] + \text{const},. \qquad (2.22)$$

If we now use the exponential of this newly defined quantity as the probability distribution that moves the searcher we obtain trajectories that have features that resemble closely those of trajectories with truly random hits.
However, even if we have reduced greatly the variability among trajectories, numerical trajectories obtained for this *average* signal are not completely deterministic. This is due to the stochastic errors involved in Montecarlo integration and how those play an important role in the initial breaking of rotational symmetry.
In other words, the searcher starts in a random direction which defines the phase of the turnings of the spiral. This random direction is not a feature of the Poisson noise of the hits, but of the noise coming from Montecarlo integration. If we had access to a perfect integrator, we would need to add noise artificially at least at an initial stage to start the search.
In figure 2.9 we compare a trajectory with random hits to a trajectory obtained with the *average* signal when those have comparable duration. We also plot the entropy of the posterior distribution. Notice how it plunges in discontinuous jumps for the random signal and how it tapers off gently for the average signal.

2.4.3 Performances

In order to evaluate the performance of the algorithm we have to look at the success probability and the time needed to reach the source of odor in case of a success. But first of all we have to give a clear definition of success and failure. This is at odds with the discrete algorithm, where success was obtained when the searcher and the source were at the same position and failure when the searcher wandered out of the lattice.
In a continuous, unbounded space these definitions do not apply. However we can define a radius $d_{\text{fail}} \gg 1$ from the source that defines the region of space out of which the search has not much hope of ever succeeding. The bigger d_{fail}, the less our results will depend on it.
The definition of a d_{success} is a bit more delicate since too small a radius would have catastrophic effects because of the pinning phenomenon we have described in the previous section; too big a radius would mean getting a lot of false positives and overestimating the performance of the algorithm. In the end we settled for $d_{\text{success}} = d_{\text{halt}}$.
There are two parameters that need to be varied in order to evaluate performance: one is the distance from the source, the other is γ that characterizes the dynamics.
Another delicate issue is the definition of time: since our algorithm has a complexity per time-step which is linear in the elapsed time, CPU time will not be proportional to simulation time and we would need to optimize one or the other in different scenarios.
We have investigated the success probability for different values of the initial distance between the searcher and the source.
We have chosen distances between 1 and 3 in units of λ, because, on one hand, larger initial distances would correspond to vanishing an exponentially vanishing probability of receiving one hit and would only lengthen the spirals without showing any interesting feature of the algorithm.
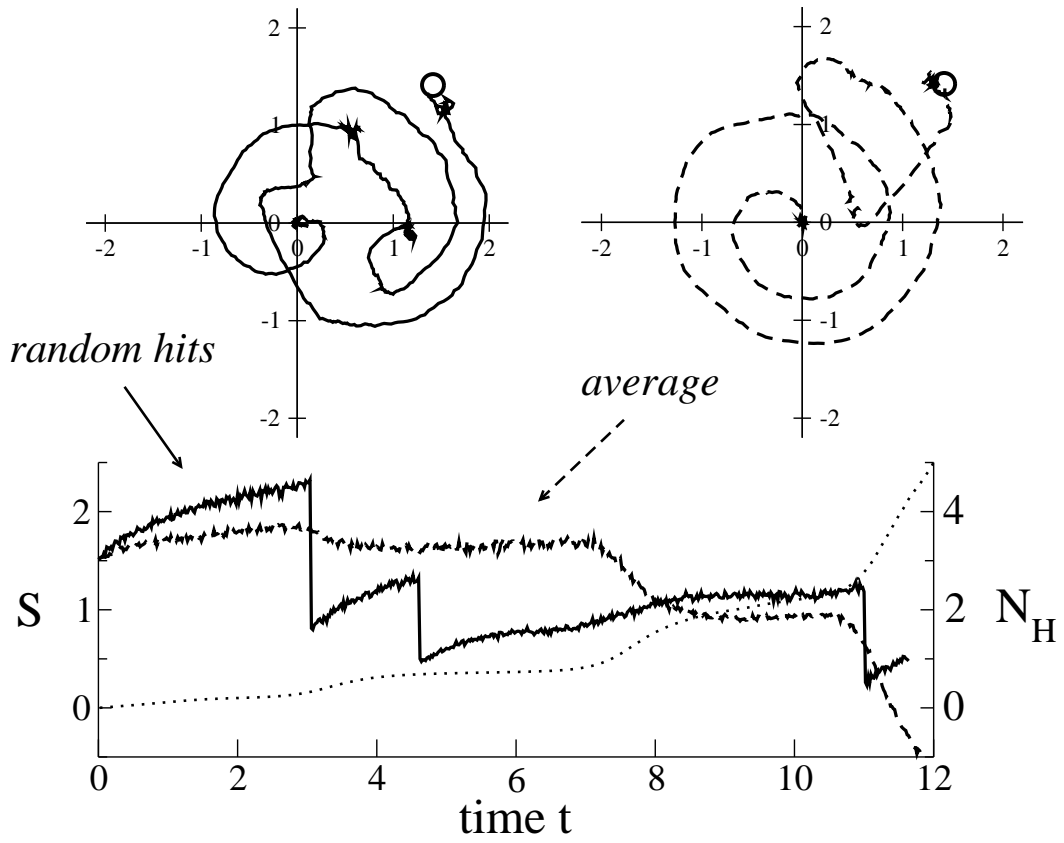
Figure 2.9: Entropy $S(t)$ (bottom, left scale) for one trajectory $\mathbf{x}(t)$ obtained with random hits (top left, full curve, 3 hits are received) and the average trajectory (top right, dashed curve). The dotted line shows the average number of hits N_H (right scale) received along the average trajectory. The source is located in $(\sqrt{2}, \sqrt{2})$ (circle).
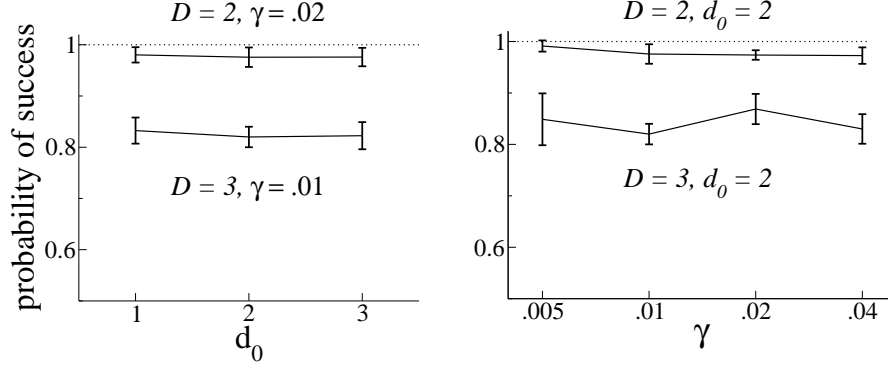
Figure 2.10: Probability of success of Infotaxis as a function of the initial distance to the source, d_0 (left), and of the friction γ (right). Top points correspond to $D = 2$ dimensions, bottom points to $D = 3$. The numbers of runs is of about 200 for each point. All probabilities where obtained with $d_{\text{fail}} = 8$.

On the other hand distances smaller than 1 are too close with the halting distance especially in three dimensions. Because of these two arguments we believe this is the only region where the behavior of this algorithm might be non-trivial.

Another important parameter is the friction coefficient γ. Overall we have observed that the success probability is affected by neither the starting distance or the friction coefficient. It is compatible with unity in two dimensions and slightly higher than 80% in three dimensions. The results are detailed in figure 2.10.

This does not surprise us much: searches are easier in two dimensions, where random walks are space filling. The result in three dimensions looks promising and it is much better than any random estimate. The interested reader can refer to the classic reference by Redner [Redner 01] for a computation of the probabilities for the associated random phenomena.

Let us now define the relevant quantities for the search time: first of all we will restrict ourselves to the successful cases. We define the success time t_{s} as the time when the algorithm halts because the searcher has entered the disk of radius d_{success}.

The CPU time can be defined in a implementation-agnostic form as $(t_{\text{s}}/\Delta t)^2$ since it will be generally proportional to this quantity. It should be noted that in the current implementation, with 10^4 Monte Carlo sampling points in spatial integrations and on a 2.4 GHz core of an Intel Core 2, $A \simeq 3$ ms.

In figure 2.11 we show, with the t_{s}'s and the CPU times for different γ's, an histogram comparing the results obtained with the average equation of the previous section with those obtained with the non-simplified equation.

It is interesting to note how if one takes into account only the t_{s} the algorithm is most efficient at low γ, however, since lower γ call for lower Δt in CPU time the algorithm is much faster for high γ.

This can be explained by remembering the dependence of the spiral spacing on γ: low γ means tighter spirals and a searcher that moves much faster linearly: while this behavior turns out to be more effective at exploring the space it is more computationally intensive because the increased scalar velocity calls for a smaller time-step.

Overall we think performance can be greatly increased either by reducing the number of Monte Carlo integration points or by reducing the number of points in the time integral. A reduction of the time points stored in memory can be obtained in two ways: the first is to add a finite memory, but if one is not careful one could end up with the searcher very
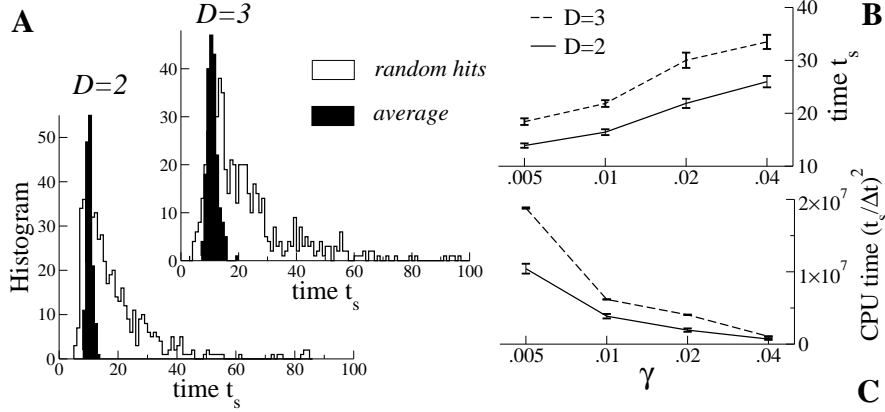
Figure 2.11: **A.** histograms of the search times t_s in $D = 2$ ($\gamma = .02$) and $D = 3$ ($\gamma = .01$) dimensions for an initial distance $d_0 = 2$ to the source. Full histograms correspond to the average trajectories, contour histograms to trajectories with random hits. **B.** Average search time t_s as a function of γ. **C.** total CPU time as a function of γ, calculated as $(t_s/\Delta t)^2$.

strongly attracted back to the origin after a certain time, because the divergence of the prior is not attenuated by the trajectory anymore.

A smarter option would be to add some sort of coarse graining in time: points become much rarer in the distant past, but they have an increasing weight in the discrete sum at the exponent in the posterior probability. We would probably lose some precision this way, but we could recover a linear-time algorithm.

Part II

DNA unzipping and sequencing

Chapter 3

Review of current sequencing technologies and their limitations

In this part we wish to show how micromanipulation experiments on DNA molecules could be exploited to give us better sequencing techniques.

In this chapter we will describe current sequencing technologies, then underline what are their current limitation and what is to be gained from single-molecule sequencing. This will be the basis and motivation for our further work.

Modern DNA sequencing was developed in the second half of the seventies by Sanger et al. [Sanger 75, Sanger 77], a few other methods were tried in the first part of the decade [Maxam 77], but since they do not have modern day equivalents we will not discuss them here

3.1 Chain-termination method

The method developed by Sanger is based on the properties of dideoxynucleotides (ddNTPs): these are modified nucleotides: where normal nucleotides would be deoxynucleotides (dNTPs) these lack the 3′ hydroxyl group on their dexyribose sugar (see figure 3.1), this means that once they are added to a growing strand of DNA, no further nucleotide can be added because they lack the ability to bind with it [Atkinson 69].

In order to be sequenced DNA needs to be single-stranded and in multiple copies each of which has a primer attached to the same point. The copies are then separated in four reactions all of which contain DNA polymerase and all four of the dNTP and only one of the ddNTP in a lower concentration.

The DNA polymerase facilitates the binding of the dNTP on the complementary bases, but once in a while a ddNTP will bind to the chain halting the process. At the end of the process we are left with different pieces of DNA all starting at the same point (where the primer was bound) and ending at random points, with the constraint that all the pieces in the reaction that contained only ddATP end at a T basis, all those in the ddCTP reaction end at a G basis and so on and so forth.

Now the molecules can be sorted according to their size with gel electrophoresis and photographed on four different lanes (one for each of the basis), a black line will appear in correspondence to each base.

Several variation to this technique exist: the ddNTP can be dyed in order for them to fluoresce or tagged with a radioactive substance, but the essential mechanism stays the same.

The main problem with this kind of method is that the quality of the sequencing traces degrade after about 1000 bp. This is due to several factors: the first and most important
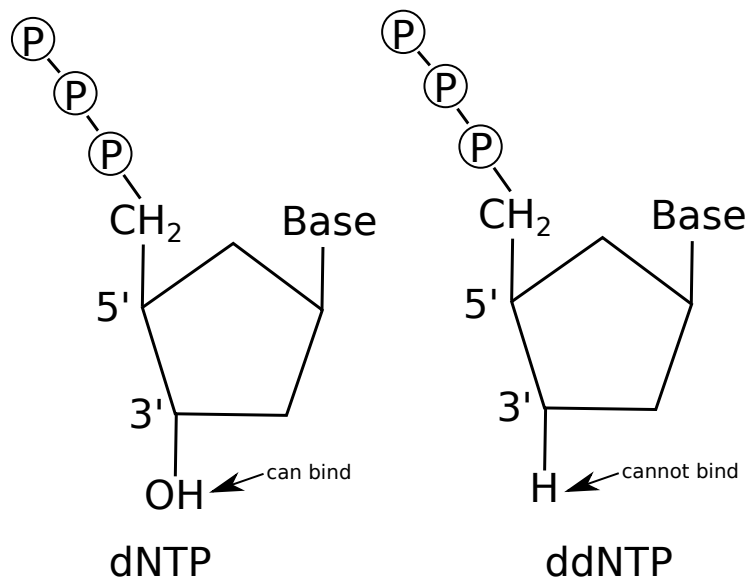
Figure 3.1: Right: a normal Nucleotide TriPhosphate where the sugar is a $2'$-deoxyribosine. Left: a NTP where the sugar is a $2',3'$-dideoxyribosine. The absence of the hydroxide on the $3'$ carbon atom means it no further nucleotide can link to it.
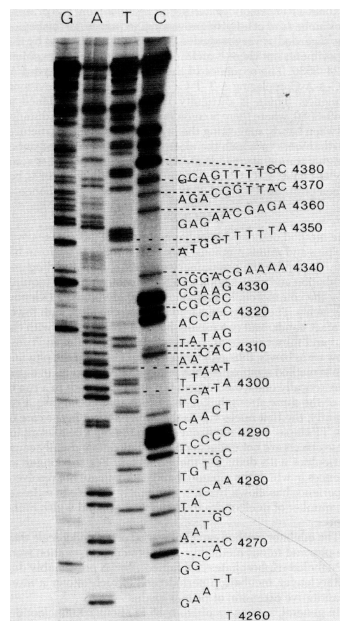

Figure 3.2: One of the figures of Sanger's seminal paper [Sanger 77] showing an autoradiograph of the four lanes of chain-termination sequencing and how they are used for sequencing.

is the nature of the random process involved in the binding of ddNTP. Suppose we are in the ddATP solution and the next base is a T, then the probability p_{dd} of the ddATP binding instead of the dATP binding does not depend on the length of the sequence. On the other hand the probability of still finding a sequence of a certain length after having encountered n T's is $(1 - p_{\mathrm{dd}})^n$ and thus decreases exponentially.

Another source of accumulating errors is the presence of two or more basis of the same kind next to each other, that is to say it is difficult to distinguish four C's in a row from five C's. This type of errors will crop up, making the alignment of the four different lanes difficult.

3.2 Pyrosequencing

Another very popular sequencing technique which is behind some current day automated sequencing methods is pyrosequencing. Developed by Ronaghi and Nyrém in the nineties [Ronaghi 96, Ronaghi 98], pyrosequencing relies on detecting the activity of DNA polymerase through the use of a chemiluminescent enzyme that will emit light whenever a new bond is formed.

A single strand of DNA reacts with DNA polymerase, a chemiluminescent enzyme and solutions of one of the four nucleotides, which are sequentially added and removed. When a nucleotide binds to the next available spot, light is emitted and we know which base has bound because only one type of nucleotide was in solution at that moment.

Pyrosequencing is inherently limited to sequences of about 500 bp (more typically less than 100 bp), but it is well suited to being automated and massively parallelized. Because of the limitations in the size of the the fragments it has been rarely used for *de novo* sequencing, instead it is either used in conjunction with other methods, or for resequencing and for the search for single nucleotide polymorphisms (SNP). Only recently read lengths of about 1000 bp have been attained by a company called 454. This will allow for *de novo* sequencing using pyrosequencing.

3.3 Sequencing by ligation

Ligation is the joining of two double stranded DNA segments through the formation of two covalent bonds. This reaction involves an enzyme called DNA ligase. The difference between DNA ligase and DNA polymerase is that DNA polymerase needs one of the two strands to be intact while DNA ligase can repair double stranded DNA.

DNA ligase can also be used to join a single strand of DNA to an otherwise intact single strand, but in this case it is very sensitive to mismatches, that is it will hardly ever join two strands which are not complementary.

Several techniques are based on this specificity, namely ligase chain reaction (LCR) [Barany 91, Wiedmann 94] and ligase amplification reaction (LAR) [Wu 89], we will not dwell here on the details, it suffices to know that these rely on oligonucleotides (short pieces of ssDNA, here typically 8-9 bases long) and their ligation to a the DNA that is being sequenced.

A number of different oligonucleotides is added to the solution where the anchor sequence is. Then the ligase will hybridize two of the bases of the oligonucleotide to the anchor sequence and emit a light signature that allow the two bases to be recognized.

Sequences are then reconstructed using two-base encoding, a technique that relies on these superposed two-base reads. Read lengths of up to 25-50 bases have been achieved [McKernan 09].

3.4 Limitations

As you might have noticed, all of the techniques outlined up to here rely on read lengths of at most 1000 bp, while whole chromosomes and genomes have lengths that exceed this by several orders of magnitude. In order to fill this gap, DNA has to be spliced and amplified to be sequenced. Amplification is usually done through a technique called polymerase chain reaction (PCR) [Mullis 86, Mullis 94].

DNA can be cut in an ordered way starting from one end and then cutting regularly. This technique is called chromosome walking and it is the best method for sequences which are too long to be sequenced in a go, but still under 10000 bp. The shorter fragments are then sequenced leaving 20 or so superposing bases on each fragment to allow for reconstruction. Longer sequences as whole chromosomes or genomes are usually dealt with a technique developed at the end of the seventies called shotgun sequencing [Staden 79].

The name derives from a metaphor: as a shotgun fires a large array of small projectiles in a random pattern, DNA is cut in random points into smaller sequences. The process is repeated multiple times as to have several copies of the same sequence cut in different points. The spliced sequences can then be sequenced one at a time and then recomposed through the use of algorithms that rely on the overlapping between different copies (see figure 3.3).

Short reads are fine when we are looking for short mutations such as SNPs or anything

Target sequence

AGTGACCATCTGGATTCGGCACGCCCATATGCCCACTTTTTTACATTAGTATTTATGG

Clones' fragments

AGTGACCATCTGGATTCGGCA
CTTTTTTACATTAGTATTTATGG
AGTGACCATCTGGATTCGGCACGCCCATATGCCCA
TTTACATTAGTATTTATGG
ATTCGGCACGCCCATATGCCCACTTTTTTACATTAGTATTTATGG
AGTGACCATCTGG
CGCCCATATGCCCACTTT

Reconstructed sequence

AGTGACCATCTGGATTCGGCACGCCCATATGCCCACTTTTTTACATTAGTATTTATGG
AGTGACCATCTGGATTCGGCACGCCCATATGCCCACTTTTTTACATTAGTATTTATGG
AGTGACCATCTGGATTCGGCACGCCCATATGCCCACTTTTTTACATTAGTATTTATGG

Figure 3.3: Shotgun sequencing: the target sequence is cloned several times and cut at random points. The smaller segments are then sequenced and the sequence is reconstructed thanks to the overlaps.

shorter than the length of the typical read, but genomes are replete with mutations that are much larger in size such as copy number variations (CNV).

Copy number variations are mutations that involve the deletion or the duplication of a section of DNA, they have lengths of at least 1 kbp and up to several hundred kbp and are very common throughout the human genome [Sebat 04, Iafrate 04].

Copy number variations seem to play a central role in cancer [Shlien 10], autism [Sebat 07] and in neurological conditions [Friedman 06, Glessner 10, Sundaram 10]. CNV are very hard to find with current sequencing methods, because reconstruction algorithms tend to miss them. The only way to effectively indentify them is to use classic sequencing techniques in conjunction with microarrays for the detection of SNPs and very complex algorithms [Koike 11].

This is one of the main reasons for developing single molecule techniques for sequencing DNA, but current efforts are not very promising: zero-mode waveguide [Levene 03] seems to be the most advanced but it still offers read lengths of about 1500 bp, that is comparable with chain termination techniques. It is a technique based on holes which are small ($\sim$ 100 nm) in all of their dimensions compared to the frequency of light used for the observation. Their optical properties allow the observation of the enzymatic activity of a single molecule.

On the other hand techniques based on nanopores look promising [Clarke 09]. Nanopores are holes with a diameter of $\sim$ 1 nm, similar to some proteins found on cellular membranes. DNA can be forced through the nanopore one base at a time. Since each nucleotide obstructs the nanopore in a different way it is possible to distinguish between nucleotides by measuring the electrical properties of the obstructed nanopore. These technique is, however, at a very early stage of development.

This is why in the following we will propose a novel approach based on single-molecule experiments of unzipping that could one day be used to sequence DNA.

The reader should keep in mind that no single method is free from the trade-off between resolution and scope, that is to say that it is impossible to attain at the same time accuracy at a single base level and very long reads.

Chapter 4

Modeling DNA unzipping

In the past two decades, the development of experimental techniques that allowed the manipulation of single biological macromolecules at the nm and pN scale has afforded us a wealth of experimental data on the physical properties of said molecules.

At the same time theoretical models have been devised to predict and model the behavior of said molecules. In particular the elasticity of both single-stranded and double stranded DNA is well know and the phase diagram of dsDNA is well understood. Experiments have permitted to denature dsDNA by applying a mechanical force, those experiments have taken the name of unzipping because the DNA is pulled apart from its two strands as a zipper (see figure 4.1.

These experiments are well understood in their single components: the ds- and ssDNA, the fork where the DNA denatures, what was lacking was a clearer picture how the delicate interplay of these different dynamics.

After an introduction to the physics of its single components, we will develop a mesoscopic model for the coupled dynamics and describe a software package for its simulation.

The goal here is to see whether the fluctuations an the correlations that compose the dynamics of linkers and beads will affect the unzipping dynamics of the force. This has already been investigated in [Manosas 05], however this approach is novel and has been published in [Barbieri 09].

4.1   Modeling fork dynamics

The thermodynamics of DNA pairing is a subject that dates back to before the first sequencing techniques were available: a first model was proposed by Tinoco and collaborators in 1971 [Tinoco 71], it gave the free energies for the two types of Watson-Crick bonds and it remarked that further study was needed to take into account stacking interactions, which had been known to be the principal cause of DNA stability for some time then [Crothers 64].

In 1973 the same group published a new letter [Tinoco 73] where new data allowed for the introduction of stacking effects, that is to say that base-pairing free energies now depended, not only on the base itself but on the previous base too. However the results were not very precise and they involved RNA hairpins rather than DNA, it wasn't until the second half of the eighties that reliable data on DNA thermodynamics became available [Breslauer 86]. More recently similar data have been obtained in unzipping experiments. [Huguet 10].

The results of all of this studies are that the free energy of a DNA base pair depends on the base pair itself and its nearest neighbor nucleotide content, that is if we now consider
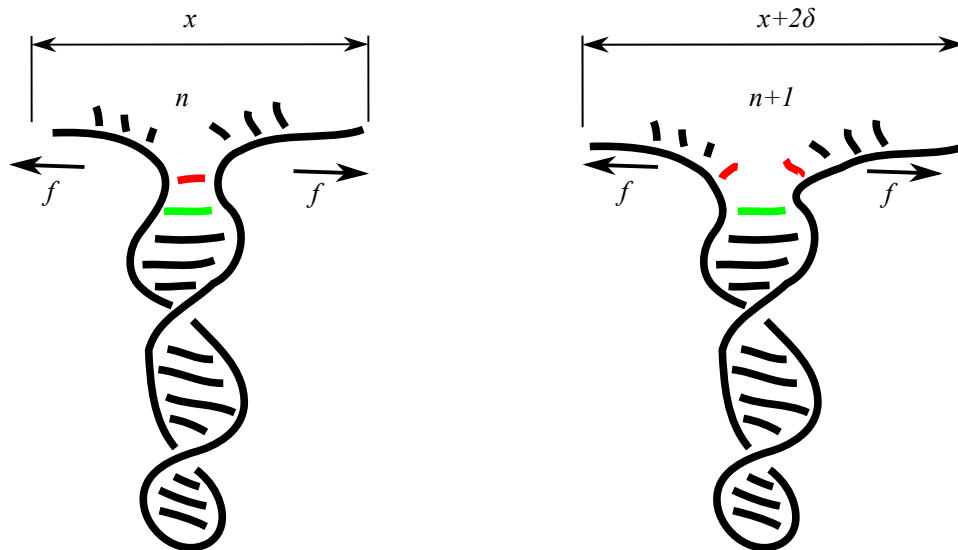
Figure 4.1: Double stranded DNA can be denatured by applying opposite forces to the two strands. In clear analogy with the zipper commonly found in clothing, this type of experiment has benn christened unzipping.

g_0	A	T	C	G
A	1.78	1.55	2.52	2.22
T	1.06	1.78	2.28	2.54
C	2.54	2.22	3.14	3.85
G	2.28	2.52	3.90	3.14

Table 4.1: Binding free energies $g_0(b_i, b_{i+1})$ (units of $k_{\mathrm{B}}T$) obtained from the MFOLD server for DNA at room temperature, pH=7.5, and ionic concentration of 0.15 M. The base values b_i, b_{i+1} are given by the line and column respectively.

a sequence of N bases of dsDNA its free energy will be given by:

$$G(B, N) = \sum_{i=1}^{N} g_0(b_i, b_{i+1}) \,, \tag{4.1}$$

where B denotes the whole sequence and $b_i = A, T, C, G$ is the ith base. Typical values of the binding energies are given in table 4.1.

What we are interested in is the phenomenon of unzipping under a force, the denaturation of dsDNA when the two strands that compose its double helix are pulled.

Let us now suppose for a moment we know the free energy of ssDNA under tension and that this is a linear function of the number of basis and otherwise depends only on the tension f applied to it. At equilibrium we will have that n bases of ssDNA have free energy equal to $ng_{\mathrm{ss}}(f)$. We will focus on the form of $g_{\mathrm{ss}}(f)$ in the following sections, it suffices to say that it needs to be an increasing function of force.

If we model only the motion of the opening fork and we do not include in the model the experimental setup (see figure (4.4): stretching the two strands of DNA away from one another we are able to apply a force and eventually open a base pair. When will this happen? The energy gain from the two new ssDNA bases must be greater than what is lost from the dsDNA energy, that is:

$$\Delta G(i) = g_0(b_i, b_{i+1}) - 2g_{\mathrm{ss}}(f) \,, \tag{4.2}$$

must be negative for the process to be energetically favored.

It is important now to put some numeric values on the quantities involved: the free energies g_0 and g_{ss} are both of the order of a few $k_{\mathrm{B}}T$, forces are expressed in units of pN and distances in units of nm. $k_{\mathrm{B}}T \simeq 4$ pN nm.

The typical range of an hydrogen bond is about 0.1 nm, since the critical force needed to break it is of about 15 pN, this works out to an energy of about 0.4 $k_{\mathrm{B}}T$ which can be neglected with respect to the few $k_{\mathrm{B}}T$ of the binding energy which is known thermodynamically. This means that the opening rate will be independent of force.

Detailed balance then gives us the closing rate, which depends only on the force fluctuation needed to bring the two strands close enough to form the hydrogen bond. We then have:

$$r_{\mathrm{o}}(n) = re^{\beta g_0(n)} \,, \qquad r_{\mathrm{c}}(f) = re^{2\beta g_{\mathrm{ss}}(f)} \,; \tag{4.3}$$

where β is the inverse temperature and r gives the timescale of the phenomenon. We will refer to r as the attempt rate; it can be estimated from the rate of self-diffusion for an object the size of a ssDNA base: $r^{-1} = \beta 2\pi\eta l^3 = 0.17$ μs, where $l = 5$ nm is the size of a base and η is the viscosity of water.

The interplay of the the stacking and pairing energy g_0 and the energy gained from the two newly formed ssDNA bases $2g_{\mathrm{ss}}$ is responsible for the formation of a complex energy
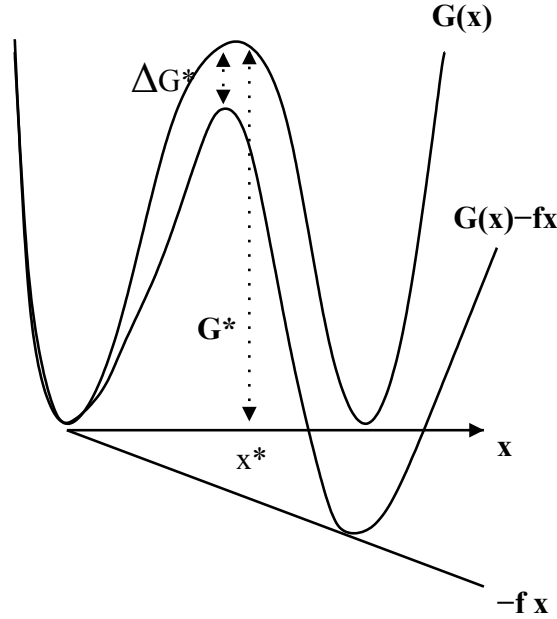
Figure 4.2: The switching rate between the two states is proportional to $\exp(\beta G^*)$, where G^* is the free energy barrier. The application of a force f tilts the distribution and lowers the barrier of $\Delta G^* \simeq -fx^*$. Actual numerical values indicate this can be neglected with respect to G^*.

landscape full of metastable minima, not dissimilar, to a one-dimensional random walk in a random environment, also known as the Sinai model. This suggest the use of methods from the statistical mechanics of disordered systems and from information theory for the description and analysis of such a system.

In figure 4.3 we show as an example the free energy derived from the first 50 base-pairs of the λ-phage DNA at two different forces.

Cocco and collaborators first worked out the opening and closing rates in [Cocco 01, Cocco 03], as a Eyring-Kramers transition state theory [Eyring 35, Kramers 40]. This theory describe the transition with a suitable continuous variable (which here is the separation x between the two bases forming the base pair), x obeys Langevin dynamics over an effective potential that is the free energy $G(x)$. This potential has two local minima at the two equilibrium position that correspond to broken/whole hydrogen bonds (see figure 4.2).

Figure 4.3: Free energy G (units of $k_{\mathrm{B}}T$) to open the first n base-pairs, for 200 randomly selected bases.

4.2   ssDNA as a modified freely jointed chain

One of the simplest polymer models possible is that of the freely jointed chain. The FJC is composed of N monomers of length d, no constraint is put on the angles formed by consecutive segments and the excluded volume is not taken into account.
The end-to-end distance is thus given by:

$$\vec{R} = \sum_{i}^{N} \vec{r}_i\,, \tag{4.4}$$

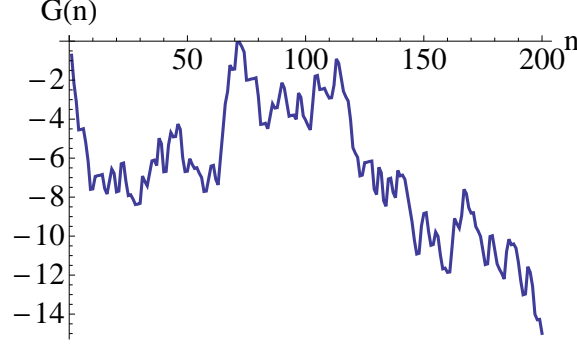where the $\vec{r}_i$ are random vectors of length d. This length is often referred to as Kuhn length.
If we are in the thermodynamic limit we can use the central limit theorem to show that the average end-to-end distance $\langle R \rangle$ vanishes and that it is distributed according to a normal distribution of variance $\langle R^2 \rangle = Nd^2$.
To get this result [Kuhn 42, James 43] we have assumed that no force was acting on one end of the chain, and it is, in fact, only valid around for end-to-end lengths of the order of $\sqrt{N}d$. In order to get the right result for high extensions we have to add a tensile force f applied in the x direction. We can now compute the average value of the x component of the ith link of the polymer as:

$$l_{\mathrm{FJC}}(f) = \langle x_i \rangle = \frac{\int_{-d}^{d} x_i \exp(\beta f x_i) dx_i}{\int_{-d}^{d} \exp(\beta f x_i) dx_i} = d\mathcal{L}(\beta f d)\,, \tag{4.5}$$

Where $\mathcal{L}(x) = \coth(x) - 1/x$ is the Langevin function. The total length of the polymer along the x axis is then given by $L_{\mathrm{FJC}}(f) = N l_{\mathrm{FJC}}(f)$. The interested reader can find further details in a classical reference such as [Flory 53].
At the beginning of the nineties it became possible to measure the elasticity of DNA with magnetic beads [Smith 92]. It then became apparent that, up to forces of 20 pN/nm the elasticity of ssDNA is well fitted by a FJC model, but even better results are obtained using a modified FJC where the monomers are extensible at high forces and where the contour length (*i. e.* the total stretched length of the polymer) is not given by the product of the number of momomers and the Kuhn length:

$$l_{\mathrm{MFJC}}(f) = l_{\mathrm{FJC}} \left(1 + \frac{f}{\gamma_{\mathrm{ss}}} \right) = d \left(\coth(\beta f b) - \frac{1}{\beta f b} \right) \left(1 + \frac{f}{\gamma_{\mathrm{ss}}} \right) \tag{4.6}$$

where $d = 0.56$ nm, $b = 1.4$ nm and $\gamma_{\mathrm{ss}} = 800$ pN [Smith 96].

4.3 dsDNA as an exstensible worm-like chain

The worm-like chain (WLC) is one of the simplest continuous models of a polymer: if we define a parametric curve in space $\vec{r}(s)$ we can define it's tangent vector as $\vec{t} = \frac{d\vec{r}(s)}{ds}$ and its curvature vector as $\vec{w} = \frac{d\vec{t}(s)}{ds}$, we can further impose that the polymer is inextensible, that is: $|vect(s)| = 1$.

Then we can give the internal energy for a polymer stretched by an external force f as:

$$\beta E = \int_0^{L_{\text{tot}}} ds \frac{A}{2} |\vec{w}(s)|^2 - \beta f \hat{t}(s) \cdot \hat{x}, \tag{4.7}$$

where A is the persistence length, that turns out to be the correlation length of the direction of the polymer at zero force.

The WLC is analytically solvable model, however the solution can only be written as an infinite series **??**. Luckily a very precise numerical fit has been proposed by Marko and Siggia in [Marko 94, Marko 95]:

$$\beta f A = \frac{l_{\text{WLC}}}{l_{\text{tot}}} + \frac{1}{4(1 - l_{\text{WLC}}/l_{\text{tot}})^2} - \frac{1}{4}, \tag{4.8}$$

where l_{tot} is the contour length of the polymer divided by the number of bases, A is the persistence length and l_{WLC} is the length of the polymer in the direction of the force f.

In the following years even more refined fits to the experimental data have been proposed such as the one by Moroz and Nelson [Moroz 97] which used a formula first proposed by Odijk [Odijk 95]. Their formula can fit the experimental data for the elasticity of dsDNA for a very large range of forces [Bouchiat 99], thanks to the relaxation of the hypothesis that $|\vec{t}(s)| = 1$, which plays an important role at high forces and the inclusion of torsional effects.

However we do not need such a large range of forces for the description of unzipping experiments; because of this that in the following we will use a simplified version of the Odijk formula, namely:

$$l_{\text{WLC}}(f) = l_{\text{tot}} \left[1 - \frac{1}{2} \left(\beta f A \right)^{-1/2} + \frac{f}{\gamma_{\text{ds}}} \right], \tag{4.9}$$

where $l_{\text{tot}} = 0.34$ nm $A = 48$ nm and $\gamma_{\text{ds}} = 1000$ pN.

4.4 Two possible ensembles

The description we have given above does not depend much on the experimental setup, the only time where we have lost some generality is in the description of fork dynamics, where we have assumed the force to be fixed; however both the polymer description and our choice for the dynamics are completely independent of details like this.

In the following we will outline two possible experimental setups (pictured in figure 4.4): in the first force is a parameter and the extension of the polymer, which is directly related to the number of open bases, is measured; in the second the distance between two optical traps can be varied as a parameter and the displacement of the beads in the traps can be measured to give a precise measurement of force.

The only detail that needs to be sorted out is the change in variable in the thermodynamic potentials that describe different setups, but this can be easily done through a Legendre transform. Before we go on we should lay out the notation we will use in the following: first of all capital letters denote extensive quantities, while lower case letters correspond to
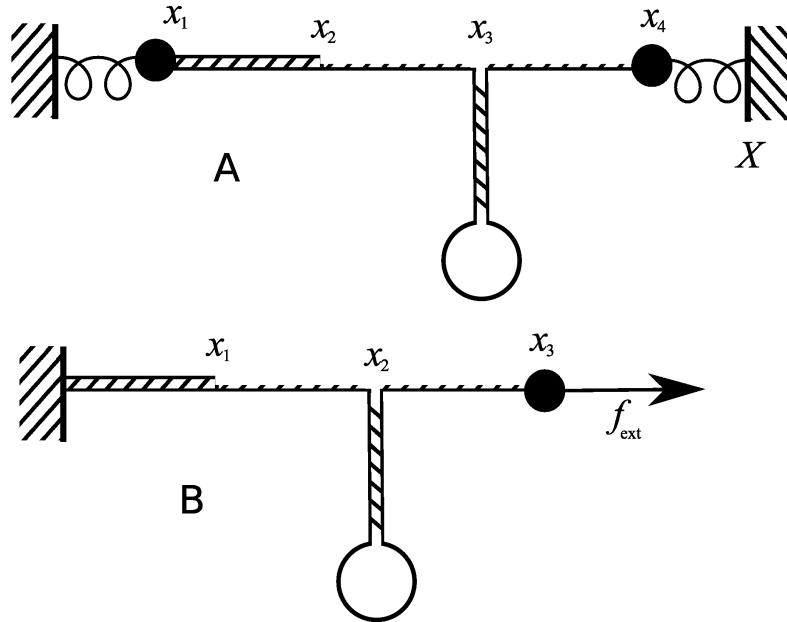
Figure 4.4: Typical experimental setups that will be described in the following. A) A setup with two optical traps (beads x_1 and x_4) drawn as springs and whose centers are the black vertical lines; B) a setup with a single magnetic bead x_3 that applies a constant force on the molecule attached to a fixed "wall". In both cases the molecular construction is made by a DNA molecule that has to be opened (therefore one should include two single-strand linkers that are the opened parts of the molecule) and one double-stranded DNA linker. The coordinates x_i are the distances of the corresponding points from the left reference position (which is the center of the left optical trap in case A and the fixed wall in case B).

the equivalent intensive quantity. x is the end to end distance of a polymer and $l = x/n$, where n is the number of momomers (bases here). For example $W(x) = nw(l)$

Let's lay out all the quantities:

- $g(f)$ is the free energy per base as a function of force.

- $l(f) = \frac{\partial g(f)}{\partial f}$ is the length as a function of force.

- $w(l) = \max_f[fl - g(f)]$ the free energy as a function of length.

- $f(l) = \frac{\partial w(l)}{\partial l}$ the force as a function of length or the inverse of $l(f)$.

- $k(l) = \frac{\partial f(l)}{\partial l}$ is the effective spring constant for a given length.

- $\frac{1}{k(f)} = \frac{\partial l(f)}{\partial f}$ is the reciprocal effective spring constant as a function of force.

4.4.1 Fixed force, magnetic tweezers

At the beginning of the 2000s Gosse and Croquette [Gosse 02] developed a technique called optical tweezing: a superparamagnetic bead with a diameter of the scale of the µm is placed under the two poles of a permanent magnet, which creates a magnetic gradient. The distance between the poles of the magnet (less than 1 mm) is fixed so that on the scale of the typical movements of the bead the gradient of the magnetic field is almost constant and so is the force applied to the bead.

Magnetic beads have a preferred direction. This is at the same time an advantage and a disadvantage: the advantage is the possibility of applying a torque to the bead, which has opened the door to experiments involving the coiling and uncoiling of DNA; on the other hand the DNA will bind on a random point of the surface and it is impossible to say exactly where. Given the relative size of the bead and of a single base of DNA, this means that the unzipping experiment can start up to 1000 bp away in two different runs. The position of the bead can be recorded optically. This type of experiments are relatively easy to set up, and the modellization of fork dynamics at fixed force is perhaps more intuitive. On the other hand fixed force experiments tend to be ill suited for sequencing purposes, since it is difficult to control the position along the energy landscape where the fork will stop.

For a given portion of the sequence there exists an average critical opening force. When the critical force is exerted the fork will fluctuate around a given number of open bases for a long time because it is in a potential well. On the other hand the top of the potential barriers that separate these wells are very hard to sample, because very little time will be spent there.

Another reason why this method is not very well suited for sequencing through unzipping is that the position of the fork for a given force depends strongly on the sequence and it is very hard to generate an unzipping protocol with varying force without a prior knowledge of the sequence.
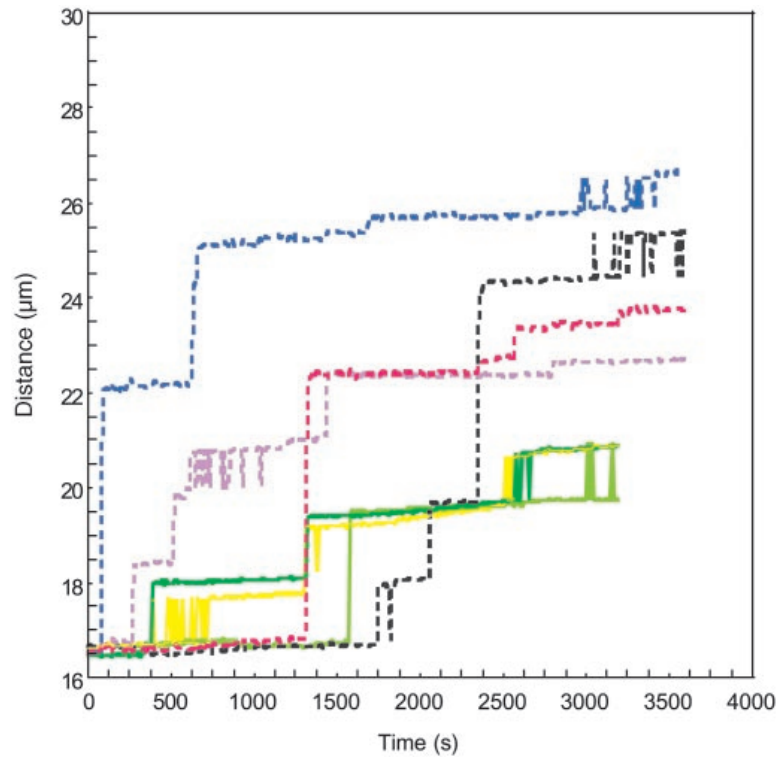
Figure 4.5: Several typical fixed force unzipping traces from [Danilowicz 03]. Solid lines correspond to a force of 15 pN, while dashed lines correspond to a force of 20 pN. The measured quantity is the distance between the center of the magnetic trap and the surface of a glass micropipette which the DNA is attached to. Horizontal plateaux correspond to minima of the free energy.

4.4.2 Fixed distance, optical tweezers

Pioneering studies on the effect of radiation pressure from laser light on micrometer-sized dielectric beads were performed at the beginning of the seventies by Ashkin [Ashkin 70]. A few years later the same Ashkin developed a single beam technique for trapping dielectric beads [Ashkin 86].

In optical tweezers a tightly focused laser beam passes through a dielectric sphere which has an optical index higher than that of the surrounding fluid. The incoming light from the laser is refracted by the bead causing a change in the momentum of the outgoing light; because of the conservation of momentum, the bead will experience a change of momentum of opposite sign.

For high enough numerical aperture of the laser there exists a stable position of the bead along the axis of propagation of laser light, on the other hand stability along the transversal directions is due to the intensity profile of the laser, which is most of the times Gaussian. In order to give a precise description of the phenomenon for the conditions most often used in micromanipulation experiments, we should take into account the full Mie theory of light scattering, since the bead size (1 µm) is very close to the wavelength of the laser employed (see for example [Mangeol 08], where the laser wavelength is 1.064 µm).

On the other hand we can give an hand-waving argument for the stability of the trap using ray optics: a particle with a refractive index higher than water will act as a positive lens, roughly speaking if the bead is placed before (after) the focal point the rays will diverge (converge).

If the lens converges the ray the light will have more momentum in the direction of propagation of the beam, conversely, if the beams have been diverged the light will lose momentum. See figure 4.6 for a schematic picture. The interested reader should refer to Kerker's book [Kerker 69] for a full treatment of the Rayleigh and Mie regimes.

The use of optical traps for the manipulation of biopolymers is compelling because it allows to fix the position of the beads and to measure the force exerted on the molecule. This is very attractive for unzipping experiment because it gives us a chance to focus on a specific region of the sequence, while in fixed force experiments the region of DNA where the fork will spend most of the time depends on the sequence itself.

The measurement of force is obtained by the observation of the displacement of the bead with respect to the center of the bead. The optical trap is well approximated by an harmonic potential around its equilibrium position. The displacement of the bead can be measured either by direct observation of the diffraction pattern through a microscope, or by measuring the deflection of the laser beam with a PSD [Wallmark 57].
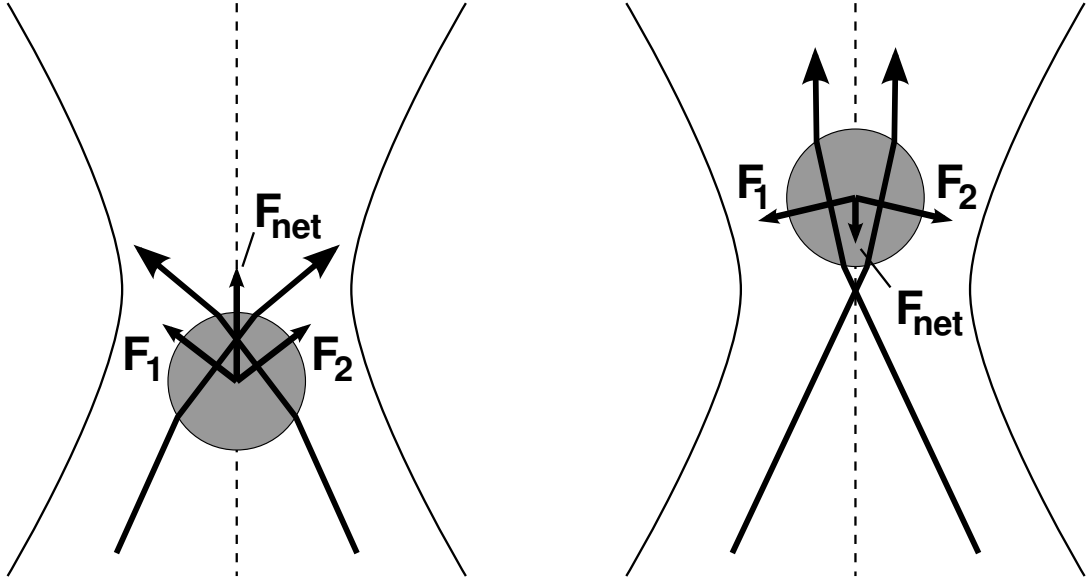
Figure 4.6: The effect of the refraction of light on a dielectric bead in the ray optic approximation. The light propagates from bottom to top. F_1 and F_2 are the forces acting on the bead because of the concentration of momentum, F_{net} their resultant.
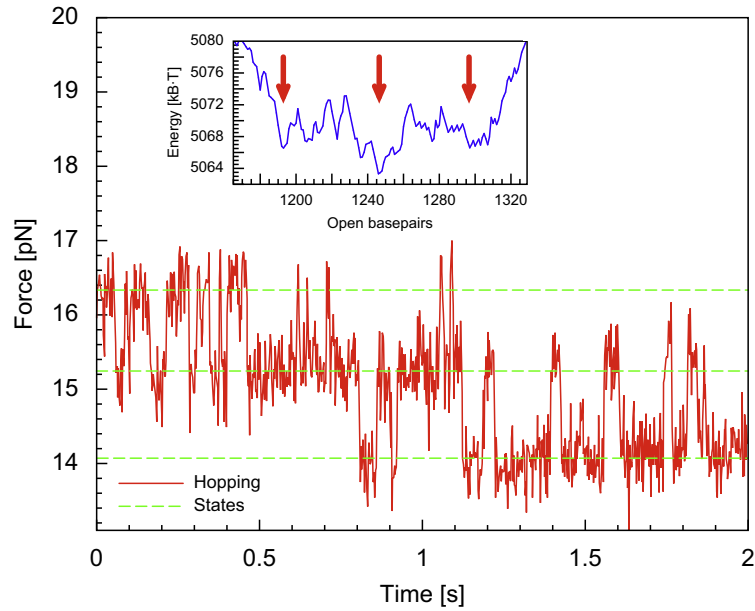
Figure 4.7: A typical fixed distance unzipping trace from [Mossa 10]. The force, measured as the displacement of the bead in the optical trap, is measured as a function of time. Notice how the three minima of the free energy correspond to the three green lines where the bead spends most of its time.

4.5 Overdamped dynamics

The motion of very small objects suspended in a liquid does not resemble much to that of objects in everyday life. The most striking features are the absence of inertial effects and Brownian noise.

A common way to quantify the ratio between inertial and viscous effects is the Reynolds number Re, which was introduced by Stokes [Stokes 51], several years before Reynolds popularized it.

It is given by:

$$\mathrm{Re} = \frac{Vl\rho}{\eta}\,, \tag{4.10}$$

where V is the mean velocity of the object with respect to the fluid, ρ is the density of the fluid, l is the linear size of the object and η is the viscosity of the fluid.

It appears in the dimensionless Navier-Stokes[1] equation for an object immersed in a Newtonian fluid as:

$$\mathrm{Re}\left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}\right) = -\nabla p + \nabla^2 \mathbf{v} + \mathbf{f} \tag{4.11}$$

where $\mathbf{v}$ is the speed of the object divided by V, p is the pressure of the fluid divided by $\eta V/l$, $\mathbf{f}$ are the external forces per unit volume divided by $\eta V/l^2$, ∇ stands for the the space partial derivatives vector multiplied by l and finally $\partial/\partial t$ is the time derivative multiplied by l/V.

The importance of the Reynolds number is that it is the only quantity needed to describe the flow of a fluid, that is to say that once the variables have been properly rescaled sistems of different size, viscosity and density will behave the same way.

It is customary to categorize the characteristics of the flow according to the Reynolds number:

- $\mathrm{Re} \gg 1$: Turbulent flow. Inertial forces are dominant.
  E.g. man swimming, the wing of a plane.

- $\mathrm{Re} \sim 1$: Laminar flow. Viscous forces dominate. *E.g.* water in a pipe.
  E.g. blood flow, fish swimming, man swimming in glycerol.

- $\mathrm{Re} \ll 1$: Creeping flow. Inertial forces are completely negligible.
  E.g. Bacteria in water, μm-sized beads in optical traps, macromolecules in solution.

Among the objects that we will consider in the following those who have the largest Reynolds number are the beads in the optical traps; for them $\mathrm{Re} \sim 10^{-6}$, because of this, the remarks we will make on their dynamic behavior will be all the more valid for objects with lower Reynolds number.

Let us suppose that a bead of diameter d, is suspended in water by an optical trap of stiffness k. Let us also suppose for the moment that the bead has the same density as the water surrounding it.

The bead obeys the Langevin harmonic oscillator equation:

$$m\ddot{x} + \gamma\dot{x} + kx = \xi(t)\,, \tag{4.12}$$

where $m = 1/6\pi\rho d^3$, $\gamma = \beta 6\pi\eta d$ and $\xi(t)$ is Gaussian noise obeying

$$\langle \xi(t) \rangle = 0, \qquad \langle \xi(t)\xi(t') \rangle = 2\gamma k_{\mathrm{B}}T\delta(t-t')\,. \tag{4.13}$$

---

[1]Please note that this is not the only way to rescale the variables in order to make the adimensional: ρV^2 has the dimensions of a pressure and can be used to the same effect, it turns out this latter is the right scaling for high Reynolds numbers, while the one in the main text is the right one for the limit of low Re.
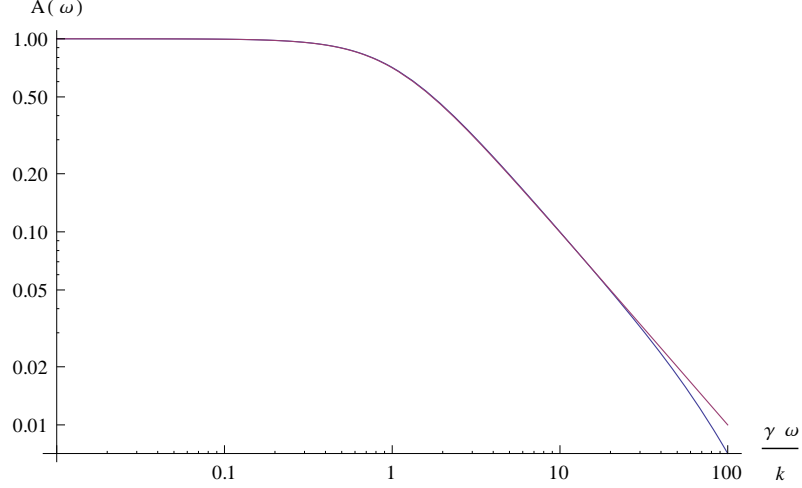
Figure 4.8: The amplitude response as a function of frequency: the blue curve corresponds to eq. (4.14 while the violet one corresponds to eq. (4.15). For this plot we have chosen $\frac{mk}{\gamma^2} = 100$, this way the cutoff frequency is well below the frequency where the mass effects become dominant.

We now consider the frequency response by performing a Fourier transform obtaining

$$A(\omega) = \frac{1}{\sqrt{(k - m\omega^2)^2 + (\gamma\omega)^2}} \, . \tag{4.14}$$

The question is: when can this be approximated by its Brownian counterpart, neglecting the mass term?
The new equation for the frequency response would read:

$$A(\omega) = \frac{1}{\sqrt{k^2 + (\gamma\omega)^2}} \, . \tag{4.15}$$

Now this response has a cutoff frequency of γ/k, what we want, in order for our approximation to be valid, is for this frequency to be much smaller than the one at which mass effects become important, that is m/γ. Summing up we want

$$\frac{mk}{\gamma^2} = \frac{l\rho k}{(6\pi)^3\eta^2} \gg 1 \, . \tag{4.16}$$

Plugging in realistic values for the stiffness of the trap $k = 0.5$ pN/nm, for the density $\rho = 1$ g/cm^3, the diameter of the bead $l = 1$ µm and the viscosity of water $8.9\,10^{-4}$ Pa s; we find the ratio to be very small: $\frac{mk}{\gamma^2} = 1.9\,10^{-4}$. This is in accordance with what we would have expected by using the Reynolds number, in fact µm-sized beads are well within the creeping flow range for speeds up to 10 cm/s.
In the following discussion we will be well justified in leaving out the mass terms from our equations and considering all of the dynamics as Brownian or overdamped.

4.6 Coupling all the dynamics together

In this section we will derive an effective mesoscopic dynamical equation for coupled heteropolymers. The details of the calculation are somewhat technical, but they offer a

different insight from the derivation published in the appendix of [Barbieri 09].

Because of the preceding discussion on overdamped dynamics we will ignore all inertial effects. In addition to this simplification we will consider the simplest polymer model: a chain of simple Hookean springs, also known as the Rouse model [Rouse Jr 53]. We will also consider the model to be effectively one dimensional.

What we hope to understand better here is how movement propagates along a heteropolymer, what kind of fluctuations and correlations are important and how. It is also of interest to know whether the polymer can be considered at equilibrium and what are the relaxations times.

We will show that in a mesoscopic description where we do not describe single monomers the noise is not decorrelated and we will propose a way to implement these characteristics in computer experiments.

4.6.1 Scaling of a homogeneous Rouse polymer

Let us now derive the equations for the simplest case: that of a homogeneous polymer. At first we will derive the equation for the free end of the polymer and then we will concentrate on a midpoint to see how the dynamics are coupled, we will see of this leads to a viscous drag matrix on the left hand side and how this translates into fluctuation dissipation relations for correlated noise.

Each monomer is characterized by its spring constant k and its viscous drag coefficient γ. Let us suppose that a chain of N identical springs is connected to a non moving wall on one end and that a constant force f is exerted on the other end. The setup is shown in figure 4.9.

The monomers will then obey this system of simultaneous equations:
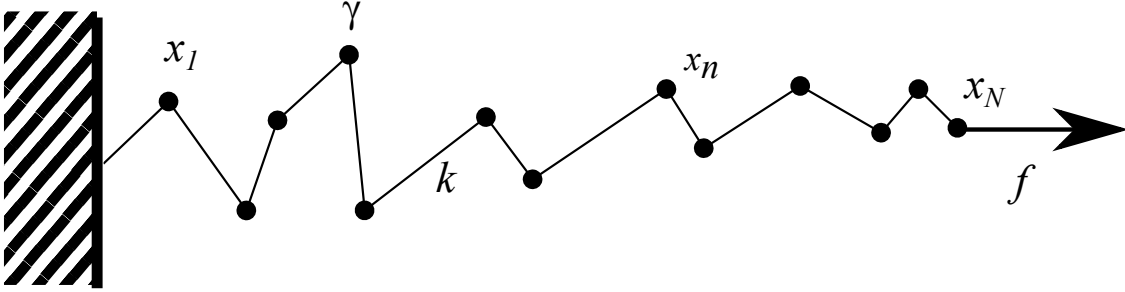
Figure 4.9: A homogeneous Rouse polymer composed of N identical springs and beads each having spring constant k and viscous drag coefficient γ. The coordinates x_i are taken along the direction of the pulling force f.

$$\gamma \dot{x}_1 = -2kx_1 + kx_2 + \eta_1$$

$$\vdots$$

$$\gamma \dot{x}_n = -2kx_n + kx_{n-1} + kx_{n+1} + \eta_n \qquad (4.17)$$

$$\vdots$$

$$\gamma \dot{x}_N = -kx_N + kx_{N-1} + f + \eta_N \,,$$

where x_n is the coordinate of the nth link. η_n are uncorrelated Gaussian noises of zero average and autocorrelation function:

$$\langle \eta_i(t)\eta_j(0) \rangle = 2\gamma k_{\text{B}} T \delta_{ij} \delta(t) \,. \qquad (4.18)$$

Equation (4.17) can be solved formally for x_1 in terms of integrals of x_2. Thus:

$$x_1(t) = \frac{1}{\gamma} \int_0^\infty e^{-\frac{2k}{\gamma}(t-t')} \left[kx_2(t') + \eta_1(t') \right] . \tag{4.19}$$

This doesn't bring us any closer to solving the system of equations, in fact iterating this procedure will only produce an integro-differential equation of order N. To make the problem tractable we have to solve it in the limit in which the ratio γ/k is small, which is reasonable given that for ssDNA at typical conditions it has the value of approximately 10^{-10} s, many orders of magnitude below experimental resolution.
Equation (4.19) thus becomes:

$$x_1(t) = \frac{1}{2} \left[x_2(t) + \eta_1(t) \right] - \frac{\gamma}{4k} \dot{x}_1(t) + o\left(\frac{\gamma}{k}\right) . \tag{4.20}$$

Substitution of this into the equation for x_2 yields:

$$\frac{5}{4} \gamma \dot{x}_2 = -(k + \frac{1}{2}k)x_2 + kx_3 + \eta_2 + \frac{1}{2}\eta_1 , \tag{4.21}$$

The fluctuation-dissipation theorem for Brownian dynamics of the form $\gamma \dot{x} = -\nabla V(x) + \eta$ states that, in order for Boltzmann equilibrium to be attained the following relation must be verified:

$$\langle \eta(t)\eta(0) \rangle = 2\gamma k_{\mathrm{B}} T \delta(t) . \tag{4.22}$$

For equation (4.21) this translates to:

$$\left\langle \left(\eta_2(t) + \frac{1}{2}\eta_1(t) \right) \left(\eta_2(0) + \frac{1}{2}\eta_1(0) \right) \right\rangle = \langle \eta_2(t)\eta_2(0) \rangle + \frac{1}{4}\langle \eta_1(t)\eta_1(0) \rangle$$
$$= 2\frac{5}{4}\gamma k_{\mathrm{B}} T \delta(t) . \tag{4.23}$$

This means that our approximation is consistent with the fluctuation-dissipation theorem. It is obvious that the iteration of this procedure will define renormalised k and γ and new η_n which will be correlated. We can write:

$$\gamma_{n-1}\dot{x}_{n-1} = -(k + k_{n-1})x_{n-1} + kx_n + \eta'_{n-1} . \tag{4.24}$$

Then solve this equation with the usual approximation finding:

$$x_{n-1} = \frac{k}{k+k_{n-1}}x_n - \frac{\gamma_{n-1}k}{(k+k_{n-1})^2}\dot{x}_n + \frac{1}{k+k_{n-1}}\eta'_{n-1} + o\left(\frac{\gamma}{k}\right) , \tag{4.25}$$

which must be inserted in the equation for x_n:

$$\left(\gamma + \frac{\gamma_{n-1}k^2}{(k+k_{n-1})^2} \right)\dot{x}_n = -\left(k + \frac{kk_{n-1}}{k+k_{n-1}} \right)x_n + kx_{n+1}$$
$$+ \eta_n + \frac{k}{k+k_{n-1}}\eta'_{n-1} , \tag{4.26}$$

thus defining recurrence relations for the coefficients:

$$k_n = \frac{kk_{n-1}}{k+k_{n-1}} ; \tag{4.27}$$

$$\gamma_n = \gamma + \frac{\gamma_{n-1}k^2}{(k+k_{n-1})^2} ; \tag{4.28}$$

$$\langle \eta'_n(t)\eta'_n(0) \rangle = \langle \eta_n(t)\eta_n(0) \rangle + \frac{k^2}{(k+k_{n-1})^2}\langle \eta'_{n-1}(t)\eta'_{n-1}(0) \rangle . \tag{4.29}$$

Applying the fluctuation dissipation theorem to the last equation shows that we have chosen the only approximation consistent with the preceding equation. That is to say that the γ's on the left hand side obey the same recurrence relations as the Brownian noises.

As we have already calculated the values of the constants for $n = 2$ we can easily solve the recurrences:

$$k_n = \frac{1}{n}\,; \tag{4.30}$$

$$\gamma_n = \frac{(2n + 1)(n + 1)}{6n}\gamma\,. \tag{4.31}$$

$$\tag{4.32}$$

This way we can rewrite equations (4.26) and (4.25) as:

$$\frac{(2n + 1)(n + 1)}{6n}\gamma\dot{x}_n = -\left(k + \frac{k}{n}\right)x_n + kx_{n+1} + \eta'_n\,; \tag{4.33}$$

$$x_{n-1} = \frac{n - 1}{n}x_n - \frac{(2n - 1)(n - 1)}{6n}\frac{\gamma}{k}\dot{x}_n + \frac{n - 1}{nk}\eta'_{n-1}\,. \tag{4.34}$$

The recurrence can be completely closed with the help of the equation for x_N as:

$$\frac{(2N + 1)(N + 1)}{6N}\gamma\dot{x}_N = -\frac{k}{N}x_N + f + \eta'_N\,. \tag{4.35}$$

Not surprisingly we recover the scalings of the Rouse model when it is solved in the continuous n limit (see for example [Doi 86]) in that it gives:

$$\frac{N}{3}\gamma\dot{x}_N = -\frac{k}{N}x_N + f + \eta'_N\,, \tag{4.36}$$

What we would like to explore now is what happens to a subpolymer, *i. e.* write down the evolution of one end of the polymer and of a midpoint, integrating out all other degrees of freedom. To do so we need to start from the $(N - 1)^{\text{th}}$ link of the polymer.

$$\gamma\dot{x}_{N-1} = -2kx_{N-1} + kx_{N-2} + kx_N + \eta_{N-1}\,. \tag{4.37}$$

which gives, after the usual procedure:

$$x_{N-1} = \frac{1}{2}\left[x_{N-2} + x_N\right] + \frac{\gamma}{4k}\left[\dot{x}_{N-2} + \dot{x}_N\right] + \frac{\eta_{N-1}}{2k} + o\left(\frac{\gamma}{k}\right)\,, \tag{4.38}$$

which can now be used in the $(N - 2)^{\text{th}}$ and N^{th} equations yielding:

$$\frac{5}{4}\gamma\dot{x}_{N-2} + \frac{1}{4}\gamma\dot{x}_N = -\left(k + \frac{k}{2}\right)x_{N-2} + \frac{k}{2}x_{N-2} + kx_{N-3} + \eta_{N-2} + \frac{\eta_{N-1}}{2} \tag{4.39}$$

$$\frac{5}{4}\gamma\dot{x}_N + \frac{1}{4}\gamma\dot{x}_{N-2} = -\frac{k}{2}x_N + \frac{k}{2}x_{N-2} + f + \eta_N + \frac{\eta_{N-1}}{2}\,. \tag{4.40}$$

If we define:

$$\gamma_n^a\dot{x}_{N-n} + \tilde{\gamma}_n\dot{x}_N = -(k + \tilde{k}_n)x_{N-n} + \tilde{k}_nx_N + kx_{N-n-1} + \tilde{\eta}_{N-n} \tag{4.41}$$

$$\gamma_n^b\dot{x}_N + \tilde{\gamma}_n\dot{x}_{N-n} = -\tilde{k}_nx_N + \tilde{k}_nx_{N-n} + f + \tilde{\eta}_N^{(n)}\,, \tag{4.42}$$

we can solve the first to get recurrence equations:

$$\begin{aligned}
x_{N-n} = {} & \frac{\tilde{k}_n}{k + \tilde{k}_n}x_N + \frac{k}{k + \tilde{k}_n}x_{N-n-1} - \left(\frac{\gamma_n^a\tilde{k}_n}{(k + \tilde{k}_n)^2} + \frac{\tilde{\gamma}_n}{k + \tilde{k}_n}\right)\dot{x}_N \\
& - \frac{\gamma_n^a k}{(k + \tilde{k}_n)^2}\dot{x}_{N-n-1} + \frac{\tilde{\eta}_{N-n}}{k + \tilde{k}_n} + o\left(\frac{\gamma}{k}\right)\,,
\end{aligned} \tag{4.43}$$

and deriving:

$$\dot{x}_{N-n} = \frac{\tilde{k}_n}{k + \tilde{k}_n}\dot{x}_N + \frac{k}{k + \tilde{k}_n}\dot{x}_{N-n-1} + O\left(\frac{\gamma}{k}\right) . \tag{4.44}$$

These last two expressions need to be used in the equation for the $(N - n - 1)^{\text{th}}$ link and in equation (4.42) to define the recurrence relations:

$$\left(\gamma + \frac{\gamma_n^a k^2}{(k + \tilde{k}_n)^2}\right)\dot{x}_{N-n-1} + \left(\frac{\gamma_n^a \tilde{k}_n k}{(k + \tilde{k}_n)^2} + \frac{\tilde{\gamma}_n k}{k + \tilde{k}_n}\right)\dot{x}_N =$$
$$-\left(k + \frac{k\tilde{k}_n}{k + \tilde{k}_n}\right)x_{N-n-1} + \frac{k\tilde{k}_n}{k + \tilde{k}_n}x_N + kx_{N-n-2} + \left(\eta_{N-n-1} + \frac{k}{k + \tilde{k}_n}\tilde{\eta}_{N-n}\right) ; \tag{4.45}$$

$$\left(\gamma_n^b + \frac{2\tilde{\gamma}_n \tilde{k}_n}{k + \tilde{k}_n} + \frac{\gamma_n^a \tilde{k}_n^2}{(k + \tilde{k}_n)^2}\right)\dot{x}_N + \left(\frac{\gamma_n^a \tilde{k}_n k}{(k + \tilde{k}_n)^2} + \frac{\tilde{\gamma}_n k}{k + \tilde{k}_n}\right)\dot{x}_{N-n-1} =$$
$$-\frac{k\tilde{k}_n}{k + \tilde{k}_n}x_N + \frac{k\tilde{k}_n}{k + \tilde{k}_n}x_{N-n-1} + f + \left(\tilde{\eta}_N^{(n)} + \frac{\tilde{k}_n}{k + \tilde{k}_n}\tilde{\eta}_{N-n}\right) ; \tag{4.46}$$

and then:

$$\tilde{k}_{n+1} = \frac{k\tilde{k}_n}{k + \tilde{k}_n} ; \tag{4.47}$$

$$\gamma_{n+1}^a = \gamma + \frac{\gamma_n^a k^2}{(k + \tilde{k}_n)^2} ; \tag{4.48}$$

$$\tilde{\gamma}_{n+1} = \frac{\gamma_n^a \tilde{k}_n k}{(k + \tilde{k}_n)^2} + \frac{\tilde{\gamma}_n k}{k + \tilde{k}_n} ; \tag{4.49}$$

$$\gamma_{n+1}^b = \gamma_n^b + \frac{2\tilde{\gamma}_n \tilde{k}_n}{k + \tilde{k}_n} + \frac{\gamma_n^a \tilde{k}_n^2}{(k + \tilde{k}_n)^2} ; \tag{4.50}$$

$$\langle\tilde{\eta}_{N-n-1}(t)\tilde{\eta}_{N-n-1}(0)\rangle = \langle\eta_{N-n-1}(t)\eta_{N-n-1}(0)\rangle$$
$$+ \frac{k^2}{(k + \tilde{k}_n)^2}\langle\tilde{\eta}_{N-n}(t)\tilde{\eta}_{N-n}(0)\rangle ; \tag{4.51}$$

$$\langle\tilde{\eta}_{N-n-1}(t)\tilde{\eta}_N^{(n+1)}(0)\rangle = \frac{\tilde{k}_n k}{(k + \tilde{k}_n)^2}\langle\tilde{\eta}_{N-n}(t)\tilde{\eta}_{N-n}(0)\rangle$$
$$+ \frac{k}{k + \tilde{k}_n}\langle\tilde{\eta}_{N-n}(t)\tilde{\eta}_N^{(n)}(0)\rangle ; \tag{4.52}$$

$$\langle\tilde{\eta}_N^{(n+1)}(t)\tilde{\eta}_N^{(n+1)}(0)\rangle = \langle\tilde{\eta}_N^{(n)}(t)\tilde{\eta}_N^{(n)}(0)\rangle + \frac{2\tilde{k}_n}{k + \tilde{k}_n}\langle\tilde{\eta}_N^{(n)}(t)\tilde{\eta}_{N-n}(0)\rangle$$
$$+ \frac{\tilde{k}_n^2}{(k + \tilde{k}_n)^2}\langle\tilde{\eta}_{N-n}(t)\tilde{\eta}_{N-n}(0)\rangle . \tag{4.53}$$

Which are quickly solved as:

$$\tilde{k}_n = \frac{k}{n} ; \tag{4.54}$$

$$\gamma_n^a = \frac{(2n + 1)(n + 1)}{6n}\gamma ; \tag{4.55}$$

$$\tilde{\gamma}_n = \frac{(n + 1)(n - 1)}{6n}\gamma ; \tag{4.56}$$

$$\gamma_n^b = \frac{(2n + 1)(n + 1)}{6n}\gamma . \tag{4.57}$$

This enables us to rewrite equations (4.41) and (4.42) as:

$$\frac{(2n+1)(n+1)}{6n}\gamma\dot{x}_{N-n} + \frac{(n+1)(n-1)}{6n}\gamma\dot{x}_N = -\left(k+\frac{k}{n}\right)x_{N-n}$$
$$+ \frac{k}{n}x_N + kx_{N-n-1} + \tilde{\eta}_{N-n}\,; \tag{4.58}$$

$$\frac{(2n+1)(n+1)}{6n}\gamma\dot{x}_N + \frac{(n+1)(n-1)}{6n}\gamma\dot{x}_{N-n} = -\frac{k}{n}x_N$$
$$+ \frac{k}{n}x_{N-n} + f + \tilde{\eta}_N^{(n)}\,. \tag{4.59}$$

Substitution of equation (4.34) in equation (4.58) yields:

$$\frac{2Nn(N-n)+N}{6n(N-n)}\gamma\dot{x}_{N-n} + \frac{(n+1)(n-1)}{6n}\gamma\dot{x}_N = -\left(\frac{k}{N-n}+\frac{k}{n}\right)x_{N-n}$$
$$+ \frac{k}{n}x_N + \frac{N-n-1}{N-n}\eta'_{N-n-1} + \tilde{\eta}_{N-n} \tag{4.60}$$

This defines a system of two coupled equations for x_N and x_{N-n} which cannot in general be decoupled because the coefficient matrices of $\begin{pmatrix} \dot{x}_{N-n} \\ \dot{x}_N \end{pmatrix}$ and $\begin{pmatrix} x_{N-n} \\ x_N \end{pmatrix}$ are not proportional to one another.

$$\begin{pmatrix} \frac{2Nn(N-n)+N}{6n(N-n)} & \frac{(n+1)(n-1)}{6n} \\ \frac{(n+1)(n-1)}{6n} & \frac{(2n+1)(n+1)}{6n} \end{pmatrix}\gamma\begin{pmatrix} \dot{x}_{N-n} \\ \dot{x}_N \end{pmatrix} =$$
$$-\begin{pmatrix} \frac{1}{N-n}+\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & \frac{1}{n} \end{pmatrix}k\begin{pmatrix} x_{N-n} \\ x_N \end{pmatrix} + \begin{pmatrix} 0 \\ f \end{pmatrix} + \begin{pmatrix} \bar{\eta}_{N-n} \\ \tilde{\eta}_N^{(n)} \end{pmatrix}, \tag{4.61}$$

where $\bar{\eta}_{N-n} = \frac{N-n-1}{N-n}\eta'_{N-n-1} + \tilde{\eta}_{N-n}$.
These equations can be rewritten in the large N limit as

$$\begin{pmatrix} \frac{1}{3} & \frac{\alpha}{6} \\ \frac{\alpha}{6} & \frac{\alpha}{3} \end{pmatrix}N\gamma\begin{pmatrix} \dot{x}_{N(1-\alpha)} \\ \dot{x}_N \end{pmatrix} =$$
$$-\begin{pmatrix} \frac{1}{1-\alpha}+\frac{1}{\alpha} & -\frac{1}{\alpha} \\ -\frac{1}{\alpha} & \frac{1}{\alpha} \end{pmatrix}\frac{k}{N}\begin{pmatrix} x_{N(1-\alpha)} \\ x_N \end{pmatrix} + \begin{pmatrix} 0 \\ f \end{pmatrix} + \begin{pmatrix} \bar{\eta}_{N(1-\alpha)} \\ \tilde{\eta}_N^{(\alpha N)} \end{pmatrix}, \tag{4.62}$$

where we have defined $\alpha = \frac{n}{N}$.

4.6.2 Scaling of a non-homogeneous Rouse Polymer

The effect of a single intermediate dishomogeneity   Let us now suppose that one of the links that compose our polymer has a much greater viscosity than its neighbours, which we leave homogeneous. We wish to investigate this kind of setup because it will give us some insight on how the attached DNA hairpin affects the fluctuations of the linkers and whether or not it decorrelates them.

What we are planning to do is to write two coupled equations as in equation (4.61), namely for the Nth and the $(N-n)^{\text{th}}$ links when the $(N-n+1)^{\text{th}}$ has a much greater viscosity than the others. In what follows we will indicate with Γ as opposed to γ the viscosity of the different link. The setup is shown in figure 4.10.

   Looking at equations (4.47-4.50) it is immediately apparent that the only one which involves the viscosity of an intermediate link is the one for γ_n^a, namely equation (4.48).
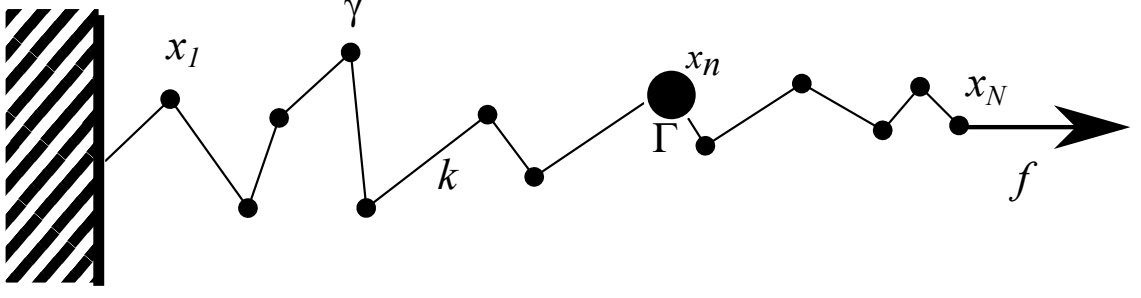
Figure 4.10: A non-homogeneous Rouse polymer composed of N identical springs and $N-1$ beads each having spring constant k and viscous drag coefficient γ. The nth bead is taken to have viscous drag coefficient Γ. The coordinates x_i are taken along the direction of the pulling force f. The node with higher viscosity represents the DNA hairpin to be opened in a typical experiment.

Retracing the steps that brought us to equations (4.61), we have to correct equation (4.55) as:

$$\hat{\gamma}_n^a = \frac{(2n-1)(n-1)}{6n}\gamma + \Gamma\,. \tag{4.63}$$

This change propagates in equation (4.58) but not in equation (4.59), and in turn equation (4.60) becomes:

$$\begin{aligned}
&\left(\frac{2n(N-3)(N-n)+N}{6n(N-n)}\gamma + \Gamma\right)\dot{x}_{N-n} + \frac{(n+1)(n-1)}{6n}\gamma\dot{x}_N = \\
&-\left(\frac{k}{N-n} + \frac{k}{n}\right)x_{N-n} + \frac{k}{n}x_N + \frac{N-n-1}{N-n}\eta'_{N-n-1} + \hat{\eta}_{N-n}\,;
\end{aligned} \tag{4.64}$$

we have to underline that the second noise term has also changed in order to fullfill fluctuation-dissipation relations.

The correlation matrix of the noise terms in equation (4.61) becomes then:

$$\begin{pmatrix} \frac{2n(N-3)(N-n)+N}{6n(N-n)}\gamma + \Gamma & \frac{(n+1)(n-1)}{6n}\gamma \\ \frac{(n+1)(n-1)}{6n}\gamma & \frac{(2n+1)(n+1)}{6n}\gamma \end{pmatrix}\,, \tag{4.65}$$

which can be rewritten in a clearer form in the limit of $N \to \infty$ with $\frac{n}{N} = \alpha$ as:

$$\begin{pmatrix} \frac{N}{3}\gamma + \Gamma & \frac{\alpha N}{6}\gamma \\ \frac{\alpha N}{6}\gamma & \frac{\alpha N}{3}\gamma \end{pmatrix}\,. \tag{4.66}$$

Block polymers Suppose we have a polymer composed of two sections: the first composed of n_1 links of viscosity γ_1 and elasticity k_1, the second of n_2 links of viscosity γ_2 and elasticity k_2. In close resemblance with what we did before we ask ourselves how this modifies the equation for the effective evolution of the floating extremity and of the point where the two sections are linked.

It is important to know this because most DNA unzipping experiments so far have relied on linkers of both single- and double-stranded DNA bonded in heteropolymers of various lengths.

Equation (4.34) involves only links of the first tipe and can thus be easily rewritten with the additional index. We may think that equations (4.58) and (4.59) share the same fate but a different index; unfortunately this is true only of the elasticities. The viscosity of
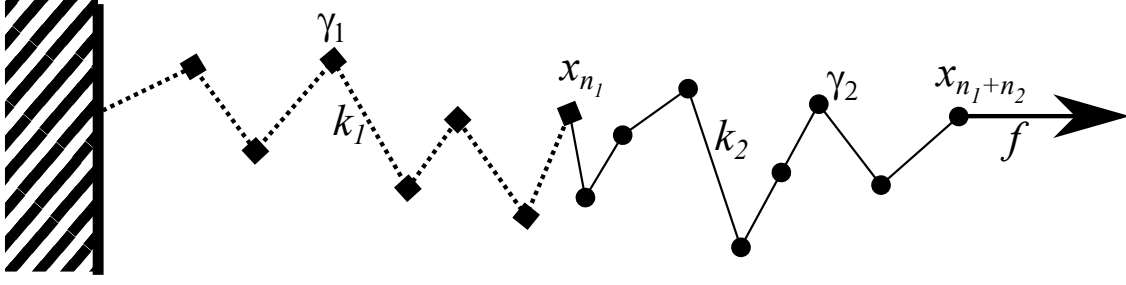
Figure 4.11: A non-homogeneous Rouse polymer composed of n_1 identical springs and beads each having spring constant k_1 and viscous drag coefficient γ_1 and n_2 more identical springs and beads each having spring constant k_2 and viscous drag coefficient γ_2. The coordinates x_i are taken along the direction of the pulling force f. One can imagine momoers of type 1 to be ssDNA and those of type 2 to be dsDNA.

the link that connects the first and the second section (that is the n_1^{th}) is of the first type. Equations (4.58) and (4.59) then become:

$$\left(\frac{(2n_2 - 1)(n_2 - 1)}{6n_2}\gamma_2 + \gamma_1 \right)\dot{x}_{n_1} + \frac{(n_2 + 1)(n_2 - 1)}{6n_2}\gamma_2 \dot{x}_{n_1 + n_2} =$$
$$-\left(k_2 + \frac{k_2}{n_2} \right)x_{n_1} + \frac{k_2}{n_2}x_{n_1 + n_2} + k_1 x_{n_1 - 1} + \tilde{\eta}_{n_1} \tag{4.67}$$

$$\frac{(2n_2 + 1)(n_2 + 1)}{6n_2}\gamma_2 \dot{x}_{n_1 + n_2} + \frac{(n_2 + 1)(n_2 - 1)}{6n_2}\gamma_2 \dot{x}_{n_1} =$$
$$-\frac{k_2}{n_2}x_{n_1 + n_2} + \frac{k_2}{n_2}x_{n_1} + f + \tilde{\eta}_{n_1 + n_2}^{(n_2)} . \tag{4.68}$$

Putting all back together gives us two matrices: one for the γ's and the other for the k's:

$$\left(\begin{array}{cc} \frac{(2n_1 + 1)(n_1 + 1)}{6n_1}\gamma_1 + \frac{(2n_2 - 1)(n_2 - 1)}{6n_2}\gamma_2 & \frac{(n_2 + 1)(n_2 - 1)}{6n_2}\gamma_2 \\ \frac{(n_2 + 1)(n_2 - 1)}{6n_2}\gamma_2 & \frac{(2n_2 + 1)(n_2 + 1)}{6n_2}\gamma_2 \end{array} \right) , \tag{4.69}$$

$$\left(\begin{array}{cc} \frac{k_1}{n_1} + \frac{k_2}{n_2} & -\frac{k_2}{n_2} \\ -\frac{k_2}{n_2} & \frac{k_2}{n_2} \end{array} \right) , \tag{4.70}$$

the former can be rewritten in the limit of n_1, n_2 large as:

$$\left(\begin{array}{cc} \frac{n_1}{3}\gamma_1 + \frac{n_2}{3}\gamma_2 & \frac{n_2}{6}\gamma_2 \\ \frac{n_2}{6}\gamma_2 & \frac{n_2}{3}\gamma_2 \end{array} \right) . \tag{4.71}$$

Validity of the approximation In the beginning of this section we have stated that the microscopic time-scale for ssDNA is given by $\gamma/k = 10^{-10}$ s. Now by looking at the scaling of the mesoscopic timescale we'll obtain the range of validity of our aproximation, that is the timescale at which a polymer will continue to behave as a single entity and the propagation time along the polymer will be negligible. The scaling of the macroscopic time is proportional to $\gamma/k N^2/3$ where N is the number of monomers.

Given that as of today the state of the art in experiments the maximum sampling frequency is of the order of 10 kHz, polymers larger than 1000 base pairs have relaxation times that are observable.

To take this into account in mesoscopic simulations we can split long polymer into smaller pieces even though this doesn't appear to have an appreciable effect on measured relaxation times. The interested reader should refer to section 5.1 of [Barbieri 09].

Moreover in [Barbieri 09, Appendix A], a more formal discussion of the normal modes of a single homogeneous polymer is given. It turns out that the factor $1/3$ that appears in our equations is an approximation of the true factor $\pi^2/4$ that would appear in an exact treatment. Here we have preferred to give this approximated result is the only one that can yield the scaling for the off-diagonal terms, and extends well to non-homomgeneous polymers.

4.6.3 Detailed balance

Now that we have described all the different pieces of the dynamics of DNA unzipping we would like to derive a coupled mesoscopic dynamics that respects detailed balance equations with the right thermodynamic equilibrium distribution:

$$P(n, x) = e^{-\beta W(x,n) - \beta G_0(n;B)}/Z, \tag{4.72}$$

where $G_0(n; B) = \sum_i^n g_0(b_i, b_{i+1})$ is the binding energy of the fork, and $W(x, n)$ is the free energy of the linkers, the beads and the traps, but we need not concentrate on the details for now.

This is not a trivial task because we have to take into account the coupling between a continuous time Markov chain (the fork dynamics n), and the Brownian dynamics of the polymers and the beads.

Let us first identify the possible events at each time step, the fork can either open, close or stay where it is at each time step, and the x variable will perform a Langevin step of size Δx.

We have identified three transitions that correspond to three detailed balance equations:

$$P(n, x)H_o(x, n, \Delta x) = P(n + 1, x + \Delta x)H_c(n + 1, x + \Delta x, -\Delta x); \tag{4.73}$$

$$P(n, x)H_c(x, n, \Delta x) = P(n - 1, x + \Delta x)H_o(n - 1, x + \Delta x, -\Delta x); \tag{4.74}$$

$$P(n, x)H_s(x, n, \Delta x) = P(n, x + \Delta x)H_s(n, x + \Delta x, -\Delta x); \tag{4.75}$$

where o, c and s denote respectively open, close and stay, and H are the transition rates. If we now suppose, as we have discussed previously, that the opening rate depends exclusively on the binding energy, and we further impose it to be a product of the opening rate and a Langevin step we get:

$$H_o(x, n, \Delta x) = r\Delta t \, e^{\beta G(n;B) - \beta G(n+1;B)}$$
$$\times \sqrt{\frac{4\pi T \Delta t}{\gamma_n}} \exp\left[-\frac{\gamma_n}{4T\Delta t}\left(\Delta x - \frac{f(x, n)\Delta t}{\gamma_n}\right)^2\right], \tag{4.76}$$

that is consistent with the definition of r_o defined previously if we integrate over Δx. This, in conjunction with equation (4.74) gives the closing rate:

$$H_c(x, n, \Delta x) = r\Delta t \, e^{\beta W(x,n) - \beta W(x+\Delta x, n-1)}$$
$$\times \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}} \exp\left[-\frac{\beta\gamma_{n-1}}{4\Delta t}\left(\Delta x + \frac{f(x + \Delta x, n - 1)\Delta t}{\gamma_{n-1}}\right)^2\right]. \tag{4.77}$$

The problem is that this rate depends on quantities computed both in x and $x + \Delta x$ and it is not Gaussian for general f. On the other hand if we impose $f(x, n) = -\frac{\partial W}{\partial x}$ and we

perform a Taylor of the terms that are calculated in $x + \Delta x$ expansion at the exponent we get for the W part:

$$
\begin{aligned}
&W(x,n) - W(x + \Delta x, n-1) = \\
&W(x,n) - W(x, n-1) + W(x, n-1) - W(x + \Delta x, n-1) = \\
&W(x,n) - W(x, n-1) + f(x, n-1)\Delta x - \frac{\partial^2 W(x, n-1)}{\partial x^2}(\Delta x)^2 + O((\Delta x)^3) \, ;
\end{aligned}
\tag{4.78}
$$

and for the Brownian step:

$$
\begin{aligned}
&-\frac{\gamma_{n-1}}{4\Delta t}\left(\Delta x + \frac{f(x + \Delta x, n-1)\Delta t}{\gamma_{n-1}}\right)^2 = \\
&-\frac{\gamma_{n-1}}{4\Delta t}\left(\Delta x - \frac{f(x + \Delta x, n-1)\Delta t}{\gamma_{n-1}}\right)^2 - f(x + \Delta x, n-1)\Delta x = \\
&-\frac{\gamma_{n-1}}{4\Delta t}\left(\Delta x - \frac{f(x, n-1)\Delta t}{\gamma_{n-1}}\right)^2 - f(x, n-1)\Delta x + \frac{\partial^2 W(x, n-1)}{\partial x^2}(\Delta x)^2 \\
&+ O(\Delta t \Delta x) + O((\Delta x)^3) \, .
\end{aligned}
\tag{4.79}
$$

Now we only have to notice that terms up to and including order Δx cancel out and that for Brownian motion $\Delta t \sim (\Delta x)^2$, to see we can rewrite the rate as:

$$
\begin{aligned}
H_c(x, n, \Delta x) = r\Delta t \, e^{\beta W(x,n) - \beta W(x, n-1)} \\
\times \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}} \exp\left[-\frac{\beta \gamma_{n-1}}{4\Delta t}\left(\Delta x - \frac{f(x + \Delta x, n-1)\Delta t}{\gamma_{n-1}}\right)^2\right] \, ,
\end{aligned}
\tag{4.80}
$$

which is now consistent with the definition of r_c by integrating over Δx.

The attentive reader should note that the force in the Brownian step is computed in $n-1$, that is once the base has been closed, this has important consequences on the implementation of the algorithm.

Finally, the rate at constant n is obtained by imposing that:

$$
\int d\Delta x H_s(x, n, \Delta x) + H_o(x, n, \Delta x) + H_c(x, n, \Delta x) = 1 \, ,
\tag{4.81}
$$

that is:

$$
\begin{aligned}
H_s(x, n, \Delta x) = [1 - r_o(x, n) - r_c(x, n)] \\
\times \sqrt{\frac{4\pi T \Delta t}{\gamma_n}} \exp\left[-\frac{\gamma_n}{4T\Delta t}\left(\Delta x - \frac{f(x, n)\Delta t}{\gamma_n}\right)^2\right] \, ,
\end{aligned}
\tag{4.82}
$$

Now the algorithm can be summarized:

```
p=randomreal()
if(p<r_open(n,x)){
    x+=f(x,n)*dt/gamma+randomgaussian()*sqrt(2*beta*dt*gamma)
    n++
}
else if(p<r_open(n,x)+r_close(n,x)){
    n--
    x+=f(x,n)*dt/gamma+randomgaussian()*sqrt(2*beta*dt*gamma)
}
else{
    x+=f(x,n)*dt/gamma+randomgaussian()*sqrt(2*beta*dt*gamma)
}
```

Note how the order of the Brownian step and the opening or closing is reversed, as we have underlined before this is essential to the satisfaction of detailed balance equations.

4.7 Results from the dynamical model

We have spent the best part of the previous chapter defining an effective mesoscopic dynamical model for DNA micromanipulation experiments. Our approach is much more complex than separately simulating fork and polymer dynamics: first because it does not imply equilibrium and secondly because it allows for cross-correlation effects between fork, beads and polymers dynamics.

In this section we wish to turn to the novel measurements that we have been able to perform thanks to this software and that were published in [Barbieri 09].

First of all we have observed that for complex polymers the expression $W(l) = Nw(l)$ for the free energy breaks down at low N. This was immediately clear when we observed that the measured sojourn times did not match the theoretical prediction from the Boltzmann distribution, however, the effect is much smaller even simply adding the nonlinear dependence in N coming from the square root term in:

$$e^{-\beta W(x,n)} = e^{-\beta Nw(x/N)}\sqrt{\frac{\beta k(x/N)\ell^2}{2\pi N}} \, , \qquad (4.83)$$

where ℓ is a dimensional constant of no importance, and k was defined previously as the second derivative of w with respect to $l = x/N$.

In figure 4.12 we show the effect of the square-root term in the case of an uniform sequence: the time spent on a basis is obtained by simulation with and without the square-root term and by its Boltzmann estimate.

Another set of quantities which is in general not available from first principles computations are correlation functions, in [Barbieri 09] we have studied in detail the dependence of the correlation functions on the number of open bases and on the length of the linkers in various experimental setups in order to determine the importance of out of equilibrium effects such as propagation times.

In the following we will concentrate on a setup similar to that used in Bockelmann's lab at ESPCI: two optical traps of stiffness 0.1 pN/nm and 0.512 pN/nm respectively, a dsDNA linker of 3120 bases and a ssDNA linker of 40 bases.

We have found that polymers which are shorter than 1000 bases show no appreciable effect due to finite relaxation times. For longer polymers we have devised a way of introducing the propagation effect: we cut up the polymer in pieces which are at most 1000 bases long and we simulate them separately.

This is shown in figure 4.13. However we have found this to have an effect on the shape of the correlation function, but not on the correlation time. In fact, if the correlation function is fit with a stretched exponential of the form: $\exp[-(t/\tau)^\beta]$, β is slightly smaller.

We have also been able to study the dependence of the relaxation time for different parts of the setup as a function of the number of open bases and to compare those with theoretical results. This is shown in table 4.2 and in figure 4.14.

The results are in very good numerical agreement except for the two beads: it turns out that the relationship between the bead and the number of open bases is more subtle than we thought. It appears that there is a very strong correlation between the fork and the bead and this effect is stronger when the optical trap is softer.

This effect is desirable, it is in fact the effect that allows us to gain information on the sequence. To better quantify the relationship between the stiffness of the trap and the
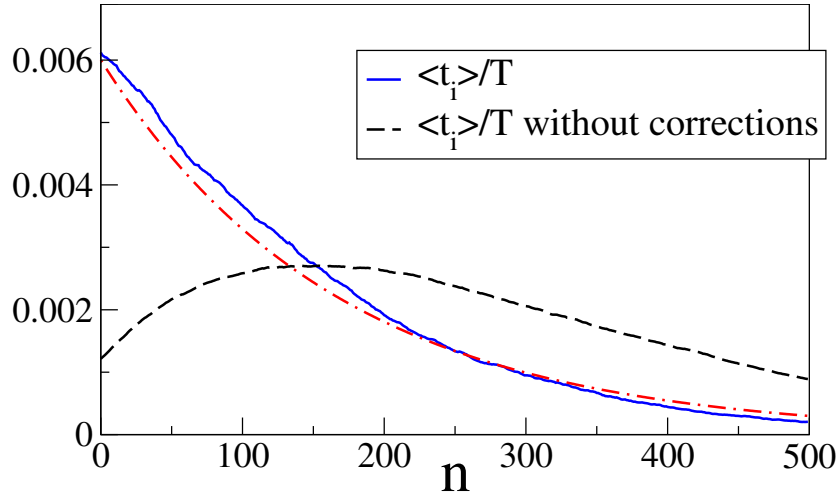
Figure 4.12: Average fraction of time spent on each base. The full (blue) curve corresponds to Eq. (4.83) while the dashed (black) curve corresponds Eq. (4.83) without the saddle-point corrections (the square-root term). The dot-dashed (red) line is $P_{\text{eq}}(n) \propto \exp[-n\,\Delta g]$ with $\Delta g = 0.006$. n is the number of open bases.

	Theoretical (s)	Numerical (s)
Single strand	$4.83 \cdot 10^{-11} N_{\text{eq}}^2$	$5.4 \cdot 10^{-11} N_{\text{eq}}^2$
Double strand	$4.96 \cdot 10^{-5}$	$\sim 3 \cdot 10^{-5}$
Spring x_1	$1.67 \cdot 10^{-4}$	$\sim 1.5 \cdot 10^{-3}$
Spring x_4	$3.26 \cdot 10^{-4}$	$\sim 7 \cdot 10^{-5}$
Fork N_{eq}	$\propto 14.2 + 0.013 N_{\text{eq}}$	$1.3 \cdot 10^{-3} + 8.4 \cdot 10^{-7} N_{\text{eq}}$

Table 4.2: Comparison between the correlation times of the setup in figure 4.4A as computed for an isolated element and the result of a complete numerical simulation. In the case of the fork, we reported as theoretical value $1/k_{\text{eff}}$, that must be multiplied by a viscosity to obtain the relaxation time; it turns out that a viscosity $\sim 8 \cdot 10^{-5}$ pN s/nm matches the theoretical and numerical results.

Figure 4.13: Correlation functions for the setup in figure 4.4A at two different values of the number of open bases, $N_{eq} = 40 + n$. ss and ds indicate the autocorrelation functions of the ssDNA and dsDNA linkers and x_1 and x_4 are the autocorrelation functions of two optical traps of different stiffnesses (x_4 being the stiffest).
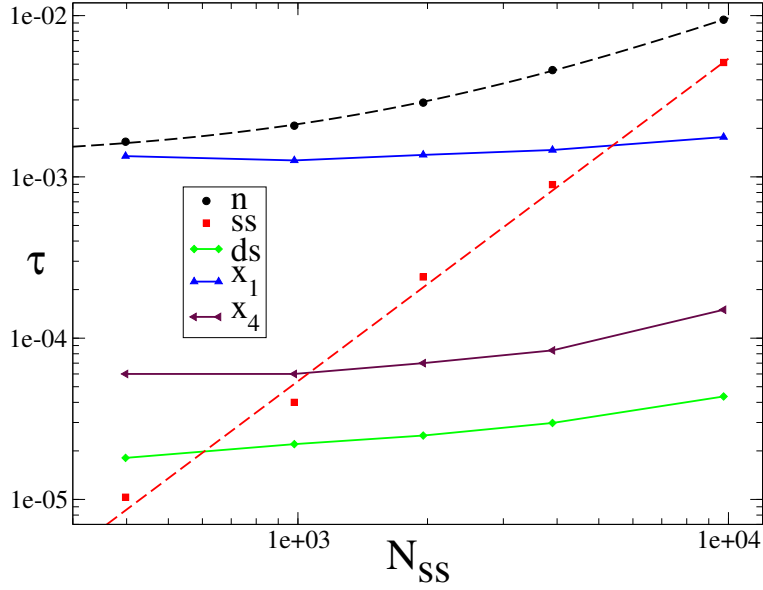
Figure 4.14: Relaxation times of the correlation functions in figure 4.13 as a function of the number of open bases. In the case of the single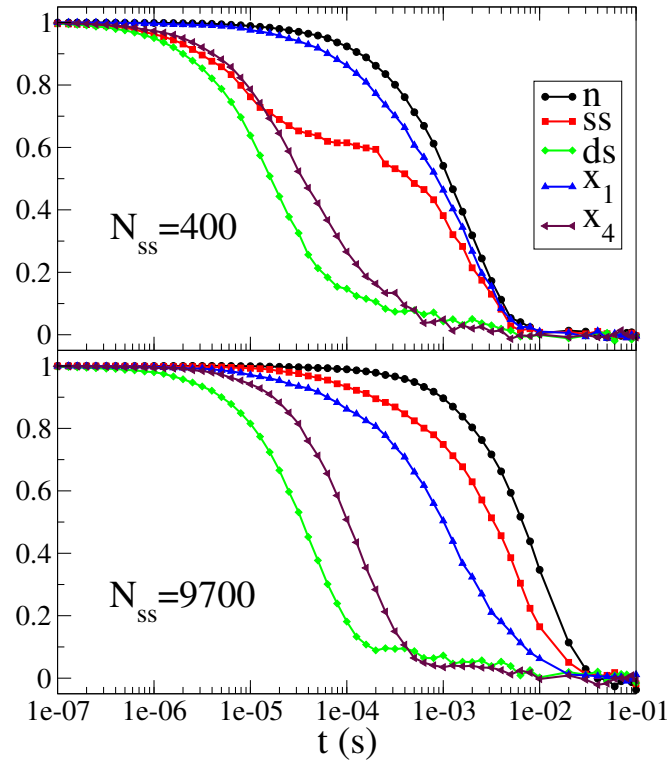 strand (ss), only the fast relaxation time is plotted. For the fork and the single strand, dashed lines indicate a fit to $\tau_n = A + BN_{eq}$ (with $A = 1.3 \cdot 10^{-3}$ and $B = 8.4 \cdot 10^{-7}$) and $\tau_{eq} = CN_{eq}^2$ (with $C = 5.4 \cdot 10^{-11}$ s). For the others, full lines are guides to the eye.

correlation of the bead with the fork we have defined the quantity:

$$I(x_4, n) = \sum_n \int dx_4 P(x_4, n) \log \left(\frac{P(x_4, n)}{P(x_4)P(n)} \right) , \qquad (4.84)$$

as the mutual information between the fork and one of the two beads.

In figure 5.19 we show the effect on the stiffness of the optical trap on the mutual information between the fork position and the bead position. We find that softer beads yield more information on the sequence. This can be intuitively understood by thinking that a softer trap gives way more easily to the excess length deriving from the opening of a base. It must be stressed, however, that this result holds only per measure, that is if one wanted to know if it were more efficient to have more rigid traps in an experiment one should take into account the autocorrelation times of the bead position. Those are in fact lower for stiffer traps allowing for a larger number of statistically independent measures per unit time.

Figure 4.15: Mutual information I between x_4 and n as a function of the trap stiffness, k. Black circles are computed on an uniform sequence, while red squares are measured on a sawtooth potential derived from a sequence that alternates stretches of 10 weak bases and stretches of 10 strong bases.

Chapter 5

Inferring the DNA sequence

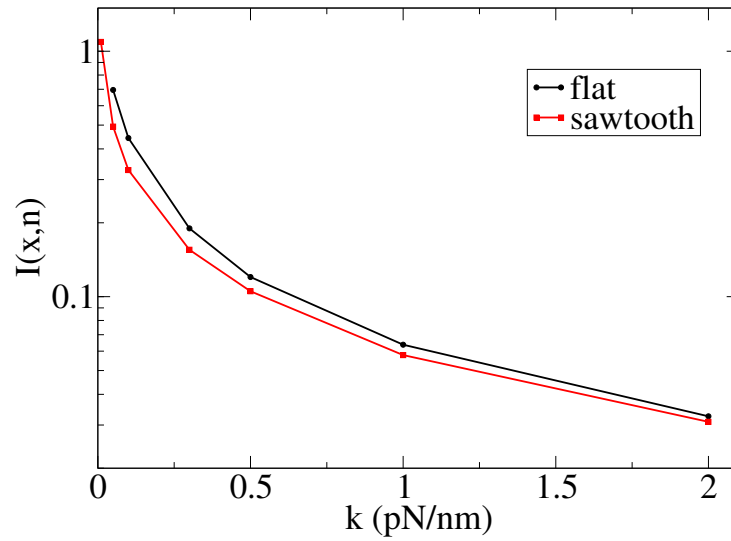As we have shown in the previous section, DNA unzipping experiments show a remarkable dependence on sequence in the force-extension signal. Several attempts have been made to reconstruct the free energy landscape from different experimental setups [Danilowicz 03, Woodside 06a, Huguet 09].

In this section we will concentrate on algorithmic and mathematical approaches to solving the inverse problem, that is characterizing the free energy landscape as a function of n and eventually sequencing DNA.

Idealized cases, where the number of open bases n is known at all time, are relatively easy to solve, but once we start adding the layers of complexity of real experiments, it becomes really difficult to extract information about the sequence.

The first algorithm we will describe is based on the very idealized situation we have just described: infinite sampling frequency and knowledge of the number of open bases.

The second supposes we can access the equilibrium value of physical quantities like the position of the beads with arbitrary precision, ignoring fluctuations. This is much more realistic than the first approach, but results are much farther from reconstructing the sequence.

Lastly we will take a look at a toy model based on the Ornstein-Uhlenbeck model, that takes into account correlations and finite sampling frequencies.

5.1   Infinite bandwidth algorithm

In what follows we suppose we have access to the results of a fixed-force experiment where the position of the fork is known at all times. Since this is not a realistic situation, the data on which to perform the inference must be simulated.

Let us suppose we have perfect knowledge at all times of the number of open bases. It is clear that this is not realistic at all: first of all because the number of open bases n is not directly measurable and secondly because in order to obtain bandwidths that are large compared to the elementary event time-scale we would need a resolution of the order of the MHz or more and current experimental setups allow for resolutions three orders of magnitude smaller.

The details of what follows were first published in [Baldazzi 06, Baldazzi 07].

In the previous chapter we have defined the opening and closing rate: respectively $r_\mathrm{o}(n)$ and $r_\mathrm{c}(f)$. For small enough time intervals Δt we can write:

$$P(n(t+\Delta t)|n(t)) = \begin{cases} \Delta t r_\mathrm{o}(n(t)) & \text{for } n(t+\Delta t) = n(t)+1\,; \\ \Delta t r_\mathrm{c}(f) & \text{for } n(t+\Delta t) = n(t)-1\,; \\ 1 - \Delta t r_\mathrm{o}(n(t)) - \Delta t r_\mathrm{c}(f) & \text{for } n(t+\Delta t) = n(t)\,; \\ o(\Delta t) & \text{otherwise.} \end{cases} \qquad (5.1)$$

This defines completely the transition probabilities from one state to another, and it can be used to define the probability of a the outcome of an experiment, that is of a complete trace. In order to do so we must define the relevant variables:

- t_n the total time spent with n open bases;

- u_n the number of transitions from n to $n + 1$;

- d_n the number of transitions from n to $n - 1$.

Given those definitions one can write the probability of an experimental trace as $\mathcal{T}$, conditioned on the sequence $\mathcal{B}$ and on the external force f, as:

$$
\begin{aligned}
P(\mathcal{T}|\mathcal{B}) &= \prod_n (\Delta t r_o(n(t)))^{u_n} (\Delta t r_c(f))^{d_n} (1 - \Delta t r_o(n(t)) - \Delta t r_c(f))^{t_n/\Delta t} \\
&= C(\mathcal{T}) \prod_n M(b_n, b_{n+1}; u_n, t_n) \,.
\end{aligned} \tag{5.2}
$$

where we have separated the part that depends on the sequence from that who does not, thus defining:

$$
C(\mathcal{T}) = (\Delta t)^{u+d} \exp(-t_{\text{tot}} r_c(f)); \tag{5.3}
$$

$$
M(b_n, b_{n+1}; u_n, t_n) = \exp\left(g_0(b_n, b_{n+1}) u_n - r e^{g_0(b_n, b_{n+1})} t_n\right); \tag{5.4}
$$

$$
\tag{5.5}
$$

where we have used the definition of r_o and we have defined $u = \sum_n u_n$, $d = \sum_n d_n$, and $t_{\text{tot}} = \sum_n t_n$.

Now we can use Bayes' theorem to compute the probability of a sequence given a trace:

$$
P(\mathcal{B}|\mathcal{T}) = \frac{P(\mathcal{T}|\mathcal{B})P(\mathcal{B})}{P(\mathcal{T})} \,. \tag{5.6}
$$

We can further assume (though it is not generally true) that all sequences are equiprobable that is $P(\mathcal{B})$ is uniform, this will lead us to a first rough estimate of the sequence given a trace.

We can maximize the expression we have given for $P(\mathcal{T}|\mathcal{B})$ over the $g_0(b_n, b_{n+1})$ without imposing that it can only take ten values to get a maximum likelihood estimate:

$$
g_0(b_n, b_{n+1}) = \log\left(\frac{u_n}{r t_n}\right) \,, \tag{5.7}
$$

This computation is not bad as a first estimate, but it amounts to searching in a continuous space when we effectively have only 4 possible values for a base. In order to find the most likely sequence $\mathcal{B}^*$ we can use the Viterbi algorithm [Viterbi 67, MacKay 05].

The procedure is as follows: let us consider the first two bases and let us define $P_2(b_2) = \max_{b_1} M(b_1, b_2; u_1, t_1)$, then $b_1^{\max}(b_2) = \arg\max_{b_1} M(b_1, b_2; u_1, t_1)$, and for $n \neq 1$ we can write:

$$
P_{n+1}(b_{n+1}) = \max_{b_n} M(b_n, b_{n+1}; u_n, t_n) P_n(b_n) \,; \tag{5.8}
$$

$$
b_n^{\max}(b_{n+1}) = \arg\max_{b_n} M(b_n, b_{n+1}; u_n, t_n) P_n(b_n) \,; \tag{5.9}
$$

this means that the optimal value for a base depends on the choice for the next base. We can solve these equations up to the last $P_N(b_N)$ which is maximized to obtain $b_N^* =$
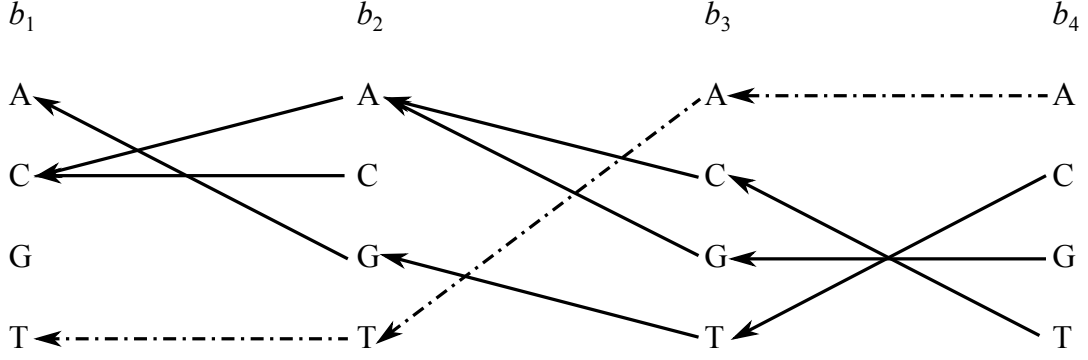
Figure 5.1: We start by choosing $b_1^{\max}(b_2)$ which amounts to choosing the best b_1 for each choice of b_2 and can be represented by an arrow going from b_2 to b_1 and then we iterate the procedure until we get to b_N (here $N = 4$). It is now possible to compute the optimum b_N, in this case A and propagate back to obtain the optimal sequence TTAA.

$\arg\max_{b_N} P_N(b_N)$ and we can then propagate back to the first value setting $b_n^* = b_n^{\max}(b_{n+1}^*)$. The algorithm is explained graphically in figure 5.1.

What is great about Viterbi algorithm is that its complexity grows linearly in N and one needs to explore only a very small subset of the 4^N possible sequences. This is a feature of message-passing algorithms in one dimension.

Another interesting feature of this framework is that unzipping experiments can be repeated several times and the different traces can be combined just by computing the product of probabilities:

$$P(\mathcal{T}_1, \mathcal{T}_1, \ldots, \mathcal{T}_M | \mathcal{B}) = \prod_{i=1}^{N} P(\mathcal{T}_i | \mathcal{B}), \tag{5.10}$$

where $\mathcal{T}_i$ is the trace of the ith experiment of a series of M.

Therefore we can combine different experiments to infer the sequence. In [Baldazzi 07] it has also been shown that the rate of error decreases exponentially with the number of measurements.

As we have said at the beginning of this section, however, this algorithm relies on two unrealistic assumptions: knowledge of the position of the fork, which is never attainable because we actually measure the position of the bead; and an infinite sampling frequency. In the following we will try to come over these two assumptions by building more complex inference algorithms.

5.2 Perfect averages algorithm

In this section we will perform a few simplifying assumptions in order to keep the equations simple looking. The reader should note, however, that these simplifications are by no mean fundamental and our results will hold even after relaxing those assumptions.

The first assumption is that we substitute Gaussian polymers for the complex behavior described in the preceding chapter and the second is that we ignore the n dependence of the spring constants. The first assumption is not of fundamental importance because it amounts to truncating the anharmonic effects in the probabilities; relaxing it would only force us to compute integrals numerically, slowing the computation down.

The second assumption is even easier to relax because the n dependence will just change the variance of the different terms in the sum in the next equation.

In general we believe that what is most important here is to have a general idea of what can and cannot be done with the spring constants set at realistic values for today's experiments. We will show that even without complex polymers and n dependence we cannot investigate the sequence at a single base level.

We define a function $\bar{u}(L|B)$ as the equilibrium average displacement of one of the beads from the center of its optical trap. L is the distance between the traps and is a parameter of the experiment and B denotes the sequence. The dependence on B will be omitted from now on.

The function $\bar{u}(L)$ has an explicit expression in terms of $g_0(n)$, that is:

$$
\begin{aligned}
\bar{u}(L) = \frac{1}{Z(B)} \sum_{n}^{N} (L - nl) \frac{k_2 k}{k_1 k_2 + k_1 k + k_2 k} \\
\times \frac{\exp\left(-\sum_{j}^{n} g_0(j) - \frac{k_1 k_2 k}{2(k_1 k_2 + k_1 k + k_2 k)}(L - nl)^2\right)}{\sqrt{k_1 k_2 + k_1 k + k_2 k}} ,
\end{aligned}
\tag{5.11}
$$

where $k_1 = 0.025$ nm^{-2} and $k_2 = 0.125$ nm^{-2} are the spring constants of the traps; $k = 0.025$ nm^{-2} is the spring constant of the linkers and the open part of the DNA and may depend weakly on n; N is the total number of bases. $l = 1$ nm is the difference in length when a base is open (two ssDNA bases, one for each side). g_0 is the binding energy of the DNA and it's given in table 4.1

For any given value of n, the number of open basis there is a characteristic length of the fluctuations of u, which corresponds to the width of the gaussian in (5.11). This length is given by:

$$
b = \frac{1}{l} \sqrt{\frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k}} ,
\tag{5.12}
$$

the reader should note that spring constants are expressed so that energies are dimensionless, that is as $k = \beta\kappa$ where β is the inverse temperature and κ a spring constant in the conventional units.

In the following (unless otherwise noted), $b = 9.38$ as it was calculated from realistic constants from Bockelmann's experiment as described in [Barbieri 09]. Other references use different setups that yield different numeric values: Woodside et al. [Woodside 06b] have a setup that would corresponds to $b = 6.46$ in the same approximation. Huguet et al. [Huguet 10, Supplementary material] have $b = 8.49$ for their setup.

In figure 5.2 we show two sequences and their corresponding free energy landscape at fixed L and the $u(L)$. The reader should note how for a fixed L, values of n as far apart as 60 bases can be visited with non negligible probability, and most of the times there exist two or more values as far apart as 20 bases which have a high probability of being visited.

If we now define a trial function which depends linearly on a set of coefficients c_i:

$$
g_{\text{trial}}(n|c_i) = \sum_{i=1}^{M} c_i \Omega_{b'}(n - b'i) ,
\tag{5.13}
$$

where $\Omega_{b'}$ is some one-dimensional function of width b', which we do not need to specify now to keep the discussion as general as possible. We should discuss in the following how b' is related to b.

We can also define u_{trial} which is $\bar{u}$ where g_0 has been substituted by g_{trial}, and the cost function:

$$
C(c_i) = \frac{1}{2} \sum_{i=M_0}^{M+M_0} \left(\bar{u}(ibl) - u_{\text{trial}}(ibl)\right)^2 ,
\tag{5.14}
$$
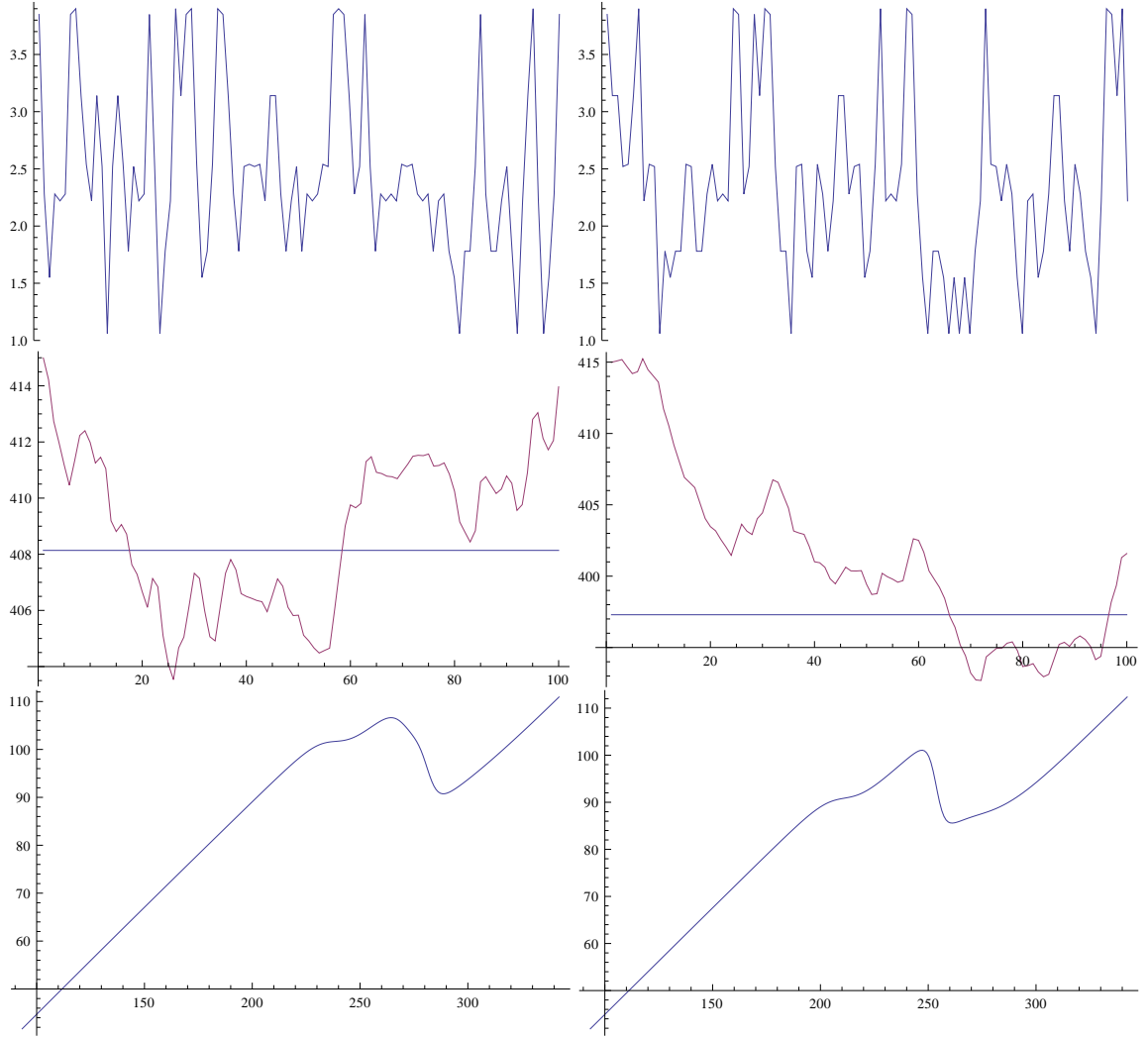
Figure 5.2: Two different random sequences. On top the $g_0(n)$. In the center the free energy defined as $w(n, L) = \sum_{j}^{n} g_0(j) + \frac{k_1 k_2 k}{2(k_1 k_2 + k_1 k + k_2 k)}(L - nl)^2$ as a function of n for $L = 270$, the horizontal line marks the energy level $\tilde{E}$ such as $\exp(-\beta(\tilde{E} - E_0)) = 0.01$, that is sites that are visited (at equilibrium) one hundredth of the time the lowest energy site is. On the bottom the $u(L)$.
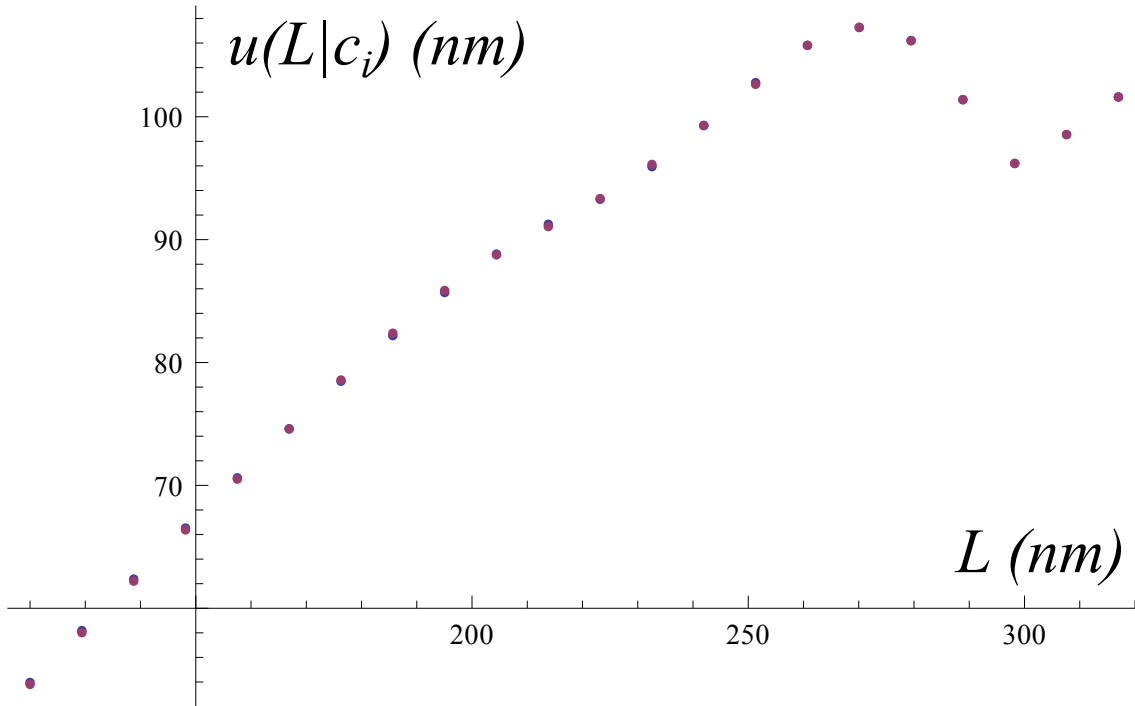
Figure 5.3: $\bar{u}(L)$ (blue) and $u_{\text{trial}}(L|c_i)$ (violet)

where $M_0 = \min_{b,b'}[g_0(b,b')]/k_{\text{eff}}$ and $M_0 + M = \max_{b,b'}[g_0(b,b')]/k_{\text{eff}} + N$. This amounts to taking a measure every bl in the interval where there could be some effect from the sequence, for larger (smaller) i all the bases will be closed (open).

The objective in defining this is to find the set of c_i that approximates the best a set of experimental measures.

$\min_{b,b'}[g_0(b,b')]/k_{\text{eff}} = 93.28$ nm, and $\max_{b,b'}[g_0(b,b')]/k_{\text{eff}} = 343.2$ nm for the set of parameters specified previously. The reader should notice that the difference between these two numbers is rather large compared to the size of one open base pair (1 nm).

In effect most of the times we take many more measures than it is necessary, because for a given sequence the central limit theorem says it is unlikely that such extremes are ever reached, on the contrary the relative fluctuations of the size of the interval of interesting L will scale as $1/\sqrt{N}$.

However, this is not a big computational problem because the computation time will not depend as much on the number of measures, as on the number of parameters (the c_i) which is fixed. On the other hand taking measures in where the response of the system is purely elastic does not change the landscape over which we are optimizing.

It is now possible to minimize the cost function over the c_i.

We will now show some results we have obtained for a random sequence of 50 base pairs and $\Omega_b(x) = \theta(x + b/2)\theta(b/2 - x)$ is the boxcar function of width b. There is very good agreement between $\bar{u}$ and u_{trial}, but if we plot g_0 and g_{trial} the agreement is less good. At some points consecutive values of c_i, that is c_i and c_{i+1}, wander off to values which make it differ greatly from g_0. To quantify the difference between g_0 and $g_{\text{trial}}(n|c_i)$ we can define another cost function:

$$D(c_i) = \sum_n^N (g_0(n) - g_{\text{trial}}(n|c_i))^2. \tag{5.15}$$

It now seems natural to define the set of parameters that minimize this new cost function as d_i and compare the $g_{\text{trial}}(n|c_i)$ and $g_{\text{trial}}(n|d_i)$ as we do in figure 5.4. Where $g_{\text{trial}}(n|d_i)$
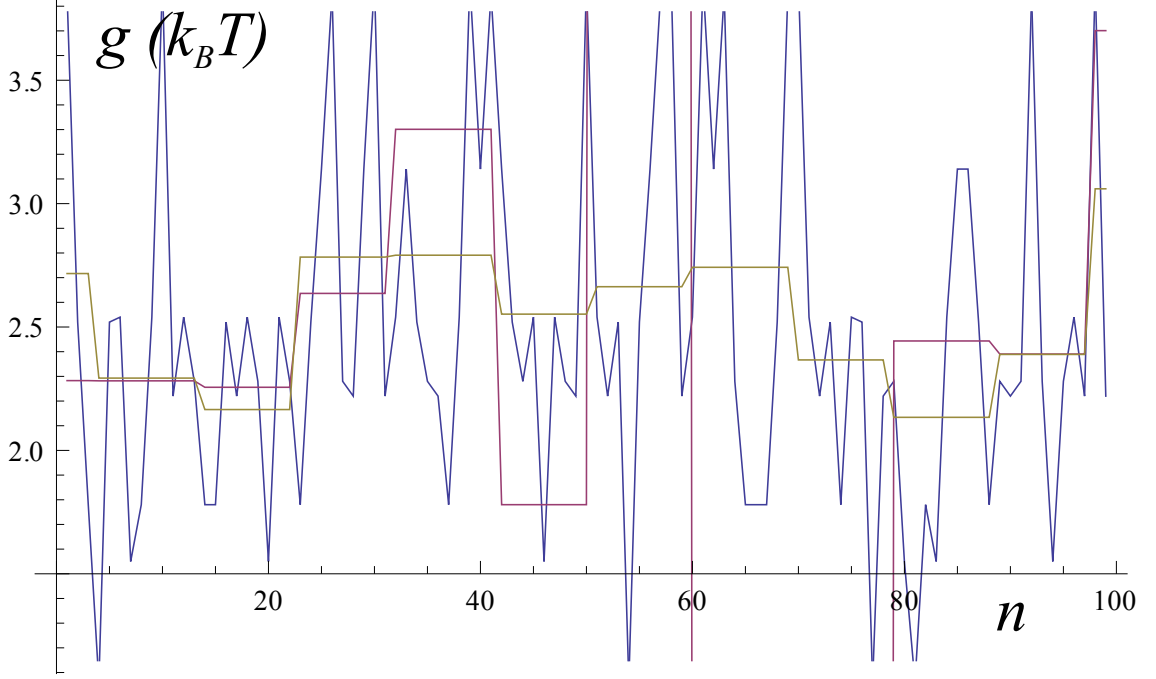
Figure 5.4: $g_0(n)$ (blue), $g_{\text{trial}}(n|c_i)$ (violet) and $g_{\text{trial}}(n|d_i)$ (brown)

is given by:

$$g_{\text{trial}}(n|d_i) = \sum_i^N d_i \Omega_{b'}(n - b'i)\,, \tag{5.16}$$

In practice this amounts to the average of $g_0(n)$ over the step of the trial function, in fact for a given step we have to minimize $\sum_{j=\lceil ib-b/2 \rceil}^{\lfloor ib+b/2 \rfloor}(d_i - g_0(j))^2$, that is:

$$d_i = \frac{1}{|\omega_i|} \sum_{j \in \omega_i}^N g_o(j)\,, \tag{5.17}$$

where $|\omega_i|$ is the cardinality of ω_i, the number of bases that make up a step (it can take either $\lceil b \rceil$ or $\lfloor b \rfloor$ as values).

This way we have shown that $g_{\text{trial}}(n|d_i)$ is a box average of $g_0(n)$ which is very different from a moving average, and since the $g_{\text{trial}}(n|c_i)$ has the exact same structure it makes sense to compare the two.

One might also want to know how $u_{\text{trial}}(L|d_i)$ compares to $\bar{u}(L)$. We can see that in figure 5.5 and the agreement is definitely worse than what it was than when the fit was obtained witht the cost function C.

5.2.1   Prior

As one can see in figure 5.4: two adjacent steps can sometimes grow in opposite directions to non-physical values.

To avoid this kind of problems we have added a prior to center the values of the steps around the average:

$$\tilde{C}_\gamma(c_i) = \frac{1}{2} \sum_{i=M_0}^{M+M_0} (\bar{u}(ibl) - u_{\text{trial}}(ibl))^2 + \gamma \sum_n^N (g_{\text{trial}}(n|c_i) - \bar{g}_0)^2\,, \tag{5.18}$$
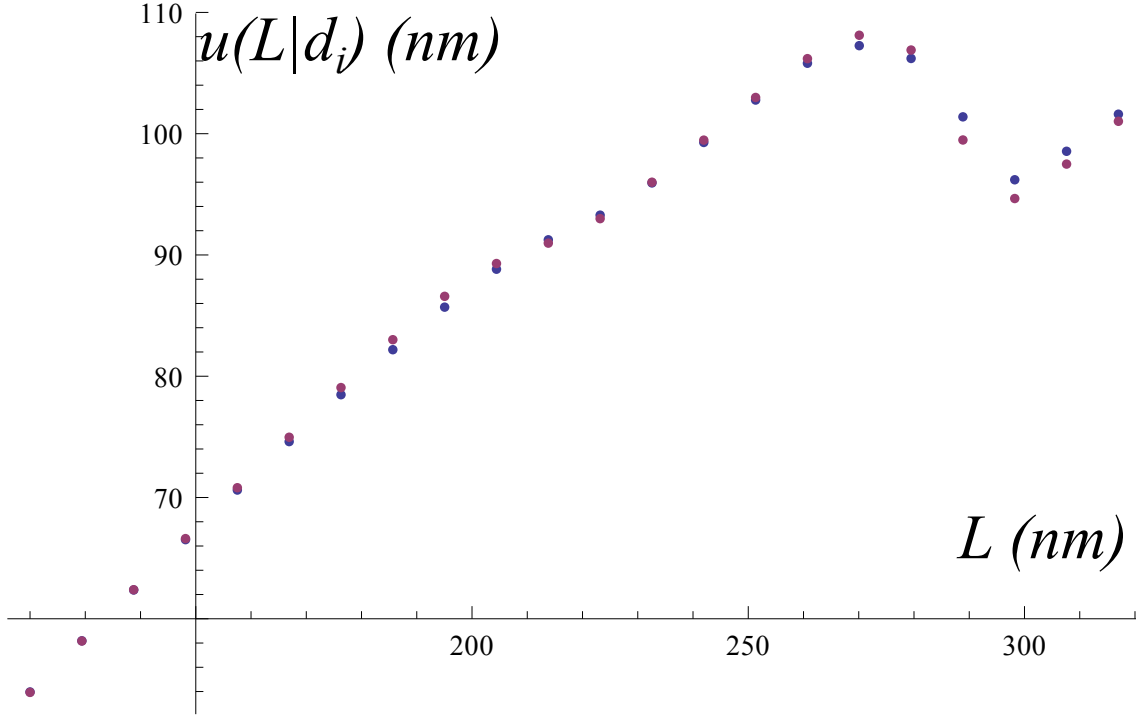
Figure 5.5: $\bar{u}(L)$ (blue) and $u_{\text{trial}}(L|d_i)$ (violet)

Where γ is a constant we use to increase or decrease the effect of the prior. Ideally we hope to obtain a reasonable fit for values of γ smaller than the biggest eigenvalue of the Hessian of C when derived with respect to the c_i's.

The problem is that sometimes we find no good fit no matter the value of γ. This is shown in figure (5.6), as the reader can easily see, the best fit for C does not coincide with the best fit for D. A decreasing D as a function of γ indicates that the best fit is dominated by the prior.

prior, to put it in other words: the best fit is the trivial one: $c_i = \bar{g}_0$ for all i.

Some other times we have a non trivial minumum over γ, and things look definitely better as in figure (5.7).

We have tried a prior that would take into account that the potential g_0 can only take 10 values, so we have chosen the form:

$$-\sum_i^M \sum_j^{10} \exp\left(-\frac{(c_i - g_j)^2}{2\sigma^2}\right) , \qquad (5.19)$$

where the g_j are the ten possible values that g_0 can take. It is important to note that this strategy makes sens only when the trial function has a stepsize of exactly one.

What we have realized is that when b has reasonable values, around those of current state of the art experiments (~ 10), this strategy yields no advantage over the prior we have tried in the preceding section.

At the same time one might think that, for smaller values of b, say when it's closer to one, this prior might help us reconstruct the original sequence, but the reconstruction is actually just as good.

We have yet to find a regime in which this prior makes a difference.

In conclusion we have found that most of the times a small value of γ (i. e. 10^{-4}) gives pretty good results, otherwise there are very clear signs that the fit has not converged.

Figure 5.6: The top two panels show the value of the cost functions C and D as a function of varying γ. The bottom panel shows the $g_{\text{trial}}(n|c_i)$ for $\gamma = 0.0158$ (brown), the real g_0 (blue) and the $g_{\text{trial}}(n|d_i)$ (purple). The value of D/N for the d_i is 0.399.
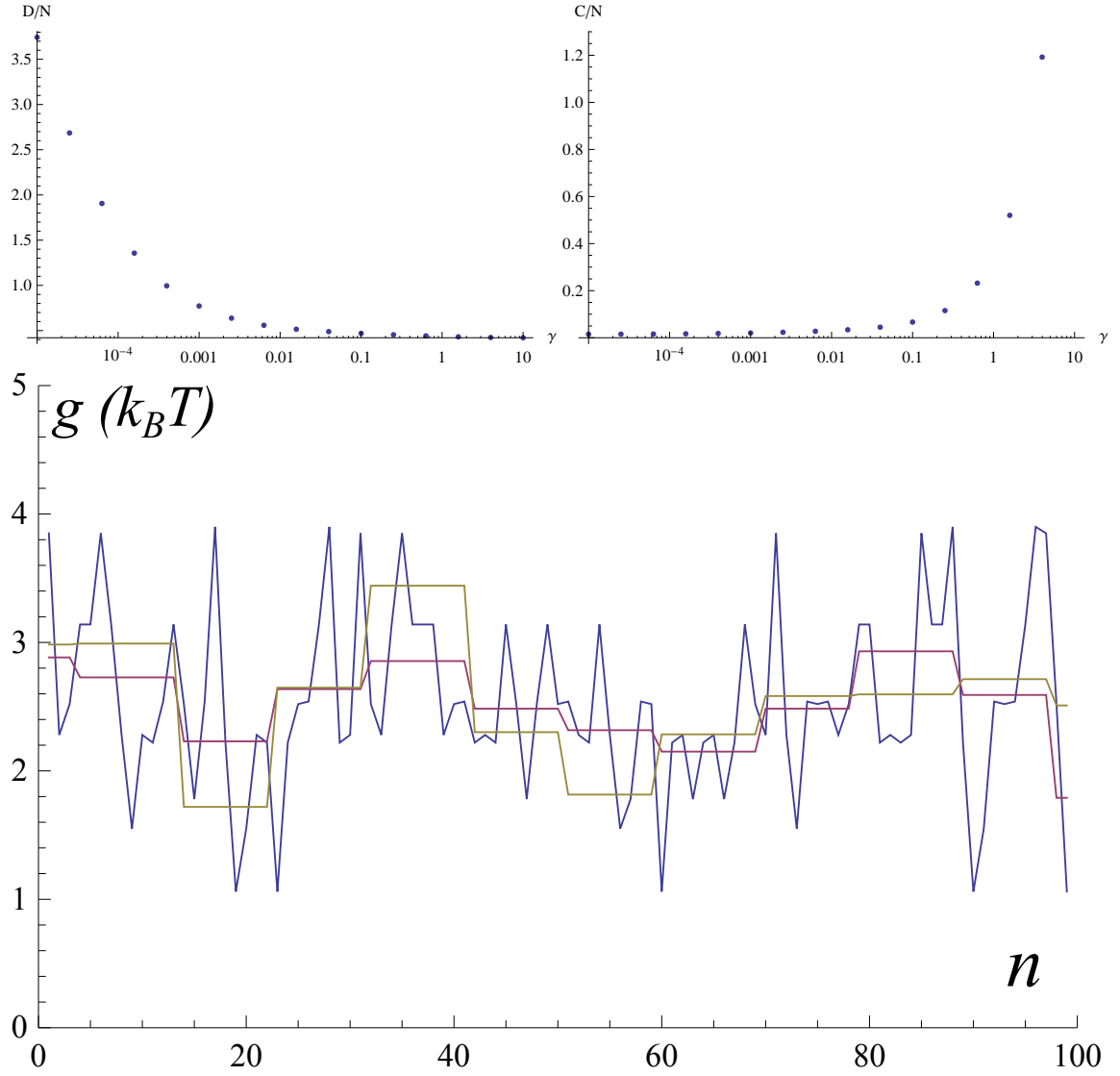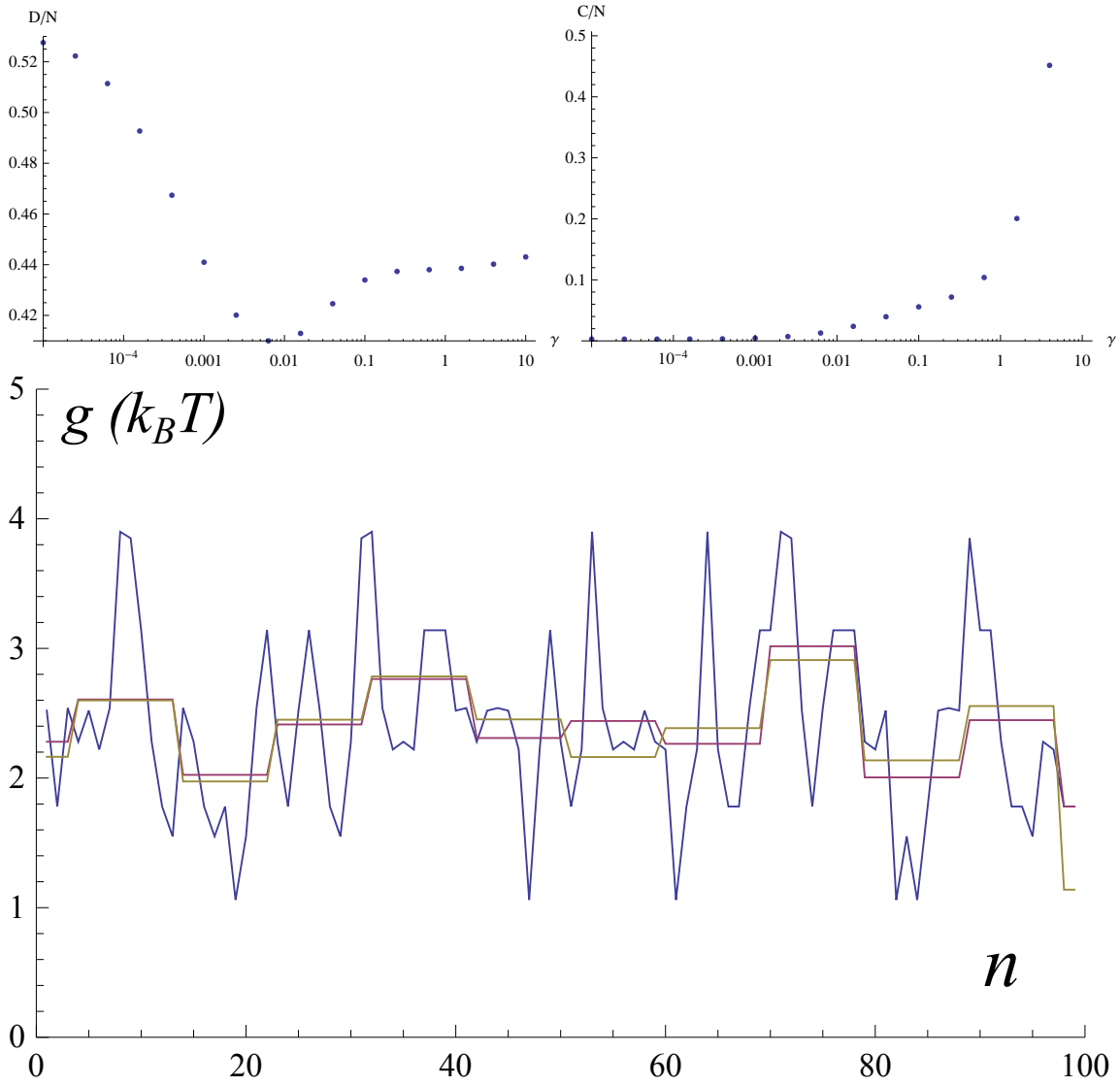
Figure 5.7: The top two panels show the value of the cost functions C and D as a function of varying γ. The bottom panel shows the $g_{\text{trial}}(n|c_i)$ for $\gamma = 0.0063$ (brown), the real g_0 (blue) and the $g_{\text{trial}}(n|d_i)$ (purple). The value of D for the d_i is 0.387.

Figure 5.8: The figure shows the $g_{\text{trial}}(n|c_i)$ for $\gamma = 0.000399$ (brown), the real g_0 (blue) and the $g_{\text{trial}}(n|d_i)$ (purple).

5.2.2   Optimal value of the step-size

The question is whether this can be further ameliorated by choosing a smaller stepsize. If we chose a stepsize $b' = b/2$ we obtain the best fit for $\gamma = 0.000399$ and a value of D/N of 0.316, while the d_i yield $D/N = 0.26$. The results are shown in figure 5.8.

If we further decrease the stepsize to $b/4$ there is not much to be gained: for $\gamma = 0.00016$ we obtain $D/N = 0.343$ which is larger than what we obtained for $b/2$ while the value for the d_i has further decreased to $D/N = 0.16$. The results are displayed in figure 5.9.

We now wish to study more systematically the optimal value of b', to do so we have computed the optimal c_i and d_i for 100 random sequences of 100 base pairs. The results are shown in figure 5.10: $D(d_i)$ gets better and better with smaller stepsize and for $b' = b/8 \simeq 1.17$ it is close to zero. On the other hand $D(c_i)$ seems to taper off to a value of approximately $0.4N$. We can now define another function we can use to evaluate the goodness of fit:

$$E(c_i|d_i) = \sum_{n}^{N} (g_{\text{trial}}(n|c_i) - g_{\text{trial}}(n|d_i))^2 . \tag{5.20}$$

E can be thought of as the distance between the fit of the u (c_i) and the boxed average (d_i), which is the best attainable fit for a give step-size.

We expect E to have a non trivial minimum where $D(c_i)$ starts to saturate, representing the point where the fit obtained through the u is closest to the average over the steps. The results are shown in figure 5.11.

This kind of metric can be a good gauge of what would happen when b is smaller, we have b's which are a half and a quarter of the original. We have obtained this by making l respectively twice and four times as long.

The results for several b' and $b = 4.69$ are shown in figure 5.12and the results for $b = 2.35$ in figure 5.13. Please note that we have excluded points where b' would have been less
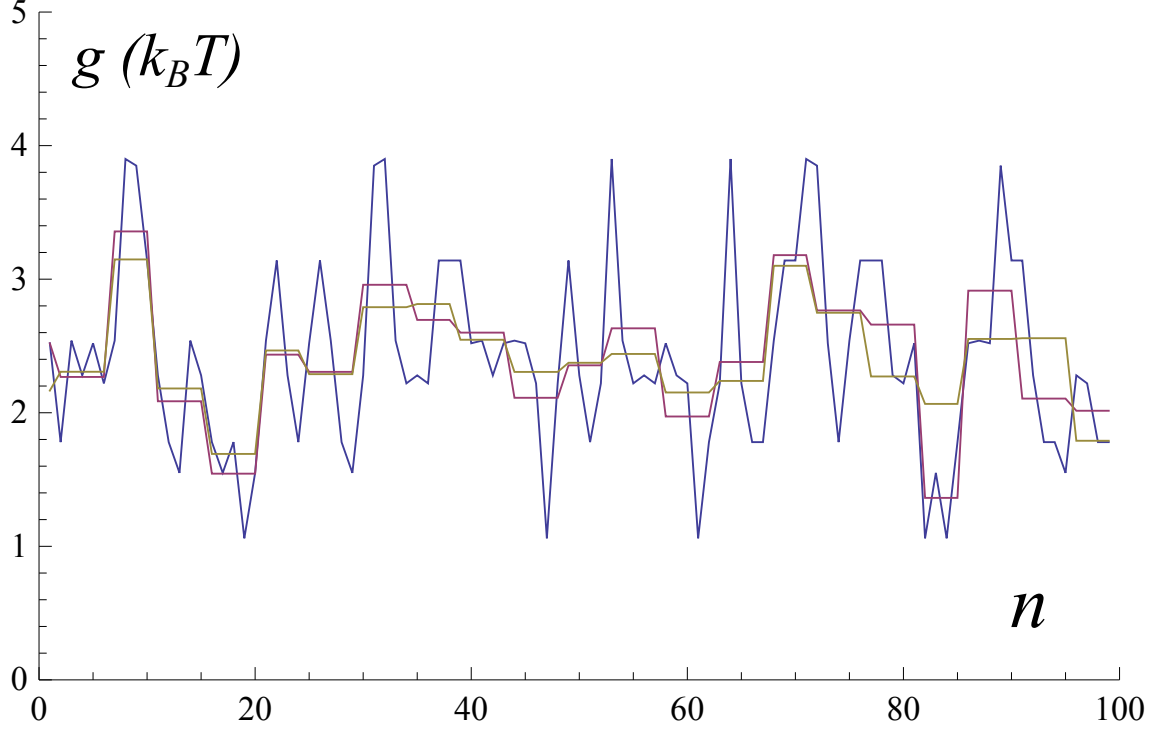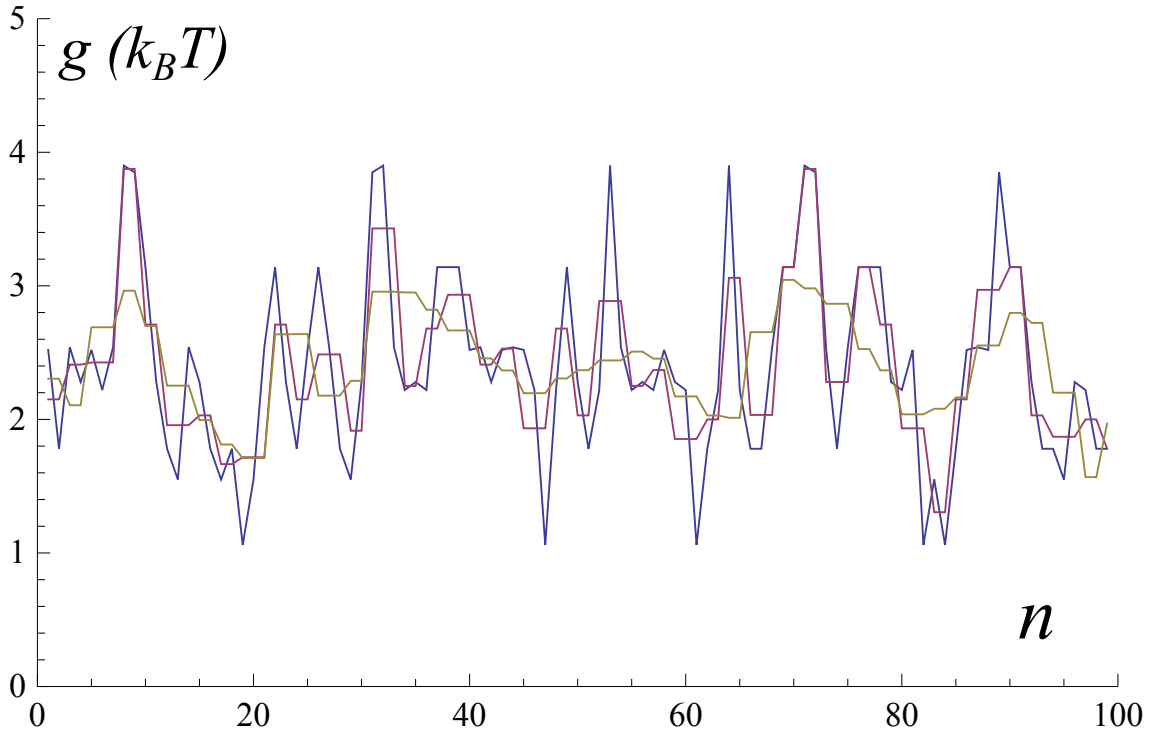
Figure 5.9: The figure shows the $g_{\text{trial}}(n|c_i)$ for $\gamma = 0.00016$ (brown), the real g_0 (blue) and the $g_{\text{trial}}(n|d_i)$ (purple).
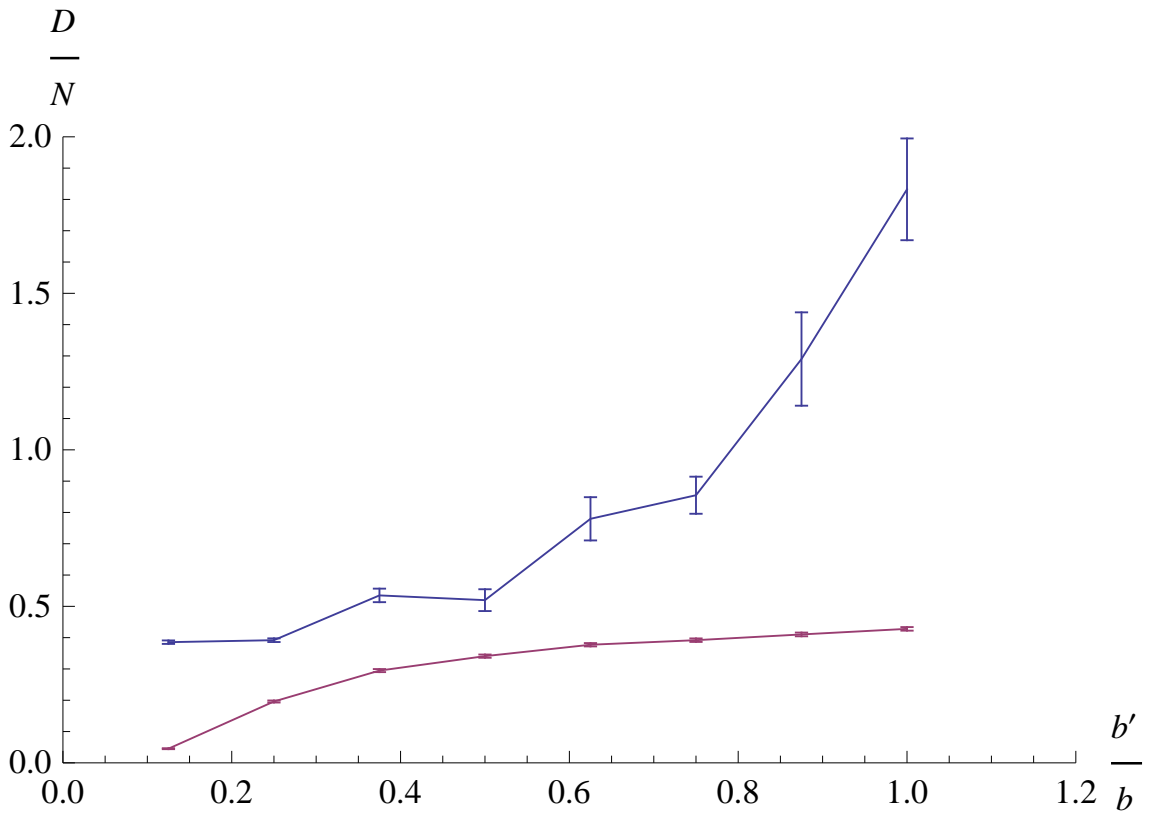
Figure 5.10: $D(c_i)$ (blue) and $D(d_i)$ (purple) as functions of b' averaged over 100 random sequences, the error-bars are the standard deviation of the mean
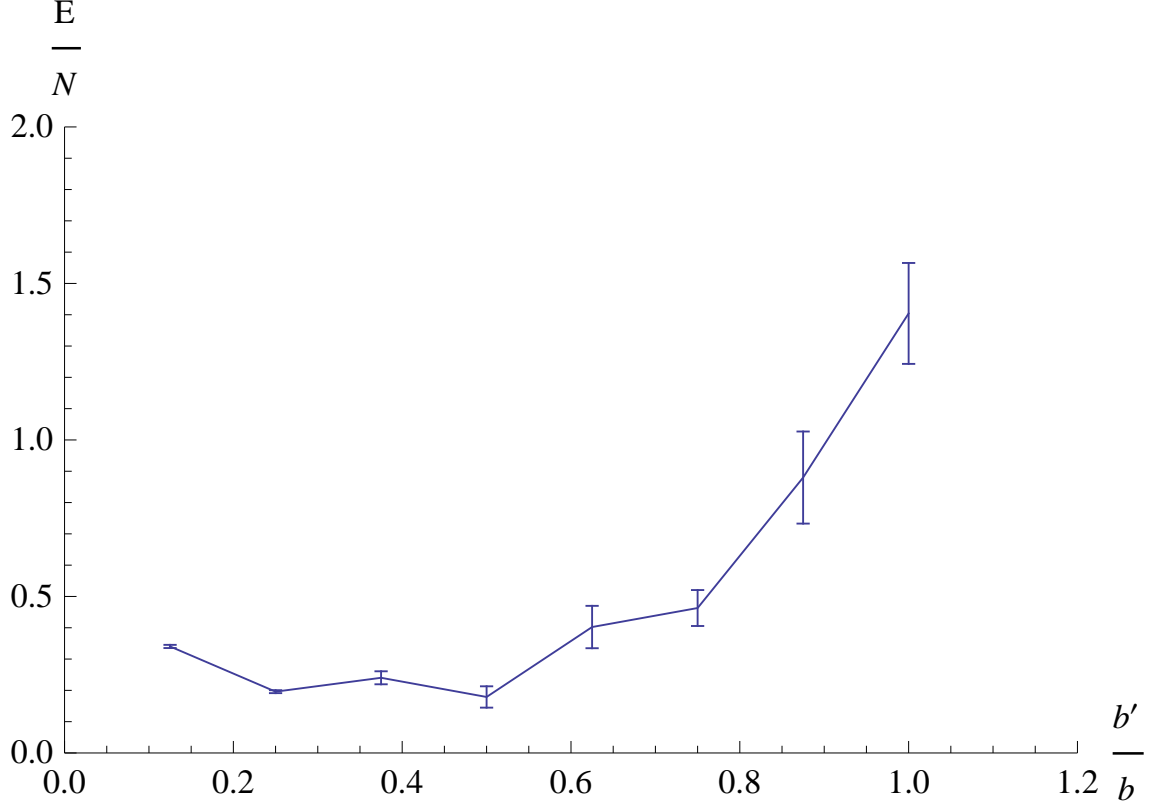
Figure 5.11: $E(c_i|d_i)$ as a function of b' averaged over 100 random sequences, the error-bars are the standard deviation of the mean

than one.

We also include the minimum of the average of D and E over 100 sequences obtained for a given b, regardless of the value of b' that corresponds to it in figure 5.14

5.2.3   Comparison with the moving average

This part stems from the observation that the $g_{\text{trial}}(n|c_i)$ when $b' = 1$ looks a lot like a smoothed version of the $g_0(n)$ we have thus defined $g_\sigma(n)$ a Gaussian filter as the convolution product between the $g_0(n)$ and a Gaussian kernel of width σ.

We then look for the σ that minimizes the following cost function:

$$F(c_i, \sigma) = \sum_n^N (g_\sigma(n) - g_{\text{trial}}(n|c_i))^2 , \qquad (5.21)$$

where the c_i are, as usual, the set of parameter that minimize the C cost function.

After several runs we have found that the optimal value of σ is roughly increasing with increasing b, but that different sequences can lead to quite different optimal σ. One would expect σ to be linearly related to the optimal b', but there is too much of a sequence dependence to conclude that. Two examples are shown in figure 5.15.

5.2.4   Difference with the naïve prediction

One possible way to perform inference on through the measurement of $u(L)$ at equilibrium is to approximate the expression in equation (5.11) through a saddle point. That is to say we find the base n^* that has the biggest contribution for a given length L and neglige all
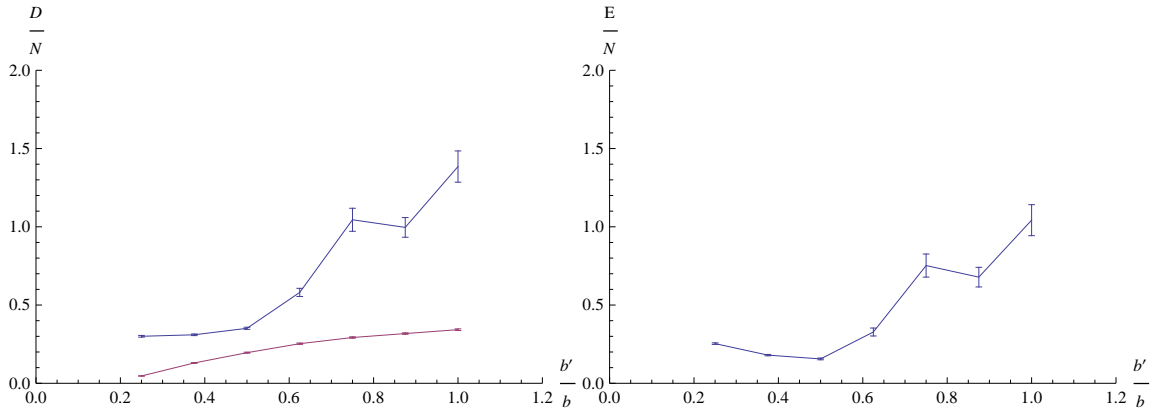
Figure 5.12: Value of cost functions when $b = 4.69$. The cost functions are shown as a function of b' averaged over 100 random sequences, the error-bars are the standard deviation of the mean. Left: $D(c_i)$ (blue) and $D(d_i)$ (purple) . Right: $E(c_i|d_i)$
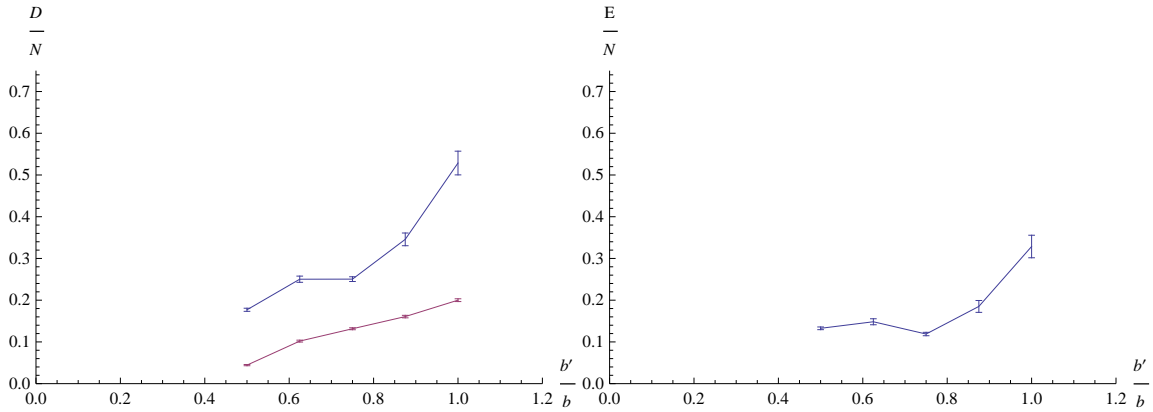
Figure 5.13: Value of cost functions when $b = 2.35$. The cost functions are shown as a function of b' averaged over 100 random sequences, the error-bars are the standard deviation of the mean. Left: $D(c_i)$ (blue) and $D(d_i)$ (purple) . Right: $E(c_i|d_i)$
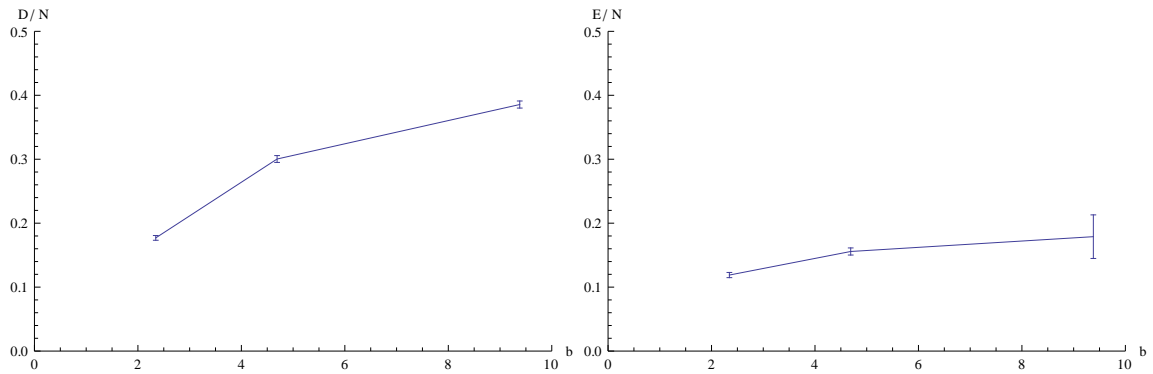
Figure 5.14: Value of the cost functions D (left) and E (right) for different values of b. The plotted value is the minimum of the average over b' (see figures 5.10, 5.11, 5.12, 5.13. Error bars are standard deviations over 100 sequences.
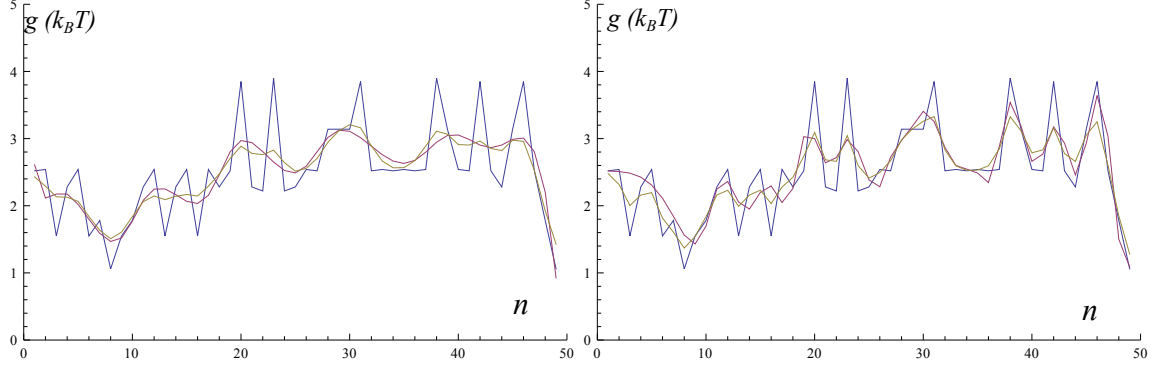
Figure 5.15: Left: $b = 9.38$, $g_0(n)$ (blue), $g_{\text{trial}}(n|c_i)$ (violet), $g_\sigma(n)$ (brown) for $\sigma = 2.75$. Right: $b = 2.35$, same color code, but $\sigma = 1.95$.

other contributions.

$$u(\bar{L}) = \sum_n^N u(n, L) P(n, L) \simeq u(n^*, L), \qquad (5.22)$$

where $P(n, L)$ is the exponential in equation (5.11), and $u(n, L) = k_{\text{eff}}/k_1(L - nl)$. Now, n^* is given by maximising $P(n, L)$, by solving:

$$g_0(n^*) = k_1 l u(n^*, L), \qquad (5.23)$$

and this equation looks as though we could use it to infer the $g_0(n^*)$ through the value of $u(n^*, L)$ which should be close to $u(L)$. The point where all this doesn't add up is the choice of a suitable L: we'd like to find $L(n^*)$ to know which L contributes the most to a given n^*. To do so we solve:

$$g_0(n^*) = k_{\text{eff}}(L - n^* l). \qquad (5.24)$$

Ideally, we'd like the solution of this to depend strongly on n^*, but not through $g_0(n^*)$ which is unknown, what we find instead is that with current state of the art experiments $g_0(n^*)/k_{\text{eff}}$ is two orders of magnitude larger than l.

This means that the $L_0(n^*)$ that solves this equation is not a nice, linear function of n^*, but instead depends very strongly on the sequence. This translates itself into a wild n^*-dependent dephasing between the naïf prediction and the Gaussian average of the sequence. For short (e.g. 100 bases) sequences this dephasing effect is dominant.

What this really points to is that the saddle point approximation is not suitable for a case where k_{eff} is so small, since it should, in principle, diverge.

However if we take a constant shift $L_0 = \bar{g}_0/k_{\text{eff}}$ we can show how bad this technique is compared to what we have proposed, by comparing it to a Gaussian runnin average with $\sigma = 25$ (which is much bigger than the ~ 2 we have used before). The results are shown in figure 5.16.

This technique can be used to extract rapidly the average of the sequence on ~ 25 bases on long sequences, since it is much faster than what we have proposed.

5.2.5 Scaling of computational time as a function of sequence length

As of now the algorithm scales as the cube of the number of basis. In principle it is possible to split a long sequence in smaller batches, and fit the separately, but some practical problems must be adressed.

First of all we have to take into account what we have discussed in the previson section,
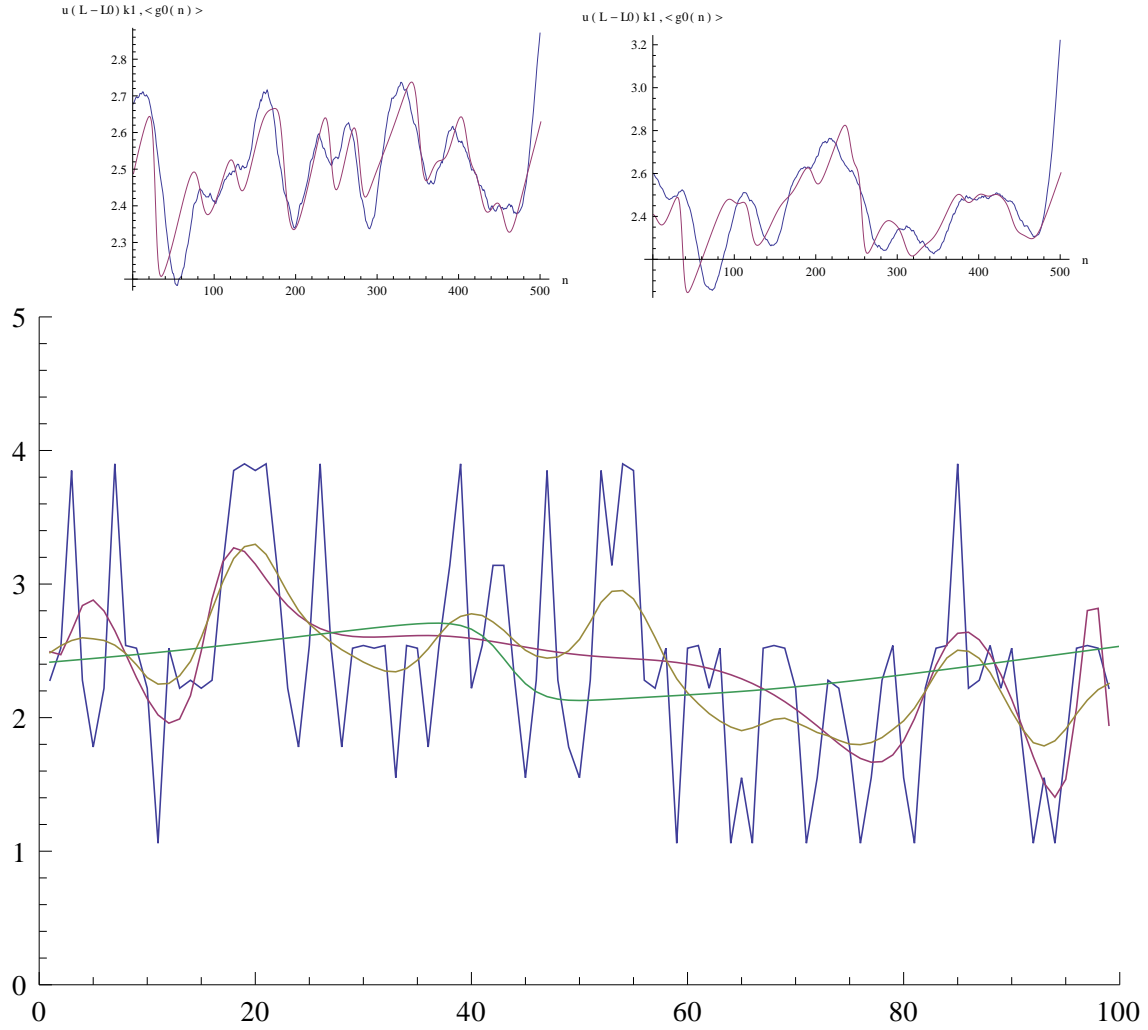
Figure 5.16: Here we show two different sequences 500 bases long and one which is only 100 bases long. The naïf estimate $u(L + L_0)k_1$ (violet) is compared to a Gaussian running average with $\sigma = 25$. On the 100 bases long sequence we compare the real sequence (blue), to the fit obtained for $b' = 1$ (violet), the moving average with $\sigma = 5.65$ (brown) and the naïf prediction (bright green)
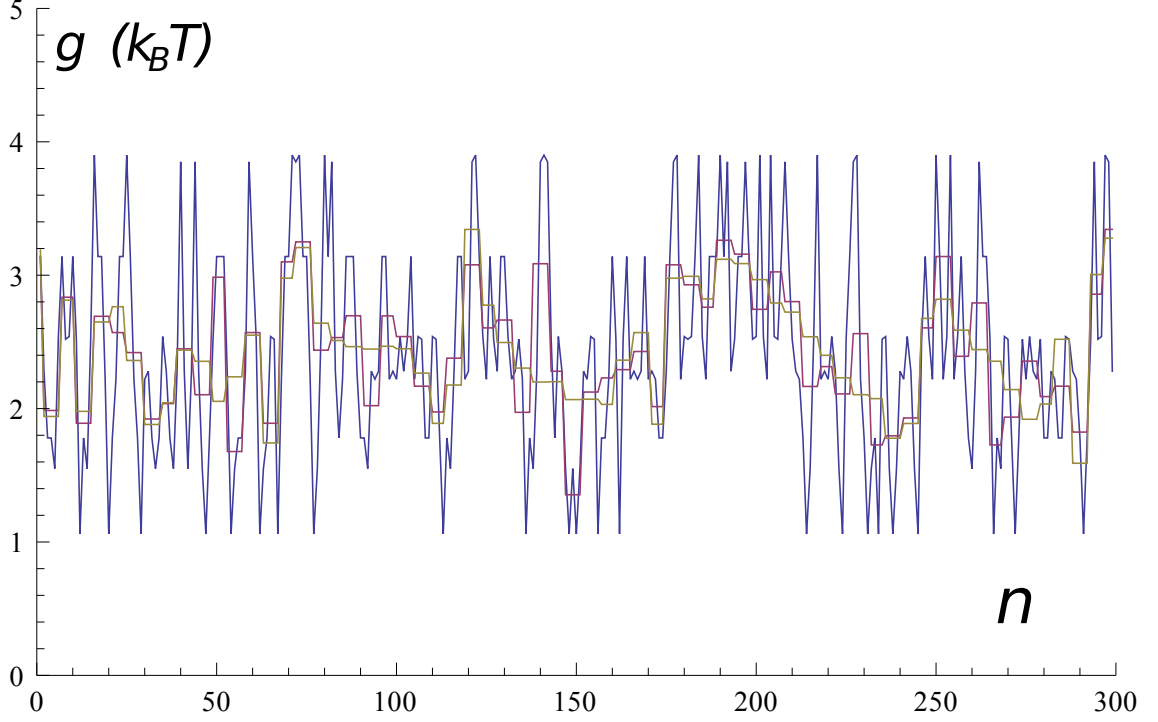
Figure 5.17: For 300 bases, $b = 9.38$ and $b' = b/2$: $g_0(n)$ (blue), $g_{\text{trial}}(n|c_i)$ (violet) and $g_{\text{trial}}(n|d_i)$ (brown)

that is: it is impossible to know the relationship between L and n without knowing the g_0 with a sufficient degree of precision.

So suppose we want to fit a section of the $u(L)$ curve, say from L_1 to L_2, it is impossible to say what are the n's that correspond to that interval with precision and we may end up adding a few hundreds left and right just to be sure, thus killing any advantage we might have had splitting unless the sequence is some 40 kbp long.

And here is where the second problem comes into play: up to now we have considered k to be roughly constant, but k really depends on n albeit weakly. A change in n of the order of 1 kbp on the other hand would not be negligible anymore and would lower the value of k, and thus of k_{eff} of an order of magnitude.

This is currently a limitation of all current single molecule experiments. whenever opening too long a molecule the linkers become too elastic to yield meaningful insights on the g_0. In figure 5.17 we display a fit of a sequence 300 bp long with $b = 9.38$ and $b' = b/2$. This computation takes Mathematica 7 a little more than 20 minutes on a Intel core 2 processor and uses up, about 1.5 GB of RAM. It involves a search in a 64-dimensional parameter space.

5.2.6 Estimation of the error bars

In least squares fitting it is costumary to estimate the variance of the variables through the Hessian of the cost function at the minimum. Let $H_{ij} = \frac{\partial^2 C}{\partial c_i \partial c_j}$ calculated at the minimum. Then the variances are given by

$$\sigma_{c_i}^2 = \sigma^2 (H)_{ii}^{-1}, \tag{5.25}$$

where σ^2 is the true residual variance, which is unkown, but is usually estimated as C^*/n, where C^* is the value of the cost function at the minimum and n is the number of degrees of freedom.

If we do so without taking into account the prior H is not positive definite and we end up with negative variances. Because of this we use the full cost function with the prior. Three examples of the results is shown in figure 5.18.

The reader should note how for an unchanged b, there is not much gain in lowering b'. On the other hand when b is smaller the fit is much better and this is reflected in the error-bars.

5.2.7 Entropy

Let us suppose we consider the cost functions we have defined in the preceding sections as thermodynamic quantities. As such we can draw a physical analogy and estimate the number of sequences that correspond to a given value of the cost function through the entropy.

By defining a pseudo-temperature we can define the partition function as:

$$
\begin{aligned}
Z &= \sum_{\{b_n\}} \exp\left[-\frac{\lambda^2}{2} E(c_i|d_i) \right] \\
&= \sum_{\{b_n\}} \exp\left[-\frac{\lambda^2}{2} \sum_n \left(g_{\text{trial}}(n|c_i) - g_{\text{trial}}(n|d_i) \right)^2 \right] \\
&= \sum_{\{b_n\}} \exp\left[-\frac{\lambda^2}{2} \sum_i |\omega_i| \left(c_i - \frac{1}{|\omega_i|} \sum_{j \in \omega_i} g_0(b_j, b_{j+1}) \right)^2 \right] \\
&= A \int d\vec{x}\, e^{-\sum_i \frac{1}{2|\omega_i|} x_i^2} \sum_{\{b_n\}} \exp\left[-i\lambda \sum_i x_i \left(c_i - \frac{1}{|\omega_i|} \sum_{j \in \omega_i} g_0(b_j, b_{j+1}) \right) \right] \\
&= A \int d\vec{x}\, e^{-\sum_i \frac{1}{2|\omega_i|} x_i^2 - i\lambda x_i c_i} \sum_{\{b_n\}} \exp\left[i\lambda \sum_n \frac{\Omega_{b'}(n - b'i)}{|\omega_i|} x_i g_0(b_n, b_{n+1}) \right] \\
&= A \int d\vec{x}\, e^{-\sum_i \frac{1}{2|\omega_i|} x_i^2 - i\lambda x_i c_i} \sum_{\{b_n\}} \prod_n \exp\left[i\lambda \frac{\Omega_{b'}(n - b'i)}{|\omega_i|} x_i g_0(b_n, b_{n+1}) \right],
\end{aligned}
\tag{5.26}
$$

where $A = \prod_i (2\pi|\omega_i|)^{-\frac{1}{2}}$.

Using the transfer-matrix method we can recognise $\exp\left[i\lambda \frac{\Omega_{b'}(n-b'i)}{|\omega_i|} x_i g_0(b_n, b_{n+1}) \right]$ as a 4×4 matrix which appears $|\omega_i|$ times identical and then involves a different x_i.

Because of this we can rewrite the previous equation as:

$$
Z = A \int d\vec{x}\, e^{-\sum_i \frac{1}{2|\omega_i|} x_i^2 - i\lambda x_i c_i} \sum_{b_0, b_N} \vec{b}_0 \cdot \prod_i^M [\mathbf{T}(i\lambda x_i)/|\omega_i|)]^{|\omega_i|} \cdot \vec{b}_N,
\tag{5.27}
$$

where:

$$
\mathbf{T}_{b_n, b_{n+1}}(t) = \exp\left[g_0(b_n, b_{n+1})t \right] = \begin{pmatrix} 5.93^t & 4.71^t & 12.43^t & 9.21^t \\ 2.89^t & 5.93^t & 9.78^t & 12.68^t \\ 12.68^t & 9.21^t & 23.1^t & 46.99^t \\ 9.78^t & 12.43^t & 49.4^t & 23.1^t \end{pmatrix}
\tag{5.28}
$$

By rearranging the terms one can decouple the integrals

$$
Z = A \sum_{b_0, b_N} \vec{b}_0 \cdot \prod_i^M \left[\int dx_i\, e^{-\sum_i \frac{1}{2|\omega_i|} x_i^2 - i\lambda x_i c_i} [\mathbf{T}(i\lambda x_i)/|\omega_i|)]^{|\omega_i|} \right] \cdot \vec{b}_N,
\tag{5.29}
$$
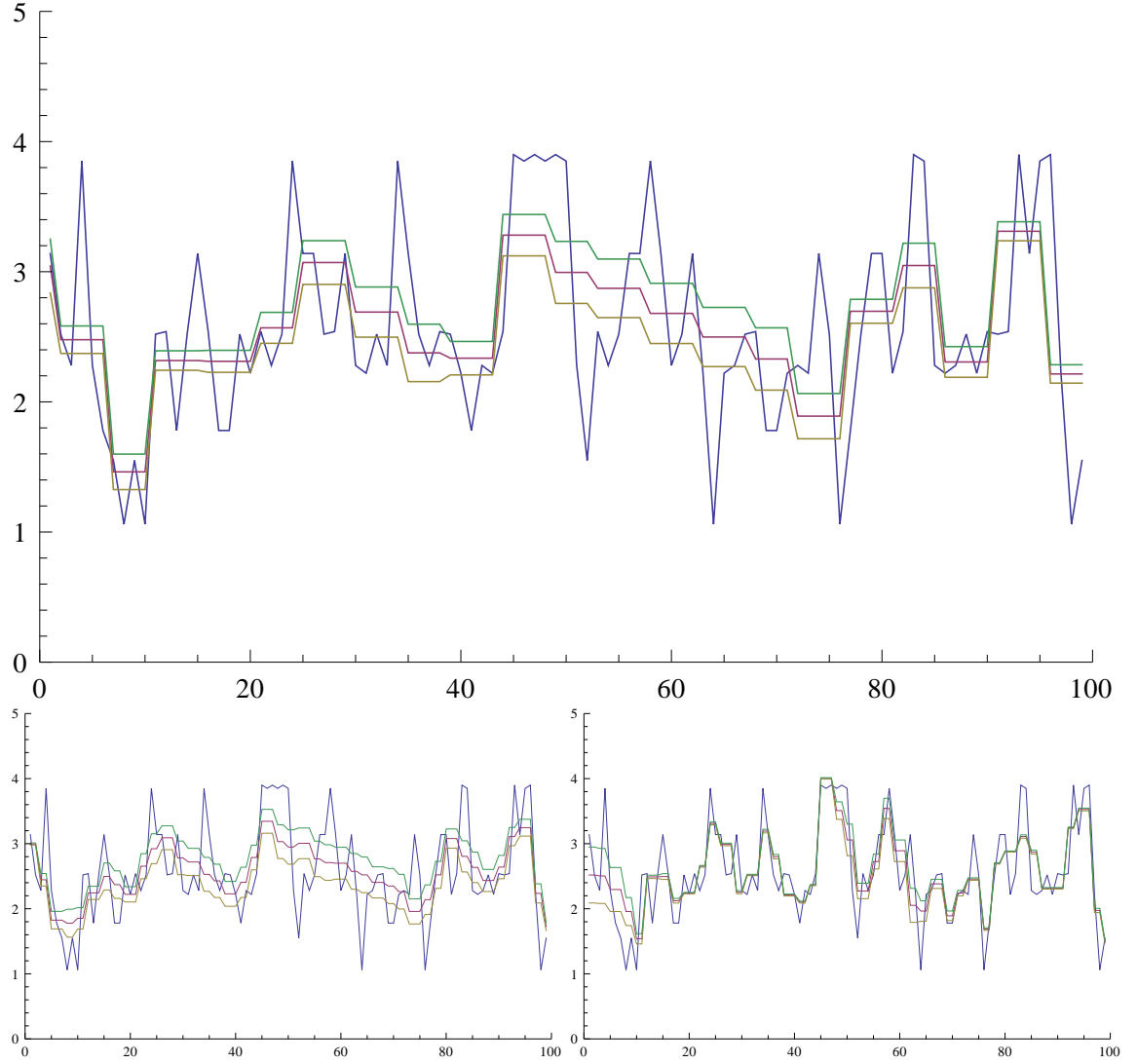
Figure 5.18: For the three panels: g_0 in blue. The other three curves are the g_{trial}, and the $g_{\text{trial}} \pm \sigma$. For the top panel we have $b = 9.38$, $b' = b/2$, $N = 100$. $\gamma = 0.1$. On the bottom right we have changed the value of b' to $b/4$ and $\gamma = 10^{-4}$. On the bottom right we have changed l to nm so that $b = 2.35$ and we have kept $b' = b$, the fit is obtained for $\gamma = 10^{-4}$.
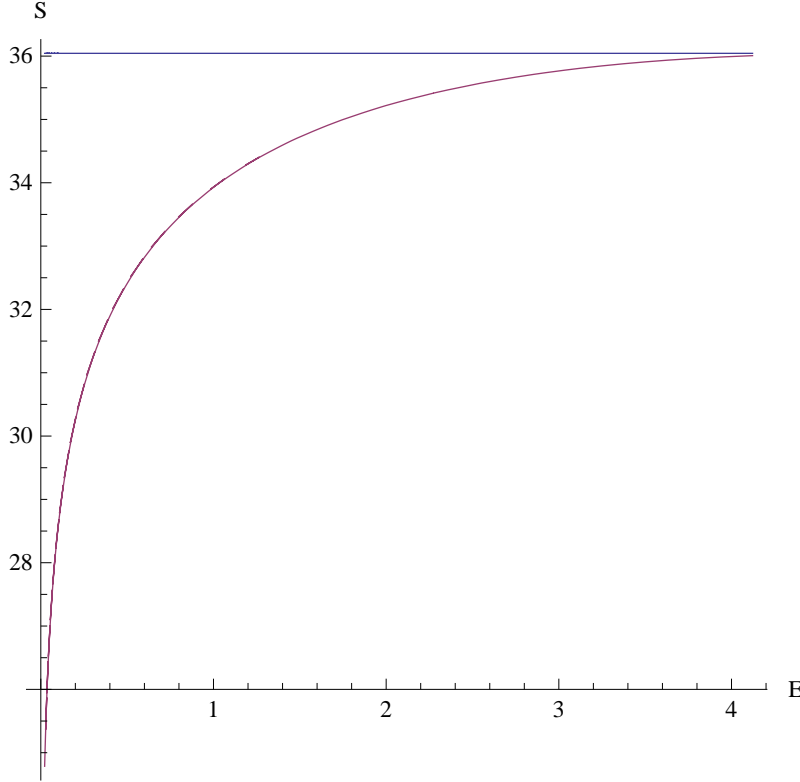
Figure 5.19: Entropy for a sequence of 26 bases (25 values of g_0) and 5 measures. The entropy is computed for $c_i = 2.52$ for every i. At high energies the entropy saturates to $S(\infty) = 26 \log(4) \simeq 36$, the right value which is the logarithm of the number of possible sequences of 26 bases.

Once we have computed the one dimensional integrals we can multiply the N matrices and sum over the first and last base.

We can then change variable ($\beta = \lambda^2/2$) and compute the thermodynamic quantities as:

$$E(\beta) = -\frac{\partial}{\partial \beta} \log Z \tag{5.30}$$

$$F(\beta) = -\frac{1}{\beta} \log Z \tag{5.31}$$

$$S(E) = \max_{\beta}(\beta(E - F(\beta))) \tag{5.32}$$

The entropy as a function of internal energy can be also obtained with a parametric plot, as it's shown in figure 5.19. Even through the simplifications obtained thanks to the transfer matrix method, the computation of entropy is very taxing and we didn't have enough memory for computing the entropy for cases where the c_i are not all identical or for longer sequences.

5.2.8   A different approach

In this section we will expound a different approach for tackling the same problem as developed by Jörg, Monasson and Cocco and is yet to be published.

The main difference is that this formalism allows for the description of the same system in a vector space and defines measures as orthogonality constraints.

This formalism has also allowed Jörg et al. to compute interesting statistical properties of this system, but we will not dwell on the details here.

Equation (5.11) can be rewritten by multiplying both sides by $Z(B)$:

$$\sum_n v_n(B)[\alpha(L - nl) - \beta\bar{u}(L)] \exp\left[-\frac{\kappa}{2}(L - nl)^2\right] = 0, \tag{5.33}$$

where

$$v_n(B) = \exp\left[-\sum_j^n g_o(j)\right] \tag{5.34}$$

$$\alpha = \frac{kk_2}{\sqrt{(k_1k_2 + kk_1 + kk_2)^3}}, \tag{5.35}$$

$$\beta = \frac{1}{\sqrt{k_1k_2 + kk_1 + kk_2}} \tag{5.36}$$

and

$$\kappa^{-1} = \frac{1}{k} + \frac{1}{k_1} + \frac{1}{k_2}. \tag{5.37}$$

This equation only makes sense if we use for the $\bar{u}(L)$ the measures we have obtained from an experiment.

Equation (5.33) can be rewritten as:

$$\sum_n v_n(B)p_n(L) = 0, \tag{5.38}$$

where $p_n(L) = [\alpha(L - nl) - \beta\bar{u}(L)] \exp\left[-\frac{\kappa}{2}(L - nl)^2\right]$ is a vector that depends on a on a given measure $\bar{u}(L, B)$.

This can be thought of as the scalar product $\vec{v}(B) \cdot \vec{p}(L)$, suggesting a geometrical interpretation: we have to choose the sequence B so that it is orthogonal to all the vectors given by the measures encoded by the vectors $p(L)$ for different values of L.

The problem of finding the optimal vector $\vec{v}(B)$ can be rephrased as a minimization problem over a quadratic form by squaring both sides:

$$\left[\sum_n v_n(B)p_n(L)\right]^2 = \sum_{m,n} v_m(B)p_m(L)v_n(B)p_n(L) =$$
$$\sum_{m,n} v_m(B)K_{m,n}(L)v_n(B) = \vec{v}^\dagger(B)\mathbf{K}(L)\vec{v}(B) = 0, \tag{5.39}$$

where $K_{m,n}(L) = p_m(L)p_n(L)$. Different measures are easily taken into account by adding the terms obtained for different L's:

$$\vec{v}^\dagger(B)\left[\sum_{i=1}^M W_i\mathbf{K}(L_i)\right]\vec{v}(B) = 0, \tag{5.40}$$

where W_i are arbitrary positive weights.

5.3 Dynamical algorithm

5.3.1 A toy model: coupled Ornstein-Uhlenbeck processes

Real unzipping measurements do not grant us access to the instantaneous force (or displacement) signal. What is actually measured is a signal which is time averaged over a

period of a few milliseconds.

In this section we wish to explore the effects of time averaging on a simple stochastic system. We will compute the probability of observing a series of time averages given a set of parameters and thanks to the Bayes theorem we will be able to chose the most likely set of parameters given a set of measures.

Let us consider an Ornstein-Uhlembeck process [Uhlenbeck 30]:

$$\gamma \dot{x} = -k(x - y) + \eta \,, \tag{5.41}$$

where η is a Gaussian noise with zero mean and variance $\langle \eta(t)\eta(t') \rangle = 2k_B T \gamma \delta(t - t')$.

We wish to consider its temporal average $\bar{x}$ over a certain time and to infer from it the physical quantities γ and k.

The solution of the model is well known and it's the stochastic function:

$$x(t) = x_0 e^{-\frac{k}{\gamma}t} + y(1 - e^{-\frac{k}{\gamma}t}) + \frac{1}{\gamma} \int_0^t dt' e^{-\frac{k}{\gamma}(t-t')} \eta(t') \,. \tag{5.42}$$

That is a Gaussian process with mean and variance given by:

$$\langle x(t) \rangle = x_0 e^{-\frac{k}{\gamma}t} + y(1 - e^{-\frac{k}{\gamma}t}) \,, \tag{5.43}$$

$$\langle x(t)^2 \rangle - \langle x(t) \rangle^2 = \frac{k_B T}{k} \left(1 - e^{-2\frac{k}{\gamma}t} \right) \,. \tag{5.44}$$

If we now consider the time average over a time t of the same stochastic function we obtain another stochastic function of the form:

$$\bar{x}(t) = \frac{1}{t} \int_0^t dt' x(t') = \frac{\gamma}{kt}(x_0 - y)(1 - e^{-\frac{k}{\gamma}t}) + y$$
$$+ \frac{1}{\gamma t} \int_0^t dt' \int_0^{t'} dt'' e^{-\frac{k}{\gamma}(t'-t'')} \eta(t'') \,. \tag{5.45}$$

That is a Gaussian process with mean and variance:

$$\langle \bar{x} \rangle = \frac{\gamma}{kt}(x_0 - y)(1 - e^{-\frac{k}{\gamma}t}) + y \,, \tag{5.46}$$

$$\langle \bar{x}(t)^2 \rangle - \langle \bar{x}(t) \rangle^2 = \frac{2k_B T \gamma}{k^2 t} + \frac{k_B T \gamma^2}{k^3 t^2} \left(-3 + 4e^{-\frac{k}{\gamma}t} - e^{-\frac{2k}{\gamma}t} \right) \,, \tag{5.47}$$

and additionaly we should consider:

$$\langle \bar{x}(t)x(t) \rangle - \langle \bar{x}(t) \rangle \langle x(t) \rangle = \frac{k_B T \gamma}{k^2 t} \left(1 - e^{-\frac{k}{\gamma}t} \right)^2 \,. \tag{5.48}$$

All this can be summarized defining a covariance matrix as a function of a dimensionless time $\tau = kt/\gamma$:

$$\mathbf{C} = \frac{k_B T}{k} \begin{pmatrix} 1 - e^{-2\tau} & \frac{\left(1-e^{-\tau}\right)^2}{\tau} \\ \frac{\left(1-e^{-\tau}\right)^2}{\tau} & \frac{2}{\tau} + \frac{1}{\tau^2}\left(-3 + 4e^{-\tau} - e^{-2\tau}\right) \end{pmatrix} \,, \tag{5.49}$$

But since the process is Gaussian we can write the full probability starting from the means vector and the covariance matrix:

$$P(x(t), \bar{x}(t)|x_0) = \frac{1}{2\pi\sqrt{\det \mathbf{C}}} \exp\left(-\frac{1}{2}\vec{x}^\dagger \mathbf{C}^{-1} \vec{x} \right) \,, \tag{5.50}$$
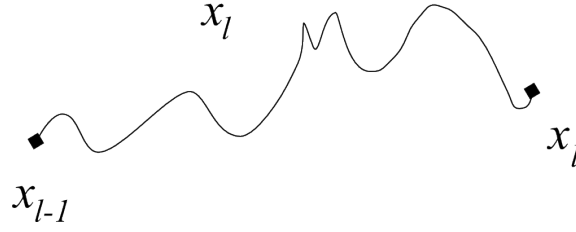
Figure 5.20: The evolution of the Ornstein-Uhlenbeck process during a time step and its average.

where $\vec{x} = \begin{pmatrix} x(t) - \langle x(t) \rangle \\ \bar{x}(t) - \langle \bar{x}(t) \rangle \end{pmatrix}$ and $\mathbf{C}^{-1}$ is the inverse of $\mathbf{C}$, that is:

$$
\begin{aligned}
\mathbf{C}^{-1} &= \frac{k}{k_B T \left(\tau \left(1 + e^{-\tau} \right) - 2 \left(1 - e^{-\tau} \right) \right)} \\
&\times \begin{pmatrix} \frac{2\tau - 3 + 4e^{-\tau} - e^{-2\tau}}{2(1 - e^{-\tau})} & -\tau \left(1 - e^{-\tau} \right) \\ -\tau \left(1 - e^{-\tau} \right) & \tau^2 \left(1 + e^{-\tau} \right) \end{pmatrix}
\end{aligned}
\tag{5.51}
$$

What we have just wrote defines the evolution of the system through an amount of time t; let us now just suppose that this is just a step in the evolution of the sistem, that is, at time $(l-1)\Delta t$ the system is in x_{l-1} and it evolves to x_l in $l\Delta t$ as shown in figure 5.20. In this time interval its time average is defined as:

$$
\bar{x}_l = \frac{1}{\Delta t} \int_{(l-1)\Delta t}^{l\Delta t} dt' x(t') .
\tag{5.52}
$$

If we set: $x_{l-1} = x_0$, $x_l = x(t)$, $\bar{x}_l = \bar{x}(t)$ and $\tau = k\Delta t/\gamma$ we can recycle the previous expression to define a *propagator*:

$$
\begin{aligned}
P(x_l, \bar{x}_l | x_{l-1}) &= \frac{1}{2\pi\sqrt{\det \mathbf{C}}} \\
&\times \exp\left(-\frac{1}{2} (x_l - \langle x_l \rangle, \bar{x}_l - \langle \bar{x}_l \rangle) \mathbf{C}^{-1} \begin{pmatrix} x_l - \langle x_l \rangle \\ \bar{x}_l - \langle \bar{x}_l \rangle \end{pmatrix} \right) .
\end{aligned}
\tag{5.53}
$$

So, as long as the Ornstein-Uhlenbeck process is a Markov process, we can write the joint probability of the process as:

$$
P(\{\bar{x}_l, x_l\}_{l=1}^L | x_0) = \prod_{l=1}^L P(x_l, \bar{x}_l | x_{l-1}) .
\tag{5.54}
$$

This is pictured in figure 5.21 and can easily rewritten as a single exponential:

$$
P(\{\bar{x}_l, x_l\}_{l=1}^L | x_0) = \frac{1}{(2\pi\sqrt{\det \mathbf{C}})^L} \exp\left(-\frac{k}{2k_B T} Q \right) ,
\tag{5.55}
$$

where Q is:

$$
\begin{aligned}
Q = \sum_{l=1}^L \Bigg[&A x_l^2 + B x_l x_{l-1} + C x_{l-1}^2 - D(x_l + x_{l-1})\bar{x}_l - (x_l - x_{l-1})y - \tau \bar{x}_l y \\
&+ E \bar{x}_l^2 + \tau \frac{y^2}{2} \Bigg] ,
\end{aligned}
\tag{5.56}
$$
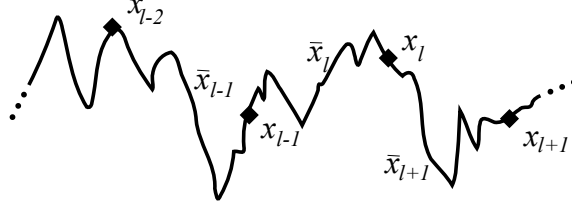
Figure 5.21: The evolution of the Ornstein-Uhlenbeck process during several time-steps.

where:

$$A = \frac{2\tau - 3 + 4e^{-\tau} - e^{-2\tau}}{2\left(\tau\left(1 + e^{-\tau}\right) - 2\left(1 - e^{-\tau}\right)\right)\left(1 - e^{-\tau}\right)} \tag{5.57}$$

$$B = \frac{2 - 4\tau e^{-\tau} - 2e^{-2\tau}}{2\left(\tau\left(1 + e^{-\tau}\right) - 2\left(1 - e^{-\tau}\right)\right)\left(1 - e^{-\tau}\right)} \tag{5.58}$$

$$C = \frac{2\tau e^{-2\tau} + 1 - 4e^{-\tau} + 3e^{-2\tau}}{2\left(\tau\left(1 + e^{-\tau}\right) - 2\left(1 - e^{-\tau}\right)\right)\left(1 - e^{-\tau}\right)} \tag{5.59}$$

$$D = \frac{2\tau\left(1 - e^{-\tau}\right)}{2\left(\tau\left(1 + e^{-\tau}\right) - 2\left(1 - e^{-\tau}\right)\right)} \tag{5.60}$$

$$E = \frac{\tau^2\left(1 + e^{-\tau}\right)}{2\left(\tau\left(1 + e^{-\tau}\right) - 2\left(1 - e^{-\tau}\right)\right)} \,. \tag{5.61}$$

We would like now to integrate out the x_i's in order to obtain the joint probability distribution for the time averages only:

$$P(\{\bar{x}_l\}_{l=1}^L | x_0) = \int_{-\infty}^{\infty} \prod_{l=1}^{L} dx_l P(x_l, \bar{x}_l | x_{l-1}) \,. \tag{5.62}$$

In order to perform this integral in full generality we need to change variables in order to diagonalize the quadratic form Q and factorize the integrals.

Q can be diagonalised by a discrete Fourier tranform, provided we force periodic boundary conditions (by imposing $x_0 = x_L$), that is:

$$X_q = \frac{1}{\sqrt{L}} \sum_{l=1}^{L} x_l e^{\frac{-2\pi i q l}{L}} \tag{5.63}$$

$$x_l = \frac{1}{\sqrt{L}} \sum_{q=1}^{L} X_q e^{\frac{2\pi i q l}{L}} \,, \tag{5.64}$$

the choice of prefactors ensures the unitarity of the transformation which is known to be orthogonal. Q is thus transformed into:

$$Q = \sum_{q=1}^{L} \left[\left(A + C + B\cos\left(\frac{2\pi q}{L}\right)\right) X_q^2 - D\left(1 + \cos\left(\frac{2\pi q}{L}\right)\right) X_q \bar{X}_q \right.$$
$$\left. + E\bar{X}_q^2 \right] - \sqrt{L}\tau y \bar{X}_0 + L\tau \frac{y^2}{2} \,. \tag{5.65}$$

Thus integrating over the X_q yields:

$$
\begin{aligned}
\tilde{Q} &= \sum_{q=1}^{L} \left[E - \frac{D^2 \left(1 + \cos\left(\frac{2\pi q}{L}\right) \right)^2}{4 \left(A + C + B \cos\left(\frac{2\pi q}{L}\right) \right)} \right] \bar{X}_q^2 - \sqrt{L}\tau y \bar{X}_0 + L\tau \frac{y^2}{2} \\
&= \sum_{q=1}^{L} \frac{\tau^2 \left[1 + e^{-\tau} - \frac{\left(1-e^{-\tau}\right)^3 \cos^4\left(\frac{\pi q}{L}\right)}{\tau\left(1+e^{-2\tau}\right)-\left(1-e^{-2\tau}\right)+\cos\left(\frac{2\pi q}{L}\right)\left(1-2\tau e^{-\tau}-e^{-2\tau}\right)} \right]}{2 \left(\tau \left(1 + e^{-\tau} \right) - 2 \left(1 - e^{-\tau} \right) \right)} \bar{X}_q^2 \\
&\quad - \sqrt{L}\tau y \bar{X}_0 + L\tau \frac{y^2}{2} \ .
\end{aligned}
\tag{5.66}
$$

We can now use Bayes' theorem to interpret the probability in eq (5.62) as the likelihood of a set of measures being generated by a given τ.

With standard computational techniques one can compute the log-likelihood in time $O(L^2)$. In figure 5.22 we show the results for simulated runs of different lengths. It is easily shown how the prediction of τ improves with more points, but it's already reasonably good with only 200 points.

One could also use the width of this curve to compute $k/(k_{\mathrm{B}}T)$ and by knowing the value of the temperature compute γ and k.

What is compelling about this algorithm is that we are exploiting all the information available: fluctuations, correlations and not only the averages.
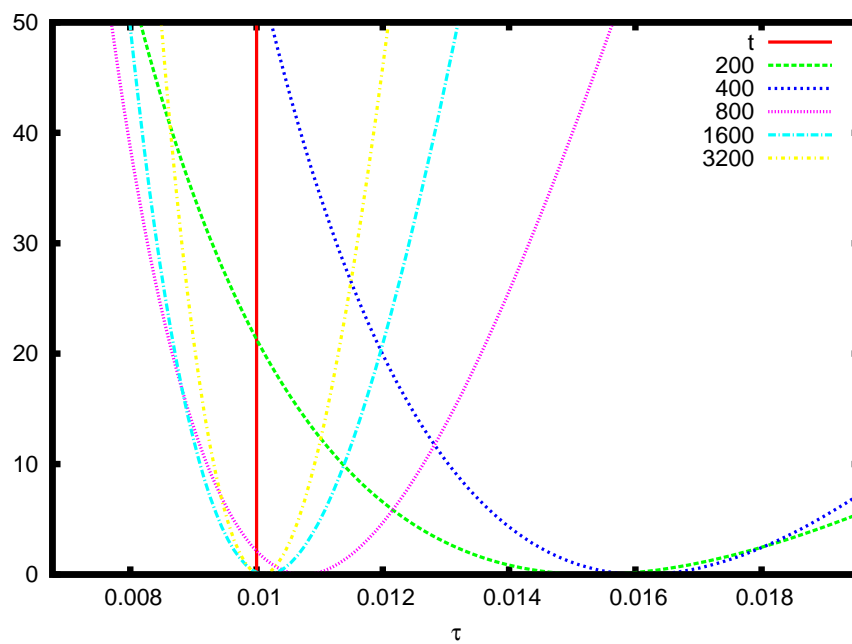
Figure 5.22: The log-likelihood as a function of τ for different L (indicated in the legend), in red we show the actual value of τ that generated the data. The log-likelihoods have been offset by a constant value and changed into its opposite for cosmetic reasons. The minimum of the displayed curve is the most likely value of τ.

Conclusions and outlook

Infotaxis

In the part devoted to infotaxis we have developed a continuous version of the algorithm and analyzed its behavior and performance in two and three dimensions.

We have shown the probability of success not to depend on the distance from the source when this latter is of the order of magnitude of λ, the characteristic length of the odor advection phenomenon.

Furthermore we have shown the search time to grow with γ, the parameter that regulates the speed of the searcher in response to the gradient. However, we have observed simulations with small γ to be computationally more taxing.

The computational time needed to perform a single step is still very large: this is due to the need of computing many Monte Carlo integrations over the whole space, but also from the strategy we have chosen that increases the time complexity of the algorithm to $O(t^2)$. In order to bring back the complexity of the algorithm to $O(t)$ we have toyed with finite memory, as in forgetting earlier events, but this has the effect of removing the exponential term that discounted the probability at the starting point and at the early hits. This is to be avoided because it will attract the searcher very strongly back to where it started.

We think the solution to this is to coarse-grain past events by decimating older events and increasing the weight of the points left. This could leave us with a constant number of points and a precision in integration that's only slightly reduced. We think that the coarse-graining could be performed on the fly according to the position of those points compared with the most recent position of the searcher.

Another exciting new direction we think could be explored is to think infotaxis as the first and simplest strategy in the class of those based on information theory: infotaxis performs choices by looking at the immediate next step, what would happen if we looked several steps ahead?

Such a strategy would translate to adding higher derivatives to the differential equation that regulates the movement of the searcher: the first such step adding an inertial term:

$$\tau^2 (\nabla_x \nabla_x V_t(x)) \ddot{x} + x\gamma \dot{x} = -\nabla_x V_t(x) \,.$$

This inertial term with a mass tensor proportional to the local curvature of the potential at the position of the searcher could have beneficial effects to the performance of the searcher.

Finally we think another interesting direction to take would be to build a meta-heuristic for searches that mimics the behavior we have observed in infotaxis without the need of performing the full entropy calculation. For example we could rethink a technique such as the one developed in [Balkovsky 02] to work in continuous space and three dimensions.

DNA unzipping and sequencing

In chapter 5 we have shown several approaches to the inference of DNA sequencing through micromanipulation experiments. The first issue that stands between us and a successful algorithm is the fact that the number of open bases is not directly known, but acts as a hidden variable, while the position of the bead can be measured directly; the second problem is that the temporal resolution in experiments is very low compared to the time-scale of the opening and closing of the fork.

The second section of this chapter deals with the first problem: the fact that the fork position n is unknown. It does so in a limit which is not completely realistic by imposing that equilibrium is perfectly attained and that we can sample the equilibrium distribution up to an arbitrary precision. In a real experiment there will be many sources of noise and if we take averages for a long enough time we will end up measuring drifts in temperature and trap position which will change the equilibrium distribution.

The approach of the third section, on the other hand takes into account the fact that we could in principle be out of equilibrium and that an infinite sampling frequency is out of the question, but it does so by relying on a very simple model, arguably the simplest non-trivial stochastic process in continuous time and space.

In order to devise a more realistic algorithm we should combine this two approaches, but a few difficulties stand in our way: suppose we took the dynamic approach we have used with the Ornstein-Uhlenbeck and tried to apply it to a more complicated system, our experience tells us that even relaxing the periodic boundary conditions in time makes it hard to diagonalize the covariance matrix analytically.

On the other hand we could try to adapt the idea developed for the perfect averages approach and use them in conjunction with the dynamic algorithm: we could describe the potential on the hidden variable by a simple potential that depends only on a few parameters, that can in turn be fitted, but it is hard to say how a numeric approach can be combined to the dynamical algorithm.

Ultimately we think that many improvements can be brought into play for the experimental procedure if one bears in mind sequencing by unzipping as the ultimate goal. One example are advances in manipulation techniques through holography, allow for the manipulation of multiple beads with a single laser beam [Curtis 02] which could allow for the simultaneous rotation of complex objects. Setups similar to a microscopic bobbin or spindle could one day become feasible if one could find a way to prevent ssDNA from forming secondary structures when confined.

Another idea suggested to us by Prof. A. Libchaber is the use of proteins that bind to ssDNA stiffening it, bringing us somewhat closer to the measurement of the actual fork position.

Part III

Publications

Dynamical modeling of molecular constructions and setups for DNA unzipping

Carlo Barbieri[1], Simona Cocco[1], Rémi Monasson[2] and Francesco Zamponi[2]

[1] LPSENS, Unité Mixte de Recherche (UMR 8550) du CNRS et de l'ENS,
associée à l'UPMC Université Paris 06, 24 Rue Lhomond, 75231 Paris Cedex 05, France
[2] LPTENS, Unité Mixte de Recherche (UMR 8549) du CNRS et de l'ENS,
associée à l'UPMC Université Paris 06, 24 Rue Lhomond, 75231 Paris Cedex 05, France

Abstract
We present a dynamical model of DNA mechanical unzipping under the action of a force. The model includes the motion of a fork in a sequence-dependent landscape, the trap(s) acting on the bead(s) and the polymeric components of the molecular construction (unzipped single strands of DNA and linkers). Different setups are considered to test the model, and the outcome of the simulations is compared to simpler dynamical models existing in the literature where polymers are assumed to be at equilibrium.

1. Introduction

Over the past 15 years, various single molecule experiments have investigated DNA mechanical and structural properties [1–18] and protein–DNA interactions [19–29]. These experiments provide dynamical information usually hidden in large-scale bulk experiments, such as fluctuations on the scale of the individual molecule. The separation of the two strands of a DNA molecule under a mechanical stress, usually referred to as unzipping, was first carried out by Bockelmann and Heslot in 1997 [8]. The strands are pulled apart at a constant velocity while the force necessary for the opening is measured. The average opening force for the λ-phage sequence is about 15 pN (at room temperature and standard ionic conditions), with fluctuations around this value that depend on the particular sequence content. Bockelmann, Heslot and collaborators have shown that the force signal is correlated to the average sequence on the scale of ten base pairs but could be affected by the mutation of one base pair (bp) adequately located along the sequence [10]. Liphardt *et al* [15] and Danilowicz *et al* [16–18] have performed an analogous experiment, using a constant force setup, on a short RNA and long DNA molecules respectively (figure 1(*B*)). The distance between the two strand extremities is measured as a function of the time while the molecule is submitted to

a constant force. The separation of DNA strands has also been studied in single molecule experiments by translocation through nanopores [26, 27].

The motivation underlying unzipping experiments of DNA is (at least) twofold. First, the study of unzipping aims at a better understanding of the mechanisms governing the opening of DNA during transcription and replication by proteins such as polymerases, helicases and exonucleases [20, 21, 28, 29]. Simple theoretical models describing the opening as an unidimensional random walk on a sequence-dependent free energy landscape have been proved to describe quite well several experimental effects such as stick–slip motion in the opening at constant velocity [9, 10], the long pauses at a fixed position of unzipping at constant forces [16, 30, 31], the hopping dynamics between two or more states in unzipping at critical forces of short DNA molecules [15, 31–33] and the torsional drag effects in unzipping at large velocity [11, 34]. Moreover, statistical mechanical analyses have been successfully applied to extract from experimental data the sequence-dependent free energy landscape and the height of free energy barriers [35, 36].

Second, unzipping experiments could potentially be useful to extract information on the sequence itself [37]. Recently, single molecule sequencing has been achieved by monitoring a DNA/RNA polymerase in the course of
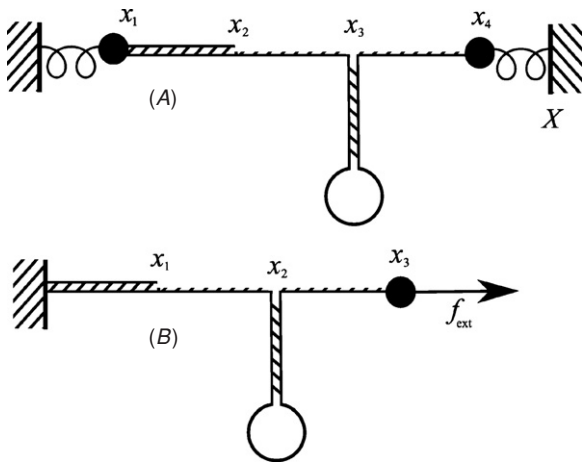
Figure 1. Typical experimental setups that will be described in the following. (*A*) A setup with two optical traps (beads x_1 and x_4) drawn as springs and whose centers are the black vertical lines and (*B*) a setup with a single magnetic bead x_3 that applies a constant force to the molecule attached to a fixed 'wall'. In both cases, the molecular construction is made by a DNA molecule that has to be opened (therefore, one should include two single-strand linkers that are the opened parts of the molecule) and one double-stranded DNA linker. The coordinates x_i are the distances of the corresponding points from the left reference position (which is the center of the left optical trap in case (*A*) and the fixed wall in case (*B*)).

DNA synthesis from a ssDNA template [33, 38]; such single molecule sequencing could become competitive with standard DNA sequencing because they do not require, *a priori*, amplification through polymerase chain reactions. A fundamental question on the possibility of extracting information on the sequence from unzipping experiments is the influence of the experimental setup on the measures and the limitations imposed by the latter [37, 39]. Indeed, characteristic spatio-temporal limitations are the finite rates of data acquisition, the relaxation time of the bead, the limited spatial resolution, the thermal drift and more generally the noise in the instruments. Moreover, the dynamics of the opening fork (figure 1) is influenced by the single strands (open parts) of the molecule and the linkers, and cannot be deduced directly from the observation of the bead from which the force or the position is measured.

The accuracy of unzipping experiments at fixed velocity has improved a lot over the last decade. Initially performed with an optical fiber [8], experiments were then based on the use of simple optical traps [10]. Nowadays, double optical traps [13, 36] allow us to considerably reduce the drift of the setup and to achieve a temporal resolution of the order of 10 kHz, a sub-nanometric spatial resolution, and a precision on measured forces of the order of fraction of pN. Unzipping at fixed force has been performed by a magnetic trap with a low temporal resolution (from 60 Hz to 200 Hz) due to the time needed to extract the position of the bead, the spatial precision being of the order of 10 nm Hz$^{-1/2}$ [28, 29], or by an optical trap also with a low temporal resolution (about 10 Hz) imposed by a feedback mechanism needed to keep the force constant [15]. Recently, a new dumbbell dual optical trap has been developed. It operates without feedback and can

maintain the force constant over distances of about 50 nm [33] with a temporal resolution of 10 kHz and a spatial resolution of 0.1 nm Hz$^{-1/2}$.

Limitations due to the experimental systems were first addressed in [39]. This paper stated the impossibility of inferring the sequence due to ssDNA fluctuations: fluctuations increase with the number of opened base pairs and can become larger than the length of about 1 nm corresponding to the spatial resolution of one open base pair. This problem could however be solved by integrating out the single-strand dynamical fluctuations. Several works have studied the effects of the setup on the hopping dynamics of small RNA molecules [32, 33, 39, 40]. The following effects have been underlined. First, the free energy landscape changes when adding a harmonic potential to the free energy, due to the bead and handles [10, 32, 33, 40]. Therefore, for a given force, the measured separation of the extremities depends on the stiffnesses of the trap and handles. Moreover, the opening and closing rates depend on the stiffness of the optical trap; in particular when the experimental system gets softer the fluctuations of the force gets smaller, and the hopping rates approach their fixed-force values.

In this paper, we introduce a model for the coupled dynamics of the opening fork, the ssDNA strand, the linkers and the bead in the optical or magnetic trap. Essential notions and existing literature are reviewed in section 2. Our dynamical model is presented in section 3. Our program allows us to simulate a generic setup, characterized by bead dimensions, optical stiffness (absent in the case of magnetical tweezers), linker composition (dsDNA or ssDNA) and lengths, and the length of molecule to be unzipped. All the parameters that characterize the different dynamical components can be adjusted in the simulation. The model is then used to simulate fixed-force (section 4) and fixed-extension (section 5) numerical unzippings.

2. Free energies, time scales and effective dynamics

We discuss hereafter the thermodynamic properties of the various parts of the experimental setup (DNA sequence, open part of the molecule, single- or double-strand linkers), as well as the relevant time scales. Finally, we briefly review previous dynamical studies where the linkers and the open portion of the molecules are assumed to be at equilibrium.

2.1. Thermodynamics of the components

2.1.1. Polymeric models for the linkers and open molecule. A polymer model is specified by its free energy as a function of the extension x for a given number n of monomers; we call this quantity $W(x, n)$. When x and n are large, W is an extensive quantity; hence, $W(x, n) = nw(x/n) = nw(l)$, where $l = x/n$ is the extension per monomer. We also define

$$f(l) = \frac{\partial W(x, n)}{\partial x} = w'(l),$$
$$l(f) = \text{inverse of } f(l), \tag{1}$$
$$g(f) = \max_l [fl - w(l)] = fl(f) - w[l(f)],$$

which are, respectively, the force at fixed extension, the average extension at fixed force and the free energy at fixed force. Note that $g(f)$ is simply the integral of $l(f)$. Hence, a polymer model is completely described from the knowledge of the extension versus force characteristic curve, $l(f)$. In the following, we will use some classical models for this function.

- *Gaussian (Hook) model.*

$$l_{\text{Hook}}(f) = \frac{f}{k^m}, \tag{2}$$

where the stiffness constant k^m is related to the temperature T and the average squared monomer length (at zero force) b^2 through $k^m = k_B T/b^2$.

- *Freely-jointed chain (FJC) model.*

$$l_{\text{FJC}}(f) = \coth\left(\frac{fb}{k_B T}\right) - \frac{k_B T}{fb} \tag{3}$$

is the extension (per monomer) of a chain of rigid rods of length b, free to rotate around each other. Comparison of this model with force–extension curves for single-stranded DNA shows that a better fit is obtained from a modified FJC:

$$l_{\text{MFJC}}(f) = d\left(1 + \frac{f}{\gamma_{\text{ss}}}\right) \times l_{\text{FJC}}(f), \tag{4}$$

which takes into account the elasticity effects on the rod length. Standard fit parameters are $d = 0.56$ nm, $b = 1.4$ nm and $\gamma_{\text{ss}} = 800$ pN.

- *Extensible worm-like chain (WLC) model.*

$$l_{\text{WLC}}(f) = L\left[1 - \frac{1}{2}\left(\frac{k_B T}{fA}\right)^{1/2} + \frac{f}{\gamma_{\text{ds}}}\right] \tag{5}$$

is the formula for the high-force extension of an elastic chain with persistence length equal to A. Experiments show that it is an excellent description of double-stranded DNA at high forces, with $L = 0.34$ nm, $A = 48$ nm and $\gamma_{\text{ds}} = 1000$ pN.

2.1.2. Free-energy landscape for the sequence. Let $b_i = A, T, C$ or G denote the ith base along the $5' \rightarrow 3'$ strand (the other strand is complementary) and $B = \{b_1, b_2, \ldots, b_N\}$. The free-energy excess when the first n bp of the molecule is open with respect to the closed configuration ($n = 0$) is [31]

$$G(n; B) = \sum_{i=1}^{n} g_0(b_i, b_{i+1}), \tag{6}$$

where $g_0(b_i, b_{i+1})$ is the binding energy of the bp number i; it depends on b_i (pairing interactions) and on the neighboring bp b_{i+1} due to stacking interactions. g_0 is obtained from the MFOLD server [41, 42], and listed in table 1 for 150 mM NaCl, room temperature and pH 7.5. The values of the free energies should be changed for different ionic conditions and temperatures.

As an illustration, we plot the free energy $G(n; \Lambda)$ of the first 50 bases of the λ-phage sequence, $\Lambda = (\lambda_1, \lambda_2, \ldots, \lambda_N)$, in figure 2 after subtraction of $ng_{\text{ss}}(f)$ for forces $f = 15.9$ and 16.4 pN. $g_{\text{ss}}(f)$ is the work to stretch the two opened single strands when one more bp is opened, and calculated
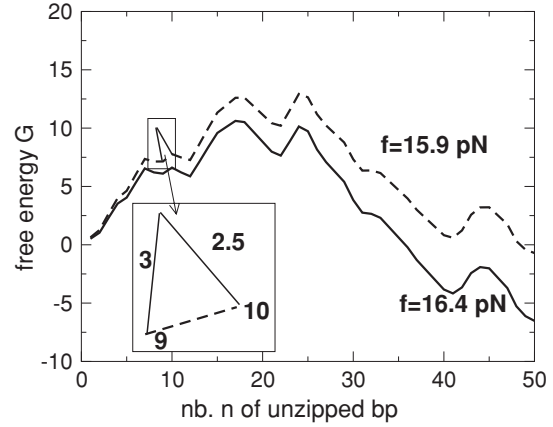
Figure 2. Free energy G (units of $k_B T$) to open the first n base pairs, for the first 50 bases of the DNA λ-phage at forces 15.9 (dashed curve) and 16.4 pN (full curve). For $f = 15.9$ pN, the two minima at bp 1 and bp 50 are separated by a barrier of 12 $k_B T$. Inset: additional barrier representing the dynamical rates (21) to go from base 10 to 9 (barrier equal to $2g_{\text{ss}} = 2.5$ $k_B T$) and from base 9 to 10 (barrier equal to $g_0(b_9, b_{10}) = 3$ $k_B T$); see text.

Table 1. Binding free energies $g_0(b_i, b_{i+1})$ (units of $k_B T$) obtained from the MFOLD server [41, 42] for DNA at room temperature, pH = 7.5 and an ionic concentration of 0.15 M. The base values b_i and b_{i+1} are given by the line and column, respectively.

g_0	A	T	C	G
A	1.78	1.55	2.52	2.22
T	1.06	1.78	2.28	2.54
C	2.54	2.22	3.14	3.85
G	2.28	2.52	3.90	3.14

from the modified FJC model (4). The subtraction allows us to compare the increase in the free energy due to the opening of the sequence to the gain resulting from the release of ssDNA polymers at a given force.

At these forces, the two global minima in figure 2 are located in $n = 1$ (closed state) and $n = 50$ (partially open state). Experiments on a small RNA molecule, called P5ab [15], have been performed at the critical force f_c such that the closed state has the same free energy as the open one: $G(N; \Lambda) = Ng_{\text{ss}}(f_c)$. They showed that, as the barrier between these two minima is not too high, the molecule switches between these two states; see section 2.3.

2.2. Fluctuations at equilibrium

2.2.1. Case of a single polymer. We now consider the orders of magnitude of the fluctuations of the polymer. When submitted to a force of $f = 15$ pN, the average extension of the polymer is $\bar{x} = nx^m$ with $x^m = l(f)$. We use for single-stranded DNA the MFJC model, and for double-stranded DNA the WLC model, with the parameters discussed in section 2.1.1; then we get $x_{\text{ss}}^m = 0.46$ nm and $x_{\text{ds}}^m = 0.33$ nm for ss- and dsDNA respectively. At thermal equilibrium, the extension will fluctuate around these average values. The fluctuations are controlled by the *microscopic effective spring constant* $k^m(l) = w''(l) = 1/l'(f)$. For ds-

Table 2. Fluctuations of single-stranded DNA at $f = 15$ pN and $T = 16.7\,°C$; $\delta\bar{x}/\bar{x} = 0.37/\sqrt{n}$, $\delta\bar{f}/\bar{f} = 1.57/\sqrt{n}$, $\tau = 4.83 \times 10^{-11}\mathrm{s}n^2$.

n	$\delta\bar{x}/\bar{x}$	$\delta\bar{f}/\bar{f}$	τ (s)
10	0.117	0.496	4.8×10^{-9}
40	0.058	0.248	7.7×10^{-8}
100	0.037	0.157	4.8×10^{-7}
400	0.018	0.078	7.7×10^{-6}
1000	0.012	0.050	4.8×10^{-5}

Table 3. Fluctuations of double-stranded DNA at $f = 15$ pN and $T = 16.7\,°C$; $\delta\bar{x}/\bar{x} = 0.17/\sqrt{n}$, $\delta\bar{f}/\bar{f} = 4.83/\sqrt{n}$, $\tau = 5.1 \times 10^{-12}\mathrm{s}n^2$.

n	$\delta\bar{x}/\bar{x}$	$\delta\bar{f}/\bar{f}$	τ (s)
100	0.017	0.483	5.1×10^{-8}
400	0.0085	0.241	8.1×10^{-7}
1000	0.0054	0.153	5.1×10^{-6}
4000	0.0027	0.076	8.1×10^{-5}
10000	0.0017	0.048	5.1×10^{-4}

and ssDNA we find, respectively, $k_{ds}^m = 1311$ pN nm^{-1} and $k_{ss}^m = 138$ pN nm^{-1} according to the above models. For a polymer with n monomers, the stiffness is $k = k^m/n$ since the effective spring constant is given by $k(x, n) = \frac{\partial^2}{\partial x^2}W(x, n) = k^m(x/n)/n$.

Alternatively, the force f exerted on the polymer will fluctuate around its average value $\bar{f}$ if its extremities are kept at a fixed distance x from each other. These fluctuations of force (in the fixed-extension ensemble) and extension (in the fixed-force ensemble) are easily computed by a quadratic expansion of the free energy around the average, i.e. when approximating the polymer with a spring of stiffness k^m/n, with the result

$$\langle \delta x^2 \rangle = \frac{k_B T}{k^m}n, \qquad \langle \delta f^2 \rangle = \frac{k_B T k^m}{n}. \qquad (7)$$

Defining $\delta\bar{x} = \sqrt{\langle \delta x^2 \rangle}$ and $\delta\bar{f} = \sqrt{\langle \delta f^2 \rangle}$, we get

$$\frac{\delta\bar{x}}{\bar{x}} = \sqrt{\frac{k_B T}{k^m(x^m)^2}}\frac{1}{\sqrt{n}}, \qquad \frac{\delta\bar{f}}{\bar{f}} = \sqrt{\frac{k_B T k^m}{\bar{f}^2}}\frac{1}{\sqrt{n}}. \qquad (8)$$

As expected, the relative fluctuations of both force and extension become smaller and smaller as the number n of monomers increases. Some values are reported in tables 2 and 3.

2.2.2. Case of several polymers (fixed-distance setup). Now consider the case of several polymers, e.g. linker and open part of the molecule attached one after the other. In a fixed-force experiment, the components of the setup are independent (at the level of the saddle-point approximation) and the fluctuations in the extensions simply add up. In the fixed-distance setup, however, correlations between the extensions make the analysis more complicated. As a concrete example, we consider the setup in figure 1(*A*). The linker joining x_1 and x_2 is a double-stranded DNA segment of N_{ds} bases. The two linkers joining (x_2, x_3) and (x_3, x_4) are single-stranded DNA segments of $N_{ss} = N_{ss}^0 + n$ bases, where n is the number of opened base pairs.

The centers of the two optical traps are at 0 and X. We call x_1 the position of the first bead and x_4 the position of the second. The probability $P_{eq}(n, x_1, x_2, x_3, x_4) = e^{-F/k_B T}$, where the free energy F reads as

$$F(\vec{x}, n) = \tfrac{1}{2}k_1 x_1^2 + W_{ds}(x_2 - x_1, N_{ds}) + W_{ss}(x_3 - x_2, N_{ss})$$
$$+ W_{ss}(x_4 - x_3, N_{ss}) + \tfrac{1}{2}k_2(x_4 - X)^2 + G(n; B), \qquad (9)$$

where $W_{ds}(x, N_{ds}) = N_{ds}w_{ds}(x/N_{ds})$ and $W_{ss}(x, N_{ss}) = N_{ss}w_{ss}(x/N_{ss})$ are the elongation free energies of the double strand and single strand, respectively.

In order to study the fluctuations in this setup, we first find the maximum of P_{eq} assuming that $G(n; B) = ng_0$, i.e. a uniform sequence B, and treating n as a continuous variable assuming that it is large. At the maximum $x_i = \bar{x}_i$ and we define

$$x_{ds}^m = \frac{\bar{x}_2 - \bar{x}_1}{N_{ds}}, \qquad x_{ss}^m = \frac{\bar{x}_3 - \bar{x}_2}{N_{ss}} = \frac{\bar{x}_4 - \bar{x}_3}{N_{ss}}. \qquad (10)$$

The saddle-point condition $\partial_{x_i} F_A = 0$ gives the following equations, which represent the force balance condition along the chain:

$$k_1\bar{x}_1 = w'_{ds}(x_{ds}^m) = w'_{ss}(x_{ss}^m) = k_2(X - \bar{x}_4) \equiv \bar{f}. \qquad (11)$$

The derivative with respect to n gives, using equations (1) and (11), the condition

$$g_0 = 2[x_{ss}^m w'_{ss}(x_{ss}^m) - w_{ss}(x_{ss}^m)] = g_{ss}(\bar{f}), \qquad (12)$$

which allows us to find the force $\bar{f}$ transmitted along the chain. Once (12) is solved, the extensions of the beads and of the double- and single-stranded parts of DNA ($\bar{x}_1$, $X - \bar{x}_4$, x_{ds}^m and x_{ss}^m respectively) are determined by equation (11). Finally, the number of open bases $\bar{n}$ is determined by

$$\bar{x}_1 + N_{ds}x_{ds}^m + 2(N_{ss}^0 + \bar{n})x_{ss}^m + (X - \bar{x}_4) = X. \qquad (13)$$

Note that the value of $\bar{f}$ is determined only by g_0.

We work at temperature $T = 16.7\,°C$ ($k_B T = 4$ pN nm) and choose a uniform molecule with $g_0 = 2.69k_B T$, which is a representative value for the pairing free energies in table 1. We use the same models as in section 2.2.1 for the single- and double-stranded DNA, with $N_{ds} = 3120$ and $N_{ss}^0 = 40$. Then solving equation (12) we get $\bar{f} = 16.5$ pN, and from equation (11) we get $x_{ss}^m = 0.47$ nm, $x_{ds}^m = 0.33$ nm. We choose $k_1 = 0.1$ pN nm^{-1}, then $\bar{x}_1 = 165$ nm, and $k_2 = 0.512$ pN nm^{-1}, then $X - \bar{x}_4 = 32$ nm. Given these values, $\bar{n}$ is defined by X using equation (13):

$$\bar{n} = \frac{X - 1264}{0.94}, \qquad (14)$$

with X expressed in nanometers.

For the same setup, we can compute the fluctuations of n and of the elongations of the elements of the setup. In particular, the fluctuations of the bead positions are measurable in the experiment.

Let us define $\delta x_i = x_i - \bar{x}_i$ and $\delta n = n - \bar{n}$. To simplify the formalism, we also define $\delta x_{ds} = \delta x_2 - \delta x_1$, $\delta x_{ss}^L = \delta x_3 - \delta x_2$ and $\delta x_{ss}^R = \delta x_4 - \delta x_3$. A quadratic expansion of F around its minimum gives

$$\delta F \sim \frac{1}{2}k_1\delta x_1^2 + \frac{1}{2}k_2\delta x_4^2 + \frac{w''_{ds}(x_{ds}^m)}{2N_{ds}}\delta x_{ds}^2$$
$$+ \frac{w''_{ss}(x_{ss}^m)}{2N_{ss}^0 + \bar{n}}[(\delta x_{ss}^L - x_{ss}^m\delta n)^2 + (\delta x_{ss}^R - x_{ss}^m\delta n)^2]. \qquad (15)$$

Using (4) and (5), we get $k_{ss}^m = w_{ss}''(x_{ss}^m) = 152$ pN nm^{-1} and $k_{ds}^m = w_{ds}''(x_{ds}^m) = 1416$ pN nm^{-1}.

One should take care of the fact that $\delta x_1 + \delta x_4 + \delta x_{ds} + \delta x_{ss}^L + \delta x_{ss}^R = 0$; it is convenient to express δx_{ss}^R as a function of the others since its fluctuations are identical to those of δx_{ss}^L. The quadratic expansion of the function δF has the form $\delta F = \frac{1}{2}\delta \mathbf{x} A \delta \mathbf{x}$ where $\delta \mathbf{x} = (\delta x_1, \delta x_4, \delta x_{ds}, \delta x_{ss}^L, x_{ss}^m \delta n)$ and

$$A = \frac{k_{ss}^m}{N_{ss}^0 + \bar{n}} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 0 \\ 1 & 1 & 1 & 0 & 2 \end{pmatrix}$$

$$+ \begin{pmatrix} k_1 & 0 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 & 0 \\ 0 & 0 & k_{ds}^m/N_{ds} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (16)$$

The inverse of the matrix A is

$$A^{-1} = \begin{pmatrix} \frac{1}{k_1} & 0 & 0 & -\frac{1}{2k_1} & -\frac{1}{2k_1} \\ 0 & \frac{1}{k_2} & 0 & -\frac{1}{2k_2} & -\frac{1}{2k_2} \\ 0 & 0 & \frac{N_{ds}}{k_{ds}^m} & -\frac{N_{ds}}{2k_{ds}^m} & -\frac{N_{ds}}{2k_{ds}^m} \\ -\frac{1}{2k_1} & -\frac{1}{2k_2} & -\frac{N_{ds}}{2k_{ds}^m} & \frac{1}{4k_{eff}} & \frac{1}{4k_{eff}^s} \\ -\frac{1}{2k_1} & -\frac{1}{2k_2} & -\frac{N_{ds}}{2k_{ds}^m} & \frac{1}{4k_{eff}^s} & \frac{1}{4k_{eff}} \end{pmatrix}, \quad (17)$$

where

$$\frac{1}{k_{eff}^s} = \frac{1}{k_1} + \frac{1}{k_2} + \frac{N_{ds}}{k_{ds}^m}, \qquad \frac{1}{k_{eff}} = \frac{1}{k_{eff}^s} + 2\frac{N_{ss}^0 + \bar{n}}{k_{ss}^m}. \quad (18)$$

This immediately gives

$$k_B T (A^{-1})_{1,1} = \langle \delta x_1^2 \rangle = \frac{k_B T}{k_1}$$

$$k_B T (A^{-1})_{2,2} = \langle \delta x_4^2 \rangle = \frac{k_B T}{k_2}$$

$$k_B T (A^{-1})_{3,3} = \langle \delta x_{ds}^2 \rangle = \frac{k_B T N_{ds}}{k_{ds}^m} \quad (19)$$

$$k_B T (A^{-1})_{4,4} = \langle (\delta x_{ss}^L)^2 \rangle = \frac{k_B T}{4k_{eff}}$$

$$\frac{k_B T}{(x_{ss}^m)^2} (A^{-1})_{5,5} = \langle \delta n^2 \rangle = \frac{k_B T}{4k_{eff}(x_{ss}^m)^2}$$

and shows that the fluctuations of n are dominated by the weakest element of the setup; moreover, the correlation between the bead displacements δx_1, δx_4 and the fluctuations of the number of open base pairs δn is $\langle \delta n \delta x_1 \rangle = -\frac{k_B T}{2k_1 x_{ss}^m}$ and $\langle \delta n \delta x_4 \rangle = -\frac{k_B T}{2k_2 x_{ss}^m}$; the stiffer the optical trap, the weaker is the correlation between the location of the bead and the number of open bases. Examples are given in table 4.

2.3. Effective dynamical models

In the simplest dynamical models, the fork (separating the open and closed portions of the molecule) undergoes a biased random motion in the sequence landscape. The linkers are treated at equilibrium, which is correct if their characteristic time scales are much smaller than the average time needed to open or close a base pair.

Table 4. Saddle-point calculation for the setup in figure 1(A) with a uniform molecule and $k_1 = 0.1$ pN nm^{-1}, $k_2 = 0.512$ pN nm^{-1}, $N_{ds} = 3120$, $N_{ss}^0 = 40$. The force along the molecule is $\bar{f} = 16.5$; then $k_{ss}^m = 152$ pN nm^{-1}, $k_{ds}^m = 1416$ pN nm^{-1} and $k_{eff}^s = 0.07$ pN nm^{-1}.

X	$\bar{n}$	k_{eff}	$\sqrt{\langle \delta n^2 \rangle}$
1273	10^1	0.067	8.2
1358	10^2	0.062	8.5
2204	10^3	0.036	11.2
10664	10^4	0.0068	25.7

2.3.1. Time scales for the polymeric components of the setup.

In this section, we recall the typical time scales of the polymeric components in the setup. Assume that the polymers are subject to a Brownian force $\eta(t)$ which is a zero-average Gaussian process with an autocorrelation function $\langle \eta(t)\eta(0) \rangle = 2\Gamma T \delta(t)$. Let Γ be the friction coefficient of the polymer [43], that is, the ratio of the viscous force exerted by the solvent to the velocity. As will be shown in section 3, the friction coefficient scales as $\Gamma = \gamma^m n/3$ with $\gamma_{ss}^m = \gamma_{ds}^m \sim 2 \times 10^{-8}$ pN s nm^{-2}. Then, approximating $f(x, n) \sim k^m x/n$, the relaxation time for an isolated polymer of n bases is given by

$$\tau = \frac{\gamma^m n^2}{3k^m}. \quad (20)$$

Note that the factor 3 in the denominator of the above equation is an approximation for the true factor $\pi^2/4$. The validity of its approximation and the simplification it leads to will be discussed in appendix A.

It is useful to compare the amplitude of the force fluctuations with the noise. To do this, we approximate $\langle \delta f(t)\delta f(0) \rangle \sim 2\tau \langle \delta f^2 \rangle \delta(t) = 2T\Gamma_f \delta(t)$. Then, using equation (7) to estimate $\langle \delta f^2 \rangle$, we get $\Gamma_f = n\gamma^m/3 = \Gamma$, and (not surprisingly) the force fluctuations are of the same order as the noise term.

From table 2, the relaxation time of the unzipped strands is smaller than the typical base-pair opening (or closing) time as long as the number n of unzipped bases is smaller than a few hundreds. This is the case, in particular, for unzipping experiments on short RNA molecules.

2.3.2. Random walk in the sequence landscape.

Let us first model the motion of the fork alone, that is, assuming that the other components of the setup are at equilibrium. We consider a DNA molecule unzipped under a fixed force f in the sequence-landscape $G(n; B) - ng_{ss}(f)$ of figure 2. The fork, whose position is denoted by $n(t)$, can move forward $(n \to n+1)$ or backward $(n \to n-1)$ with rates (probability per unit of time) equal to, respectively,

$$r_o(b_{n+1}, b_{n+2}) = r \exp[-\beta g_0(b_{n+1}, b_{n+2})],$$
$$r_c = r \exp[-2\beta g_{ss}(f)], \quad (21)$$

where $\beta = 1/k_B T$; see figure 2. The value of the attempt frequency r is of the order of 10^6 Hz [12, 14, 31]. Expression (21) for the rates is derived from the following assumptions. First, the rates should satisfy detailed balance. Second, we impose that the opening rate r_o depends on the binding free
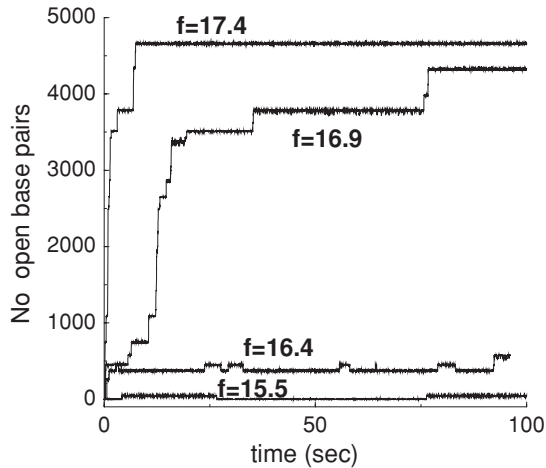
Figure 3. Number of open base pairs as a function of the time for various forces (shown in the figure). Data show one numerical unzipping (for each force) obtained from a Monte Carlo simulation of the random walk motion of the fork with rates (21).

energy, and not on the force, and vice versa for the closing rate r_c. This choice is motivated by the fact that the range for the base-pair interaction is very small: the hydrogen and stacking bonds are broken when the bases are kept apart at a fraction of an Angstrom, while the force work is appreciable on the distance of the opened bases ($\approx$ 1 nm). In contrast, to close the base pairs, one has to first work against the applied force; therefore, the closing rate r_c depends on the force but not on the sequence. This physical origin of the rates is reported in the inset of figure 2. Note that, as room temperature is much smaller than the thermal denaturation temperature, we safely discard the existence of a denatured bubble in the zipped DNA portion.

An example of unzipping dynamics for the λ-phage sequence is shown in figure 3. The characteristic pauses in the unzipping, present in experiments and corresponding to deep local minima in the sequence landscape, are reproduced. The rates (21) lead to a master equation for the probability $\rho_n(t)$ for the fork to be at site n at time t:

$$\frac{d\rho_n(t)}{dt} = -\sum_{m=0}^{N} T_{n,m}\rho_m(t), \qquad (22)$$

where the matrix $T_{n,m}$ is tridiagonal with nonzero entries $T_{m-1,m} = -r_c(f)$, $T_{m+1,m} = -r_o(m)$ and $T_{m,m} = r_o(m) + r_c(f)$. Given this transition matrix, the opening dynamics can be simulated with Monte Carlo dynamics. For small RNA or DNA molecules, the transition matrix $T_{n,m}$ can be diagonalized numerically [31]. The smallest non-zero eigenvalue gives the switching time between a closed and open configuration for a hairpin with a free energy barrier such as that plotted in figure 2.

2.3.3. Dynamics of the bead with equilibrated linkers and strands. In a typical experiment, the force is exerted on the molecule through the action of a (magnetic or optical) trap on the bead. While the external force on the bead can be considered as constant (e.g. in a magnetic trap), the force

acting on the fork fluctuates unless the trap (and the molecular construction) is very soft; see equation (8). Therefore, the fixed-force model of the previous section has to be modified. In addition the bead, of size $R \simeq 1 \mu$m, is a slow component whose dynamics need to be taken into account. Let us denote by k the stiffness of the trap and by γ the friction of the bead in the solvent of viscosity η. Typical values for these quantities are $k = 0.1$–0.5 pN nm^{-1} and $\gamma = 6\pi R\eta = 1.67 10^{-5}$ pN s nm^{-1}. Thus, the characteristic relaxation time of the bead is $\tau = \gamma/k \simeq 0.2$–1 ms.

The coupled dynamics of the fork and the bead was considered by Manosas *et al* [14] in the case of small RNA unzipping, with a single optical trap. For such small molecules the relaxation time of the unzipped strands is expected to be much smaller than the characteristic time of the bead, and the molecule can be considered at equilibrium. The dynamical scheme therefore consists in a coupled evolution equation for the location of the bead and of the fork. The bead position obeys a Langevin equation including the external force and the force exerted by the fork through the (equilibrated) linkers and unzipped strands, while the fork moves with rates (21) with a bead location-dependent force.

A main conclusion of [14] is that, in the absence of feedback imposing a fixed force on the molecule, the trap stiffness must be as low as possible to detect jumps between closed and open configurations of the RNA molecule. We will discuss the validity of this statement in an information-theoretic setting in section 5.2.

3. Dynamical modeling of the setup and its components

The assumption that the linkers and the unzipped strands are at equilibrium as the unzipping proceeds is correct for short molecules as was the case in [14]. For long DNA molecules, the relaxation time of the unzipped strands may become large and dynamical modeling of the polymers involved in the molecular construction cannot be avoided.

The purpose of this section is to describe how such a dynamical model can be implemented. We hereafter denote by 'setup' the full molecular construction that is used in a given experiment, including linkers, beads, etc, while the word 'molecule' refers to the part of DNA which has to be opened. In an idealized description, the state variable is a vector $\vec{x} = (x_1, \ldots, x_p)$ whose elements are the distances from a reference position (that can be either the center of an optical trap or a fixed 'wall' to which the polymers are attached) of the extremities of the polymeric components in the setup. In addition to $\vec{x}$, the number of open base pairs n is needed to complete the description of the state of the setup.

As discussed in section 2.1, the total free energy $F(\vec{x}, n)$ of a setup is the sum of different contributions coming from all the elements of the setup. A typical example is given in equation (9).

Our aim is thus to construct a dynamical model that holds on intermediate time scales, $t \gtrsim 10^{-6}$ s, and

(i) gives the correct equilibrium Gibbs measure $P_{eq}(\vec{x}, n) = \exp(-F(\vec{x}, n)/(k_B T))$,

(ii) reproduces the relaxation times for the different elements of the setup, as discussed below,

(iii) gives reasonable dynamical correlations between different elements of the setup.

It is worth stressing at this point that ours is a coarse-grained model which does not take into account the motion of the individual monomers. It is expected that the dynamics on time scales smaller than the typical sojourn time of the fork on a base ($\gtrsim 10^{-6}$ s) is not relevant to our study of unzipping.

3.1. Langevin dynamics for the polymers and the beads

First, we consider the dynamics of $\vec{x}$ at fixed n. In appendix A, we show that for long enough times the dynamics of the setup can be described by a system of coupled Langevin equations:

$$\Gamma_{ij}\dot{x}_j = -\frac{\partial F}{\partial x_i} + \eta_i, \tag{23}$$

where $i, j = 1, \ldots, p$, and

- the free energy $F(\vec{x})$ is the sum of a contribution coming from each element of the setup:
  - (i) each optical trap contributes $\frac{1}{2}k\Delta x^2$, where Δx is its elongation;
  - (ii) a bead in position i subjected to a constant force gives a contribution $-fx_i$;
  - (iii) a polymer gives a contribution $W_i(\Delta x, N_i)$, with Δx being its elongation and N_i its number of monomers.

  For example, the total free energies of the setups in figure 1 are

  $$\begin{aligned}
  F_A(\vec{x}) &= \tfrac{1}{2}k_1 x_1^2 + W_{\mathrm{ds}}(x_2 - x_1, N_{\mathrm{ds}}) + W_{\mathrm{ss}}(x_3 - x_2, N_{\mathrm{ss}}) \\
  &\quad + W_{\mathrm{ss}}(x_4 - x_3, N_{\mathrm{ss}}) + \tfrac{1}{2}k_2(x_4 - X)^2, \\
  F_B(\vec{x}) &= W_{\mathrm{ds}}(x_1, N_{\mathrm{ds}}) + W_{\mathrm{ss}}(x_2 - x_1, N_{\mathrm{ss}}) \\
  &\quad + W_{\mathrm{ss}}(x_3 - x_2, N_{\mathrm{ss}}) - f x_3.
  \end{aligned} \tag{24}$$

- $\vec{\eta}$ is a Gaussian white noise with zero average and variance $\langle \eta_i(t)\eta_j(0)\rangle = 2k_B T \Gamma_{ij}\delta(t)$, as requested by the fluctuation–dissipation relation.
- the matrix Γ is a tridiagonal matrix such that
  - (i) the diagonal element Γ_{ii} is the sum of three contributions:
    - (a) a term $\gamma_{i-1}^m N_{i-1}/3 + \gamma_i^m N_i/3$ coming from the adjacent polymers (if any);
    - (b) a term γ coming from the bead (if any) attached to x_i;
    - (c) a term taking into account the viscosity of the N_c base pairs of the DNA molecule attached to the fork (x_3 and x_2 in figures 1(A) and (B) respectively) that are not open; this term has the Fleury form $\gamma_{\mathrm{mol}} = \gamma' N_c^{3/5}$ and has to be added to the diagonal element of Γ corresponding to the fork position;
  - (ii) the offdiagonal elements are zero, except $\Gamma_{i,i+1} = \Gamma_{i+1,i} = \gamma_{i+1}^m \frac{N_{i+1}}{6}$ that get a contribution from the polymer joining x_i and x_{i+1}.

For instance, the setups in figure 1 correspond to the matrices:

$$\Gamma_B = \begin{pmatrix} \gamma_{\mathrm{ds}}^m \frac{N_{\mathrm{ds}}}{3} + \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{3} & \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{6} & 0 \\ \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{6} & 2\gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{3} + \gamma' N_c^{3/5} & \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{6} \\ 0 & \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{6} & \gamma + \gamma_{\mathrm{ss}}^m \frac{N_{\mathrm{ss}}}{3} \end{pmatrix},$$

$$\Gamma_A = \begin{pmatrix} \gamma + \gamma_{\mathrm{ds}}^m \frac{N_{\mathrm{ds}}}{3} & \gamma_{\mathrm{ds}}^m \frac{N_{\mathrm{ds}}}{6} & 0 & 0 \\ \gamma_{\mathrm{ds}}^m \frac{N_{\mathrm{ds}}}{6} & & & \\ 0 & & \Gamma_B & \\ 0 & & & \end{pmatrix}. \tag{25}$$

A detailed derivation of these results and in particular of the form of the matrix Γ can be found in appendix A.

3.2. Fork dynamics

The Langevin equation for the polymer dynamics at fixed n must be complemented with transition rates for the dynamics of n. To this aim, we discretize the Langevin equation with time step Δt, and at each time step we allow the opening $n \to n+1$ or closing $n \to n-1$ of a base pair at most.

The dynamics takes the form of a discrete time Markov chain, with transitions $(\vec{x}, n) \to (\vec{x}', n')$ and $n' \in \{n, n \pm 1\}$. The total free energy $F(\vec{x}, n) = F_{\mathrm{setup}}(\vec{x}, n) + G(n; B)$, where the first contribution has been discussed in the previous section and $G(n; B)$ is the pairing free energy of the molecule, as discussed in section 2.1.2. In appendix B, we show that in order to satisfy the detailed balance condition with respect to $P_{\mathrm{eq}}(\vec{x}, n) = \exp(-F(\vec{x}, n)/(k_B T))$, one should perform a single step following the procedure.

(i) Choose whether to stay ($n' = n$), to open ($n' = n + 1$) or to close ($n' = n - 1$) a base, with rates $r^{s,o,c}(\vec{x}, n)$ respectively:

$$\begin{aligned}
r^o(\vec{x}, n) &= r\Delta t\, e^{\beta[G(n;B) - G(n+1;B)]}, \\
r^c(\vec{x}, n) &= r\Delta t\, e^{\beta F(\vec{x}, n) - \beta F(\vec{x}, n-1)}, \\
r^s(\vec{x}, n) &= 1 - r^o(\vec{x}, n) - r^c(\vec{x}, n).
\end{aligned} \tag{26}$$

(ii) If the choice was to open, *first* perform a discrete Langevin step $\vec{x} \to \vec{x}'$ at fixed n and *then* increase n by one.

(iii) If the choice was to close, *first* decrease n by one and *then* perform a discrete Langevin step $\vec{x} \to \vec{x}'$ at fixed $n' = n - 1$.

(iv) If the choice was to stay, just perform a discrete Langevin step $\vec{x} \to \vec{x}'$ at fixed n.

The Langevin equation is discretized in a standard way by integrating equation (23) over a time Δt:

$$x_i(t + \Delta t) = x_i(\Delta t) + \Gamma_{ij}^{-1}\left[-\frac{\partial F(\vec{x})}{\partial x_j}\Delta t + E_j\right], \tag{27}$$

where $E_j = \int_0^{\Delta t} \eta_j(t)\,dt$ are Gaussian variables with zero average and variance

$$\langle E_i E_j \rangle = 2k_B T \Gamma_{ij}\Delta t \tag{28}$$

that are independently drawn at each discrete time step.

3.3. Free energy at finite n

In section 2.1, we discussed some models for the free energy $W(x, n)$ of a polymer with n monomers and extension x. In the limit $x, n \to \infty$ at fixed extension per monomer, $l = x/n$, the free energy enjoys an extensivity property: $W(x, n) = nw(l)$. However, in our simulations we might be interested in regimes where n is small, typically of the order of 10–40 for small RNA molecules. In this case, knowledge of the free energy per monomer, w, is not sufficient, and a more detailed expression for W is necessary to avoid inconsistencies.

As a starting point of the analysis, we consider a polymer made of N identical monomers whose endpoints are denoted by $u_i, i = 1, \ldots, N$ with $u_0 = 0$. The Hamiltonian of the chain is the sum of pairwise interactions $\varphi(u_i - u_{i-1})$ and the free energy reads, for $x = u_N$, as

$$e^{-\beta W(x,n)} = \ell_0^{-N+1} \int du_1, \ldots, du_{N-1} \, e^{-\beta \sum_i \varphi(u_i - u_{i-1})}, \quad (29)$$

where ℓ_0 is a reference microscopic length scale. From the above relation, the Chapman–Kolmogorov equation follows:

$$e^{-\beta W(x,n+m)} = \ell_0^{-1} \int dy \, e^{-\beta W(y,n) - \beta W(x-y,m)}. \quad (30)$$

We first consider for simplicity the Gaussian model, $\varphi(x) = \frac{1}{2} k^m x^2$. Then it is easy to show that

$$W(x, n) = \frac{k^m}{2n} x^2 - \frac{k_B T}{2} \log \left[\frac{k \ell_0^2}{2\pi k_B T n} \right]. \quad (31)$$

In the limit of large polymers, one obtains the free energy of a monomer of extension l through

$$w(l) = \lim_{n \to \infty} \frac{1}{n} W(x = ln, n) = \varphi(l) \quad (32)$$

as expected and consistent with the discussion of section 2.1. The logarithmic term in (31) contributes neither to w nor to the Langevin equation for x. However it does contribute to the rate to close a base pair (see equation (26)) and should be taken into account in order to recover the correct rates. An example of the effect of this term is obtained by computing the equilibrium probability of n. Consider the (unrealistic) case of a homopolymer, $G(n; B) = ng_0$, subject to a constant force and using a Gaussian model for the open part of the molecule; then

$$\begin{aligned} P_{eq}(n) &= \frac{1}{Z} \int dx \, e^{-n\beta g_0 - \beta W(x,2n) + \beta f x} \\ &= \frac{1}{Z'} e^{-n\beta g_0 + \frac{n}{k} f^2}. \end{aligned} \quad (33)$$

Therefore $P_{eq}(n)$ is a pure exponential, while if the correction were neglected one would have obtained wrong behavior at small n.

For a generic model of $\varphi(x)$, one cannot compute $W(x, n)$. Still we found that for our purposes $(n \gtrsim 40)$, a consistent approximation is obtained by keeping only the first correction to the $n \to \infty$ result, i.e. by defining

$$e^{-\beta W(x,n)} = e^{-\beta n w(x/n)} \sqrt{\frac{\beta k(x/n) \ell_0^2}{2\pi n}}, \quad (34)$$

where $k(l) = w''(l)$. One can check that this expression satisfies equation (30) with corrections in the exponent of $O(1)$, while the terms $O(\log n + \log m)$ are taken into account. Within this approximation, the error in $\log r_c(x, n)$ in equation (26) is $O(1/n^2)$ while if the first corrections are neglected it is $O(1/n)$.

In the following, we will make use of definition (34) unless otherwise stated. We will discuss an example where the effects of neglecting the corrections are clearly observable.

3.4. Details of the numerical simulations

We performed numerical simulations of the molecular constructions depicted in figure 1, with the following specifications.

- The total free energies of the two setups are given by equation (24) plus the term $G(n; B)$.
- The free energy of each polymer includes the saddle-point corrections, i.e. it is given by equation (34). The relation $l(f)$ (see section 2.1) is numerically inverted to obtain $w(l)$ and $k(l)$ that enter in equation (34).
- For the single-stranded DNA we used the MFJC model, equation (4), with $d = 0.56$ nm, $b = 1.4$ nm and $\gamma_{ss} = 800$ pN.
- For the double-stranded DNA we used the WLC model in equation (5), with a small regularization term to avoid a divergence for $f \to 0$, which is however irrelevant for values of forces to be discussed in the following, and with $A = 48$ nm, $L = 0.34$ nm and $\gamma_{ds} = 1000$ pN.
- Unless otherwise stated, the double-stranded DNA linker is made of $N_{ds} = 3120$ bps, while the two single-stranded linkers are made of $N_{ss} = 40 + n$ bases each, where n is the number of open DNA bases (in other words, we included on each side a 40-base single-stranded linker).
- We worked at fixed temperature $k_B T = 4$ pN nm, corresponding to $T = 16.7$ °C.
- We used the dynamical equations for the polymers defined above, equations (23), within the discrete procedure illustrated in section 3.2 and with transition rates (26) for the fork with the attempt rate $r = 10^6$ Hz.
- The matrices Γ corresponding to the setups in figure 1 are given in equation (25); we used $\gamma_{ds}^m = \gamma_{ss}^m = \gamma' = 2 \times 10^{-8}$ pN s nm^{-1}. We used a value $\gamma = 1.67 \times 10^{-5}$ pN s nm^{-1} for the viscosity of the beads.
- The time step was fixed at $\Delta t = 10^{-8}$ s; this value ensures a correct integration of the equation of motion in all the regimes discussed below. Even if in some cases a larger integration step could be used, we decided to keep it fixed in order to be sure that discretization biases are not present.

The values of the spring constants k_1 and k_2 and of the force f in equation (24) varied in different simulation runs, and will be specified later.

The program we used for the numerical simulations can be downloaded from http://www.lpt.ens.fr/ zamponi. A user-friendly version will be made available as soon as possible.

3.5. Limits of validity of the dynamical model

Our model of the polymer dynamics suffers from two main limitations.

First, we keep only one collective coordinate for each polymer (its extension) associated with the longest relaxation mode. Faster modes are discarded. The approximation is justified provided there is no other mode slower than the typical sojourn time on a base pair. From the discussion of section 2.3.1, the number of unzipped base pairs, n, cannot be well above a thousand.

Another upper limit on n comes from the assumption that the force is uniform along the polymer. In principle the force is a function of the time t and the location y along the polymer, which obeys a diffusion equation with a microscopic diffusion coefficient $D_{ss}^m \simeq \left(x_{ss}^m\right)^2 / \tau_{ss}^m$, where x_{ss}^m is the length of a monomer and $\tau_{ss}^m = \gamma_{ss}^m / k_{ss}^m$ is its relaxation time. Assume that, at time 0, a base pair closes and the polymer is stretched at the extremity $x = 0$ by x_{ss}^m. Then the force, initially equal to $f(x, t = 0) = k_{ss}^m x_{ss}^m \delta(x)$, will decay following the Gaussian diffusion kernel. At time t, the force density at the extremity is $f(x, t) = k_{ss}^m x_{ss}^m / \sqrt{2\pi D_{ss}^m t}$. The relaxation is over when this force excess is of the same order of magnitude as the typical thermal fluctuations δf calculated in (8), that is, for times

$$t > n \frac{k_{ss}^m \left(x_{ss}^m\right)^2}{2\pi k_B T} \tau_{ss}^m \simeq 2 \times 10^{-10} n \text{ ps.} \quad (35)$$

When $n \sim 1000$, the corresponding relaxation time is of the order of the sojourn time on a base.

In conclusion, our dynamical model is adapted to ssDNA polymers whose length ranges from a few hundred to a few thousand bases. Shorter polymers can be considered at equilibrium, while longer polymers cannot be modeled without taking into account the space dependence of forces. A simple way to tackle this difficulty consists in arbitrarily cutting long polymers into 1000-base long segments, each modeled as above. This procedure will be followed in section 5.1.

4. Unzipping at fixed force

4.1. Quasi-equilibrium unzipping

Before turning to the more interesting case of out-of-equilibrium unzipping, we focus on the case of a small molecule which is subject to a constant force close to the critical force. In this situation, the molecule is able to visit all the possible configurations.

We performed a set of numerical simulations at constant force $\bar{f} = 16.45$ pN, with the setup described in figure 1(B). The DNA molecule is a uniform segment of $N = 500$ base pairs, with pairing free energy $G(n; B) = ng_0$ and $g_0 = 2.69k_B T$. The entropic free energy per base of the two open single strands is $2g_{ss}(\bar{f}) = 2.684k_B T$. Therefore, the infinite molecule would stay close; we are slightly below the critical force. To the right and left open portions of the molecule, two single-stranded DNA linkers of $N_{ss}^0 = 40$ bases each are attached; therefore, the total length of the single-stranded linkers is $N_{ss} = N_{ss}^0 + n$, where n is as usual the
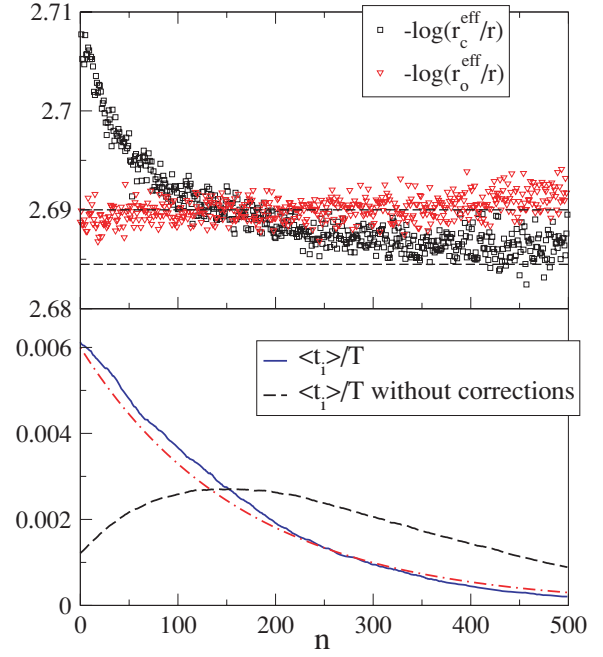
Figure 4. Bottom: average fraction of the time spent on each base. The full (blue) curve corresponds to equation (34) while the dashed (black) curve corresponds to equation (34) without the saddle-point corrections (the square-root term). The dot-dashed (red) line is $P_{eq}(n) \propto \exp[-n\Delta g]$ with $\Delta g = 0.006$. Top: effective rates (squares and triangles) estimated from the maximization of the probability in equation (36) ($r = 10^6$ Hz) without saddle-point corrections (full curve of the lower panel). The dashed lines are the asymptotic values of the rates; see text. We do not report the rates corresponding to the full equation (34) since they are essentially independent of n.

number of open base pairs. The leftmost linker is a double-stranded DNA of $N_{ds} = 3120$ base pairs, whose presence is however irrelevant for the scope of this section. The total length of the simulation was $T = 7200$ s, i.e. 2 h.

4.1.1. A test of the model. The average fraction of time spent on each base, corresponding to the equilibrium probability distribution $P_{eq}(n)$, is reported in the lower panel of figure 4. We expect that in the large n limit, $P_{eq}(n) \sim \exp[-n(g_0 - 2g_{ss}(f))] = \exp[-n\Delta g]$, with $\Delta g \sim 0.006$. This is expected to break down when N_{ss} is so small that the second-order corrections to the saddle-point in equation (34) become important. As can be seen in figure 4, the exponential form correctly describes the data.

We performed additional simulations in which the square-root term in equation (34) was removed. As one can see, in this case the small n deviations are much more pronounced. It is worth noting that for a non-Gaussian polymer, one expects a deviation from the exponential form at small enough n. However, this analysis shows that taking into account the small n corrections to $W(x, n)$ systematically reduces this effect. Estimating its real order of magnitude therefore requires an exact expression for $W(x, n)$, which could be in principle obtained from the recurrence equation (30). However, this is a complicated numerical task that goes beyond the scope of this

paper. What we want to stress here is that the inclusion of the square-root term in equation (34) gives significant differences when $n \lesssim 200$ and should therefore be included if one wants to analyze the unzipping of small molecules.

4.1.2. Effective dynamics of the fork.

In a situation where the linkers are short, such that their relaxation time is faster than the mean time spent on a base, the linkers are able to reach equilibrium before n changes. Therefore one might hope to define an *effective dynamics* for the fork, where n changes according to effective rates that depend on the variation in the free energy of the setup on closing or opening a base.

To this aim we considered the model for the fork dynamics described in section 2.3, but assuming n-dependent opening and closing rates. Within this model, the probability of a trajectory of the fork is a function of the number of upward (u_n)/downward (d_n) jumps and the time spent on base n, t_n:

$$P_{\text{eff}}[n(t)] = \prod_{n=1}^{N} \left(r_c^{\text{eff}}(n)\Delta t \right)^{d_n} \left(r_o^{\text{eff}}(n)\Delta t \right)^{u_n}$$
$$\times \left(1 - \Delta t \left(r_c^{\text{eff}}(n) + r_o^{\text{eff}}(n) \right) \right)^{t_n}. \quad (36)$$

Given the values of u_n, d_n, t_n measured along our trajectory of duration T, we can infer the effective rates by maximizing the above probability. Assuming that $r^{\text{eff}}\Delta t \ll 1$, we obtain

$$r_c^{\text{eff}}(n) = \frac{d_n}{t_n}, \qquad r_o^{\text{eff}}(n) = \frac{u_n}{t_n}, \quad (37)$$

as estimates for the effective rates. For the full expression (34), the rates are almost independent of n; on the other hand, if the first-order correction is neglected, one obtains n-dependent rates, consistent with the observation that $P_{\text{eq}}(n)$ is not exponential. These are reported in the upper panel of figure 4. In both cases, the rates are consistent with the detailed balance condition $r_c^{\text{eff}}(n)P_{\text{eq}}(n) = r_o^{\text{eff}}(n-1)P_{\text{eq}}(n-1)$.

4.2. Out-of-equilibrium opening

For long molecules, the barrier between the closed and open states may become very large, e.g. $\sim 3000\, k_{\text{B}}T$ for the 50 000 bases λ–DNA at the critical force $f_c = 15.5$ pN [31]. The time necessary to cross this barrier is huge, and full opening of the molecule never happens during experiments. To open a finite fraction of the molecule, the force has to be chosen to be larger than its critical value. The opening can then be modeled as a transient random walk, characterized by pauses at local minima of the free energy and rapid jumps in between [16].

4.2.1. Analytical calculation of the average time spent by the fork on a base.

First consider the case of a fixed force acting on the fork while all the other components are at equilibrium as in section 2.3. In the transient random walk, the opening fork spends a finite time around a position n before escaping away and never coming back again in n. The number u_n of opening transitions $n \rightarrow n + 1$ is stochastic and varies from experiment to experiment and base to base. The total number of times the fork visits the base pair n before escaping is given by the sum of the number u_n of transitions from $n - 1$ to n

and of the number $u_{n+1} - 1$ of transitions from $n + 1$ to n. Therefore, the average time spent in n is

$$t_n = \frac{\langle u_n \rangle + \langle u_{n+1} \rangle - 1}{r_o + r_c(n)}, \quad (38)$$

where $1/(r_o + r_c(n))$ is the average time spent in n before each opening or closing step. Let us introduce the probability E_{n+1}^n of never reaching back position n starting from position $n + 1$. The probability P of the number u_n of opening transitions $n \rightarrow n + 1$ during a single unzipping simply reads as

$$P(u_n) = \left(1 - E_{n+1}^n \right)^{u_n - 1} E_{n+1}^n. \quad (39)$$

From equation (39), we have that the average number of openings of bp n is

$$\langle u_n \rangle = \sum_{u_n \geqslant 1} P(u_n) u_n = \frac{1}{E_{n+1}^n}. \quad (40)$$

We are thus left with the calculation of E_{n+1}^n. For infinite force, $E_{n+1}^n = 1$ since the fork never moves backward. For finite force, we write a recursive equation for the probability E_m^n that the fork never comes back to base n starting from base $m (\geqslant n + 1)$:

$$E_m^n = q_m E_{m-1}^n + (1 - q_m) E_{m+1}^n, \quad (41)$$

where

$$q_n = \frac{e^{g_{\text{ss}}(f)}}{e^{g_{\text{ss}}(f)} + e^{g_0(b_n, b_{n+1})}} \quad (42)$$

is the probability of closing base n and $1 - q_n$ is the probability of opening it at each step. Note that for forces larger than the critical force, we have $q_n < \frac{1}{2}$: the random walk is submitted to a forward drift and is transient. The boundary conditions for equation (41) are $E_n^n = 0$ and $E_m^n = 1$ for $m \rightarrow \infty$.

For a homogeneous sequence, the escape probability is $E = (1 - 2q)/(1 - q)$. For a heterogeneous sequence by defining $\rho_m^n = \frac{E_m^n}{E_{m+1}^n}$, we obtain the Riccati recursion relation:

$$\rho_n^n = 0; \qquad \rho_{m+1}^n = \frac{1 - q_{m+1}}{1 - q_{m+1}\rho_m^n} \qquad \text{for } n \geqslant m. \quad (43)$$

Equation (43) can be solved numerically for a given sequence. Then, the escape probability starting from $n + 1$ is

$$E_{n+1}^n = \prod_{m \geqslant n+1} \rho_m^n, \quad (44)$$

and the average time spent in the base n is then obtained from (40) and (38).

4.2.2. Results from the dynamical model.

To check whether these theoretical predictions are affected by dynamical fluctuations of the bead, linkers and unzipped strands, we have carried out simulations with the model of section 3. We have carried out 160 unzippings of the λ-phage sequence at a force of 17 pN for $T = 100$ s (physical time), with the same molecular construct of section 4.1 ($N_{\text{ds}} = 3120$ base pairs of dsDNA linkers on a side plus $N_{\text{ss}}^0 = 40$ bases of the ssDNA linker at each side of the DNA to be open). For such a construct, the equilibrium extension of the polymers for n open base pairs is $2N_{\text{ss}}l_{\text{ss}} + N_{\text{ds}}l_{\text{ds}}$, where $l_{\text{ds}} = 0.3337$ nm, $l_{\text{ss}} = 0.4758$ nm and $N_{\text{ss}} = N_{\text{ss}}^0 + n$. The stiffness of the polymers

number of open bases
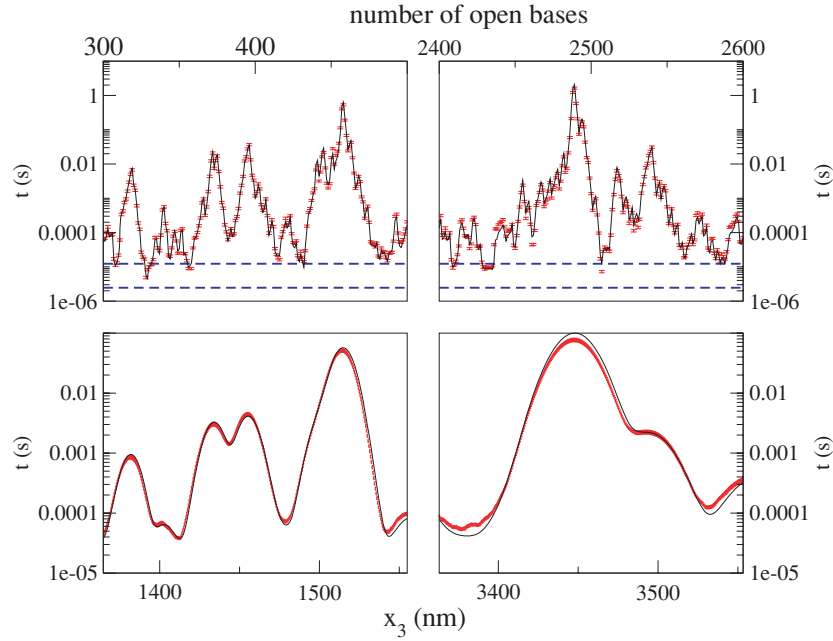
Figure 5. Top: average time spent by the fork on position n. Bottom: time spent by the whole setup at an extension between x_3 and $x_3 + \Delta x$, with $\Delta x = 0.5$ nm. The black line in both figures represents the theoretical predictions from section 4.2.1. The red points are the results from the simulation. Standard deviations are represented by error bars in the top panels and by the thickness of the red curves in the bottom panels.

is $1/k_{\mathrm{eff}} = N_{\mathrm{ss}}/k_{\mathrm{ss}}^m + N_{\mathrm{ds}}/k_{\mathrm{ds}}^m$ with $k_{\mathrm{ss}}^m = 160.5$ pN nm^{-1} and $k_{\mathrm{ds}}^m = 1450$ pN nm^{-1}. The relaxation times of the polymers are of the order of 0.1 ms for about 400 unzipped bases and 1 ms for about 2500 open bases, and are larger than the characteristic times of about 2×10^{-6} s needed to open a weak base and of about 10^{-5} s needed to open a strong base.

We plot in figure 5 the average time spent by the fork at location n for two portions of the sequence, corresponding to about 400 and 2500 open base pairs. The agreement between the theoretical and numerical estimates of the times is excellent, meaning that the fluctuations of extensions of the polymers and the dynamics of the bead induce negligible changes on the rates of opening and closing, as seen close to the critical force in section 4.1.

As experiments do not give direct acces to the time spent by the fork at location n, we show in figure 5 (bottom) the time $t(x_3)$ spent by the unzipped ssDNA between extensions x_3 and $x_3 + dx$. These times are compared to their values assuming that the positions x_3 of the beads are randomly drawn from the equilibrium measure:

$$t(x_3) = \sum_n t_n P(x_3|n), \qquad (45)$$

where t_n is calculated from (38) and $P(x_3|n)$ is calculated from an argument similar to that used in section 2.2.1 and can be written up to the quadratic order around the saddle point as

$$P(x_3|n) = \sqrt{\frac{\beta k_{\mathrm{eff}}(f)}{2\pi}} \, \mathrm{e}^{-\beta \frac{k_{\mathrm{eff}}(f)}{2}(x_3 - N_{\mathrm{ds}} l_{\mathrm{ds}}(f) - 2N_{\mathrm{ss}} l_{\mathrm{ss}}(f))^2}. \qquad (46)$$

The agreement is, again, excellent.

Figure 5 and equation (45) show that $t(x_3)$ gets contributions from the times spent by the fork on a set of bases whose number depends on the magnitude of the equilibrium

fluctuations of the linkers. These equilibrium fluctuations increase with the length of ssDNA, e.g. $\delta x_3 \simeq 5$ nm for 400 unzipped base pairs and $\delta x_3 \simeq 12$ nm for 2500 unzipped bases. Therefore, as the number n of unzipped base pairs increases, the characteristic curve of $t(x_3)$ gets more and more convoluted (compare left-bottom and right-bottom panels in 5).

In figure 6 we compare the value of the ssDNA extension from one unzipping, x_3, to its average value at equilibrium, x_3^{eq}, as a function of the number of unzipped base pairs n. The fluctuations in the extension are compatible with the equilibrium deviations. Again, no clear out-of-equilibrium effect is observed. The reason is that, even if the single strand is not relaxed in the opening time of a base, the fork goes back and forward around a given location before moving away. Therefore, the quantities we have measured are averaged on the number of times a base pair is opened and are close to their mean value even in a single unzipping. This can be deduced from figure 5 by comparing the total time spent on a base (points) with the time to open a base (dashed lines)

5. Unzipping at fixed extremities

5.1. Correlation functions

One of the main advantages of considering the dynamics of the linkers and of the beads is that it allows us to compute autocorrelation functions and to explore the interaction between different parts of the setup, a task which would be impossible from *a priori* calculations.

We have performed a few simulations with the setup shown in figure 1(A) where the spring constant of the first optical trap of extension x_1 is 0.1 pN nm^{-1} and the second

11
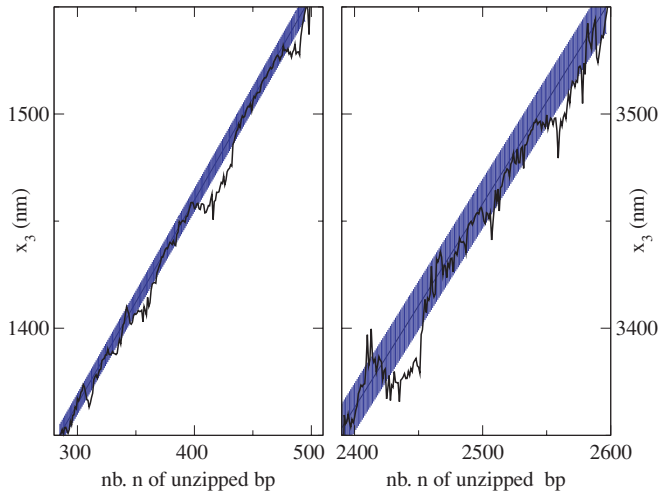
Figure 6. Total extension x_3 of the setup in figure 1(*B*) at a fixed number *n* of unzipped bases for a single unzipping (black line). If the fork visits the same base *n* twice or more, we plot the average of the extension values. The gray strip represents the average value at equilibrium, $x_3^{eq}(n)$, and the standard deviation around its value at equilibrium.

Figure 7. Correlation functions for the setup in figure 1(*A*) at two different values of the number of open bases, $N_{ss} = 40 + n$.
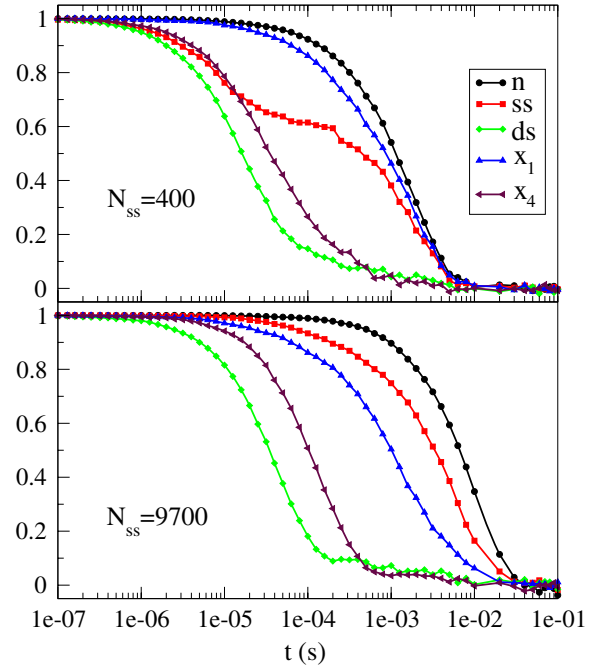
Table 5. Comparison between the correlation times of the setup in figure 1(*A*) as computed for an isolated element and the result of a complete numerical simulation. In the case of the fork, we reported as the theoretical value $1/k_{eff}$, which must be multiplied by a viscosity to obtain the relaxation time; it turns out that a viscosity $\sim 8 \times 10^{-5}$ pN s nm^{-1} matches the theoretical and numerical results.

	Theoretical (s)	Numerical (s)
Single strand	$4.83 \times 10^{-11} N_{ss}^2$	$5.4 \times 10^{-11} N_{ss}^2$
Double strand	4.96×10^{-5}	$\sim 3 \times 10^{-5}$
Spring x_1	1.67×10^{-4}	$\sim 1.5 \times 10^{-3}$
Spring x_4	3.26×10^{-4}	$\sim 7 \times 10^{-5}$
Fork N_{ss}	$\propto 14.2 + 0.013 N_{ss}$	$1.3 \times 10^{-3} + 8.4 \times 10^{-7} N_{ss}$

(x_4) has stiffness 0.512 pN nm^{-1}. The molecule in the fork is uniform with $g_0 = 2.69 k_B T$. The only parameter that is varied across simulations is the distance between the optical traps and thus the typical number of open bases. In figure 7, we show two typical cases. What is evident is that the single strand has two time scales: one which is proper to the fluctuations at *n* fixed and another which is of the same order of magnitude as the correlation time of the fork. As the number of open bases grows, the fast time scale also grows until it becomes impossible to distinguish the two.

As remarked in section 3.5, our model cannot in principle be used when the linkers are made of $n \gtrsim 1000$ monomers. To check for the importance of force propagation effects, we ran a simulation for $N_{ss} = 9700$ (bottom panel of figure 7) where we cut each linker into nine subunits of 1000 bases each plus a final unit which is connected to the opening fork. Overall, the correlation functions are not much affected by this modification and in particular the correlation times are unaffected within numerical errors. The main effect of cutting the long linkers is that the correlation function of the linker becomes more stretched (i.e. if they are fit with $\exp[-(t/\tau)^{\beta_s}]$, the exponent β_s is slightly smaller). This is to be expected since by cutting the polymer we include more relaxation modes, each with its relaxation time. A wider distribution of relaxation times implies a smaller exponent β_s. In table 5, we compare the results of the numerical simulation with the predictions of section 2.2.1 which do not take into account the interactions between different parts of the setup. While the simulated results for the single-stranded and the double-stranded DNA are not too far off from the prediction, the two springs show a much greater deviation from the theoretical estimates. This prompted us to analyze further the relationship between the fork and the bead position as will be discussed later.

The potential acting on the fork position, in the case of a uniform molecule, is dictated by the stiffness of the rest of the setup only as seen in section 2.2.1. That is to say that *n* experiences a harmonic potential with the spring constant proportional to k_{eff}; this in turn predicts correlation times that are proportional to $\frac{1}{k_{eff}}$ which has a linear dependence on *n*. This behavior is in very good agreement with the data that have been extracted from numerical simulations.

5.2. Mutual information between the bead position and fork location

Figure 9 shows the dynamical correlations of the fork and bead positions. The two beads have different correlation functions due to the difference in their stiffnesses: $k = 0.5$ pN nm^{-1} for bead 1 and $k = 0.1$ pN nm^{-1} for bead 2. After an initial decay (taking place over a time proportional to $1/k$ from section 2.3.3), the bead correlations exhibit a quasi-plateau behavior whose height is roughly proportional to $1/k$. The plateau reflects the correlation between the motion of the bead and that of the fork on time scales of the order of the equilibration
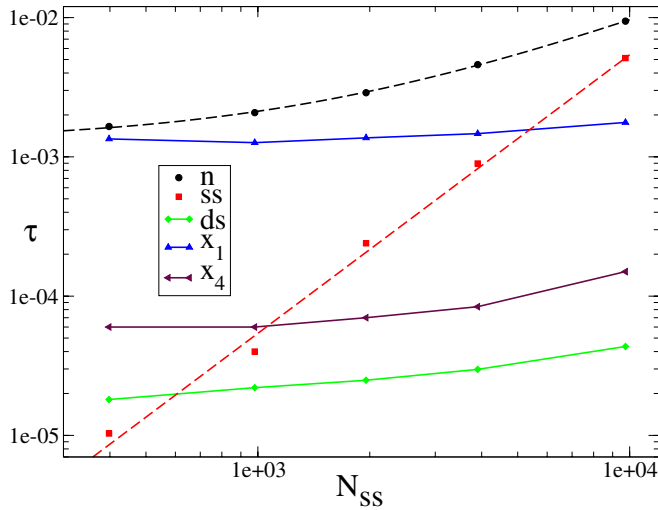
Figure 8. Relaxation times of the correlation functions in figure 7 as a function of the number of open bases. In the case of the single strand (ss), only the fast relaxation time is plotted. For the fork and the single strand, dashed lines indicate a fit to $\tau_n = A + BN_{ss}$ (with $A = 1.3 \times 10^{-3}$ and $B = 8.4 \times 10^{-7}$) and $\tau_{ss} = CN_{ss}^2$ (with $C = 5.4 \times 10^{-11}$ s). For the others, full lines are guides to the eye.

time of the fork. It appears that soft beads allow one to track the location of the fork better than stiffer beads.

In the following, we will give a closer look at the dependence of these correlations on the optical trap stiffness; to do so we construct a setup as in figure 1(*A*), but where the stiffness of the optical trap on the left is kept constant at 0.512 pN nm^{-1} while the stiffness of that on the right is varied across two orders of magnitude[3].

To give quantitative support to this statement we define the mutual information *I* between the position of the bead in the optical trap, x_4, and the number of open base pairs, n:

$$I(x_4, n) = \sum_n \int \mathrm{d}x_4 \, P(x_4, n) \log\left(\frac{P(x_4, n)}{P(x_4)P(n)}\right), \quad (47)$$

where $P(x_4, n)$ is the joint probability density for the bead to be at position x_4 while there are n open base pairs; $P(n)$ and $P(x_4)$ are the two marginals. Note that the definition of mutual information does not suffer from the problems which arise with entropy when we switch between a continuous and a discrete definition; that is to say that binning with sufficiently small bins does not change the mutual information.

I can be easily computed by keeping track of the times passed at a given bead position and the given number of open bases during a run of the simulation. As stressed before, the fact that the x_4 coordinate must be binned has negligible effects on the computation of entropy. For very large stiffnesses the amplitude of the oscillations of the bead can become very small, and thus a lack of sensitivity in the measure of the position of the bead could become an issue. Fortunately, the current state of the art in the optical trap cannot attain stiffnesses larger than, say, 1 pN nm^{-1} with

---

[3] An attentive reader might have noted that we changed the stiffness of the right bead compared to what it was in the previous section; the rationale behind this choice is to keep its value at the center of the range in which we will vary the other.
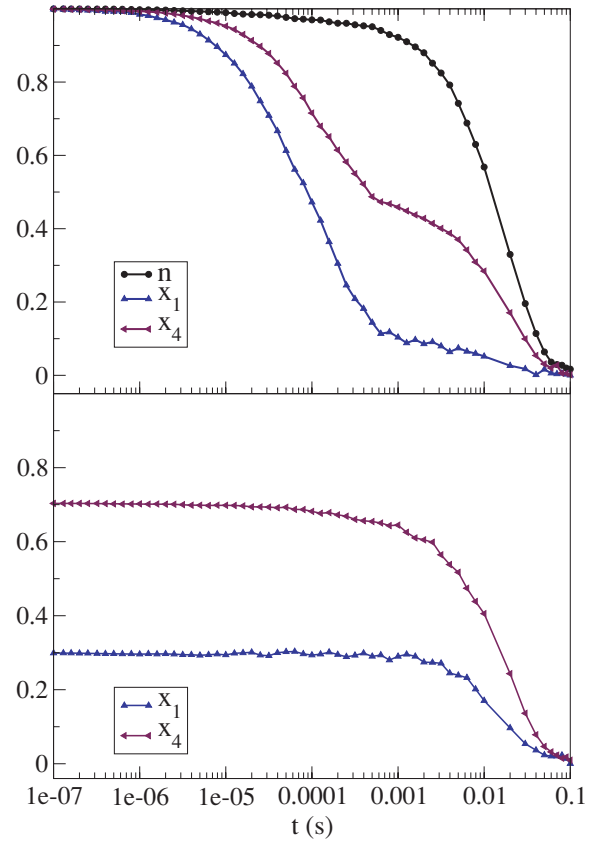
Figure 9. Top: autocorrelation functions for the setup in figure 1(*A*) when the molecule to unzip is a block copolymer composed of alternating stretches of ten strong pairs and ten weak pairs. This way the fork correlation time is greatly increased allowing us to view effects on the two traps of different optical stiffnesses. Bottom: correlation functions between one of the two beads and the number of open base pairs. Values have been normalized so that the value at zero time difference is $\rho = \langle x_i n \rangle / \sqrt{\langle x_i^2 \rangle \langle n^2 \rangle}$.

micrometer beads [32]. In this regime, the fluctuations of the bead are dominated by the stiffness of the trap and thus we can say that $\langle \delta x_4^2 \rangle \sim (\beta k_2)^{-1}$; see equation (19). Comparing the fluctuations of the bead position with the sub-nanometer precision Δ over its location yields

$$\frac{\sqrt{\langle \delta x_4^2 \rangle}}{\Delta} \simeq 10\text{--}50, \quad (48)$$

which is much larger than unity.

Figure 10 shows that the mutual information *I* only weakly depends on the sequence but strongly depends on the stiffness *k* of the trap. This behavior can be understood very intuitively. Right after a base pair opens or closes, the whole setup in a fixed-force experiment has to give way; the less rigid an element of the setup is compared to the rest, the more it will accommodate for the change in *n*.

We conclude that, in a single measurement, soft traps give more information on the fork location than stiff traps. However, *I* is the mutual information between the fork and bead locations *per measure*. As we have seen in section 5.1, the correlation times extracted from the simulations decrease with *k* and, as *k* grows, more and more uncorrelated measures
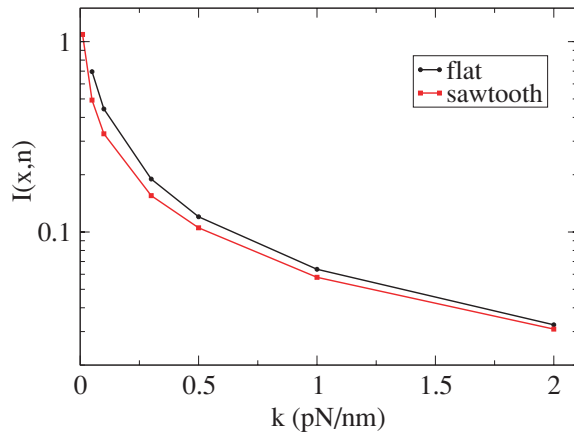
Figure 10. Mutual information *I* between x_4 and *n* as a function of the trap stiffness, *k*. Black circles are computed on a uniform sequence, while red squares are measured on the sawtooth potential described in the caption to figure 9.

can be done in the same amount of time. It is thus expected that information *per unit of time* is not maximal for small values of *k*. In other words, stiffer traps give worse quality but more frequent signals on the location of the fork. Finding the optimal value of *k* would require a detailed analysis of the correlation times of the bead and of the fork. In particular, the size of the bead would affect the optimal value for *k* through the viscosity coefficient, but not the information per measure, *I*. However this dependence should not be crucial since the bead size cannot be much varied in experiments: it can be neither too small to exert a sufficient force nor too large due to the size of the physical setup.

6. Conclusion

This paper has been devoted to the presentation of a dynamical model for the different components of the setups used in the unzipping of single DNA molecules under a mechanical action. Compared to previous studies, our model does not assume *a priori* that the polymers in the molecular construction are at equilibrium but takes into account their relaxation dynamics. It is important to stress out that the dynamical description for the linkers and the unzipped part of DNA is coarse grained: the basic unity is the polymers themselves and not the monomers they are made of.

As a consequence, each polymer is associated with a unique relaxation time. The assumption is justified as long as these times are comparable to the typical opening or closing time of a single base pair. Longer polymeric chains, e.g. ssDNA strands with a few thousand bases, need to be modeled in a more detailed way; more precisely, they should be divided into short enough segments along which the force can be considered as uniform on the time scales associated with the fork motion. Although in this paper we did not observe any important force propagation effect, these might be more important in strongly nonequilibrium situations such as opening at constant (high) velocity. We plan to simulate unzippings with such molecular constructions in the near future to understand how force propagation across the

polymeric segments can affect the effective rates for closing base pairs in such situations.

One of our results is that one has to be very careful with the expression of the free energies (entering the dynamical rates) for short polymers, be they linkers or ssDNA unzipped strands. Use of the free energy per monomer, obtained from force–extension measures on long molecules, as usually done in the literature, can lead to erroneous results. We have shown that finite-size corrections to the energetic contributions and the dynamical rates have to be taken into account.

As a main advantage, the code we have developed is versatile: we can easily change setups, for example use a fixed-force or fixed-position ensemble, and change the number and types of linkers and of traps for the beads. We have found that, in fixed-force unzippings, the opening and closing rates for the fork are not affected by the force fluctuations coming from the polymeric chains. For small linkers and a number of unzipped base pairs, indeed, force fluctuations are large but fast, and are averaged out on the characteristic opening–closing time of a base pair. For large linkers or a number of unzipped bases force fluctuations are slow but small, and therefore do not change the dynamic of the opening fork. We have also performed unzipping simulations at large forces where the opening dynamics is transient, and found that the average time spent by the unzipped strands at a given extension is accurately predicted from the time spent by the fork on a base convoluted by the equilibrium fluctuations of ssDNA. Moreover, the extension between the extremities at a fixed number of open base pairs in a single unzipping experiment is compatible with equilibrium fluctuations of ssDNA and linkers. The program could be easily adapted to unzipping at constant velocity, where non-equilibrium effects are likely to be more important.

Our study suggests that one measure of the position of the bead in soft traps gives more information on the location of the fork than in the case of stiffer traps. This statement is however to be considered with caution. Beads in stiffer traps reach equilibrium on shorter time scales, and the overall rate of information per unit time could be higher in stiffer traps. While purely qualitative at this stage, such a statement is relevant to the study of the inverse problem of unzipping, that is, inferring the sequence of the DNA molecule from the unzipping signal. We hope that the present dynamical modeling will be useful to assess the rate at which information on the sequence could be acquired from mechanical single molecule experiments.

Appendix A. Langevin dynamics of coupled polymers

One of the simplest models of polymer dynamics is that proposed by Rouse [44], where the polymer is described as

a chain of beads which are modeled as Brownian particles, linked by harmonic springs.

While it is true that this model is very crude because it ignores hydrodynamic interactions and exclude volume effects, it has the huge advantage of being largely solvable. Therefore, we will now use it as the basis for a few considerations that will then be generalized to more realistic models.

Our aim is to write a system of coupled equations for the time evolution of a certain number of marked points on a (hetero)polymer. One of these points will be for instance the location of the opening fork. In the case of a double DNA strand attached to a single strand, one point will mark the location where the two different polymers are attached (see the examples in figure 1). Note that if the marked points we focus on are far apart, only the slower modes of the system will be relevant, as the fast modes describe local relaxations of the chain. Therefore, in the following, we want to focus on a long wavelength/long time effective description of the chain.

A.1. The dynamics of a single polymer

A.1.1. The model and its normal modes. As the simplest case we consider a polymer composed of N identical springs, each with an identical link at one end. The first is connected to a wall that has infinite mass (or, better still in this framework, infinite viscosity) and on the last a force f is exerted. The Langevin equations describing such a polymer can be written as

$$\begin{cases} \gamma_m \dot{u}_1 = -2k_m u_1 + k_m u_2 + \eta_1 \\ \quad\vdots \\ \gamma_m \dot{u}_n = -2k_m u_n + k_m u_{n-1} + k_m u_{n+1} + \eta_n \\ \quad\vdots \\ \gamma_m \dot{u}_N = -k_m u_N + k_m u_{N-1} + f + \eta_N, \end{cases} \quad (A.1)$$

where η_i are white Gaussian noises of zero mean and variance:

$$\langle \eta_i(t)\eta_j(0)\rangle = 2k_B T \delta_{ij}\delta(t). \quad (A.2)$$

Let us for the moment neglect the noise term. Then, defining $\tau_m = \gamma_m/k_m$, we can formally rewrite these equations as

$$\tau_m \dot{u}_n = -2u_n + u_{n-1} + u_{n+1}, \qquad \forall\, n, \quad (A.3)$$

supplemented by the boundary conditions

$$u_0 \equiv 0, \qquad u_{N+1} \equiv u_N + f/k_m. \quad (A.4)$$

A standard way to find the normal modes of the above linear system is to search for solutions of the form $u_n(t) = u_n(0)\exp(-\lambda t/\tau_m)$. One can easily show that the general solution satisfying the first boundary condition $u_0 = 0$ has the form

$$u_n(t) \propto \sin(qn)\exp(-\lambda(q)t/\tau_m),$$
$$\lambda(q) = 2(1 - \cos(q)). \quad (A.5)$$

The second boundary condition (A.4) requires that $u_{N+1}(t) - u_N(t) = f/k_m = $ const. Since we can always add the constant value to $u_{N+1}(t)$, we can replace this boundary condition by $u_{N+1}(t) = u_N(t)$. This requires that $\sin(qN) \sim \sin(q(N+1))$; then $q = (\pi/2 + p\pi)/N$. The slowest mode then corresponds to $q = \pi/2/N$, which for large N gives a relaxation time

$$\tau(N) = \tau_m/\lambda(\pi/2/N) \sim \frac{4}{\pi^2}\tau_m N^2, \quad (A.6)$$

which proves the validity of the scaling in equation (20).

A.1.2. Recurrence equations for a fixed end. We now want to write a system of coupled equations for a certain number of points on the polymer by integrating out us we are not interested in. To begin, we focus on the end point u_N.

It is convenient to perform a Laplace transformation and write

$$u_n(t) = \int_0^\infty \mathrm{d}\lambda\, u_n(\lambda)\, \mathrm{e}^{-\lambda t/\tau_m}. \quad (A.7)$$

Then equation (A.5) becomes, in Laplace space,

$$(2 - \lambda)u_n(\lambda) = u_{n+1}(\lambda) + u_{n-1}(\lambda), \quad (A.8)$$

with the same boundary conditions $u_0(\lambda) \equiv 0$, and $u_{N+1}(\lambda) - u_N(\lambda) = (f/k_m)\delta(\lambda)$. For $\lambda \neq 0$, the latter condition reduces to $u_{N+1}(\lambda) = u_N(\lambda)$ as discussed above for the normal mode analysis.

We introduce a function

$$\zeta_{n-1}(\lambda) = u_{n-1}(\lambda)/u_n(\lambda). \quad (A.9)$$

Substituting the latter relation in (A.8), we get

$$(2 - \lambda - \zeta_{n-1}(\lambda))u_n(\lambda) = u_{n+1}(\lambda), \quad (A.10)$$

from which we get a Riccati recurrence equation

$$\begin{cases} \zeta_0(\lambda) = 0 & \text{(due to } u_0 = 0), \\ \zeta_n(\lambda) = \dfrac{1}{2 - \lambda - \zeta_{n-1}(\lambda)}. \end{cases} \quad (A.11)$$

This recurrence can be solved and the function $\zeta_n(\lambda)$ computed for all n.

Since we are interested in the large time limit, we can expand the function $\zeta_n(\lambda)$ for small λ; we obtain

$$\zeta_n(\lambda) = \frac{n}{n+1} + \frac{n(1+2n)}{6(1+n)}\lambda + \frac{n(6+19n+16n^2+4n^3)}{180(1+n)}\lambda^2$$
$$+ O(\lambda^3). \quad (A.12)$$

One obtains the effective equation for u_N by substituting the above expression in (A.10) and setting $n = N$. Keeping only the linear term in λ and the leading terms in $N \gg 1$, we get

$$\left(1 + \frac{1}{N} - \lambda\frac{N}{3}\right)u_N(\lambda) = u_{N+1}(\lambda). \quad (A.13)$$

Moving back to the time domain, we obtain

$$\tau_m\frac{N}{3}\dot{u}_N = -\frac{1}{N}u_N + (u_{N+1} - u_N), \quad (A.14)$$

which is equivalent, using the boundary condition $u_{N+1} - u_N = f/k_m$, to

$$\frac{\gamma_m N}{3}\dot{u}_N = -\frac{k_m}{N}u_N + f. \quad (A.15)$$

In this way, we got an effective equation for the endpoint of the polymer that is still a linear first-order differential equation and takes into account only the slowest mode of the chain.

There is however an inconvenience: in fact a straightforward computation shows that the relaxation time obtained from equation (A.15) is $\tau(N) = \tau_m N^2/3$ that differs by a factor $\pi^2/12$ from the correct value given by equation (A.6). The origin of this discrepancy clearly lies in the fact that the expansion we made in equation (A.12) is not convergent at fixed λ for $n \to \infty$, as successive terms in the series are of order $n^{2p-1}\lambda^p$.
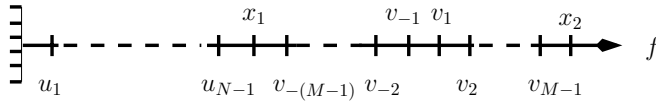
Figure A1. Two joint polymers subjected to an external force f. x_1 marks the endpoint of the first polymer made of N links whose endpoints are $u_1, u_2, \ldots, u_{N-1}, u_N \equiv x_1$. The second polymer originates from x_1 and is made of $2M - 1$ links, whose endpoints are $v_{-(M-1)}, v_{-(M-2)}, \ldots, v_{-1}, v_1, \ldots, v_{M-2}, v_{M-1}, x_2$.

Let us then go back to the computation of the normal modes of the system within this formalism. The second boundary condition $u_{N+1}(\lambda) = u_N(\lambda)$ implies $\zeta_N(\lambda) = 1$. The normal modes are the solutions of this equation with respect to λ. One can show from the exact expression of $\zeta_N(\lambda)$ that

$$\lim_{N \to \infty} N[\zeta_N(\widetilde{q}^2/N^2) - 1] = -\widetilde{q} \cot(\widetilde{q}) \equiv \widetilde{\zeta}(\widetilde{q}). \quad \text{(A.16)}$$

The zeroes of this function are $\widetilde{q} = \pi/2 + k\pi$; therefore, the solutions of $\zeta_N(\lambda) = 1$ tend for large N to $\lambda = (\pi/2 + p\pi)^2/N^2$, in agreement with the exact result of the previous section. An inspection of equations (A.12) and (A.16) shows that the small λ expansion of $\zeta_N(\lambda)$ is equivalent to performing a small $\widetilde{q}$ expansion of $\widetilde{\zeta}(\widetilde{q})$ in order to find its first zero. This indeed yields $\widetilde{\zeta}(\widetilde{q}) \sim -1 + \widetilde{q}^2/3$ that gives $\widetilde{q} = \sqrt{3}$ for the first zero that gives back $\tau(N) = \tau_m N^2/3$.

Then one can check that a higher order expansion in λ (or equivalently in $\widetilde{q}$) produces a more accurate result; indeed the series of $\widetilde{\zeta}(\widetilde{q})$ converges for $\widetilde{q} < \pi$ while the zero is located at $\widetilde{q} = \pi/2$. It is easy to show that if one truncates the series to order p, the difference between the solution and the true zero is exponentially small in p.

A.1.3. Discussion The conclusion of this section is that equation (A.15) is a correct description of the dynamics of the end of the polymer in the limit of large N and large times. While it captures the correct scaling with N of the relaxation time, the coefficient is wrong by a factor of $\pi^2/12 \sim 0.82$. Still, this is quite satisfactory for our purposes since the experimental error in the determination of τ_m is of the same order of magnitude. Better approximations can be obtained by truncating the expansion of $\zeta_N(\lambda)$ to higher orders in λ, therefore obtaining a higher order differential equation for $u_N(t)$.

In the following, we will derive the coupled equation for many marked points along the chain, limiting ourselves to the first-order truncation. This produces first-order differential equations of the Langevin type.

A.2. Dynamics of two coupled polymers

We will now show how to use this formalism to derive coupled equations for different points on a composite polymer. We continue neglecting the noise, which we will reintroduce at the end of this section.

As a simple example, let us consider the polymer drawn in figure A1. It is composed of N monomers of type 'U' linked to $2M - 1$ monomers of type 'V'. The two types of monomers might differ in the value of the microscopic spring constant,

bead viscosity, etc. If the monomers are identical, then we are just marking a point in the middle of a polymer.

The effective equation for the endpoint of polymer U can be derived following the analysis of the previous section. We denote $x_1 \equiv u_N$ and we get

$$\gamma_m^U \frac{N}{3} \dot{x}_1(t) = -\frac{k_m^U}{N} x_1(t) + k_m^V(v_{-(M-1)}(t) - x_1(t)), \quad \text{(A.17)}$$

where the last term is the 'external' force that the polymer V exerts on U.

A.2.1. Integration of the V polymer. Now we want to integrate out all the monomers $v_{-(M-1)}, \ldots, v_{M-1}$ in order to obtain the coupling between x_1 and x_2. To this aim, and in order to keep the formalism symmetric, we can start from the middle of the polymer V by integrating simultaneously v_{-1} and v_1 in order to obtain effective equations for v_{-2} and v_2, and so on. In Laplace space (note that now in equation (A.7) $\tau_m = \tau_m^V$), the equations for $v_{\pm 1}$ have the form

$$\begin{aligned} (2 - \lambda)v_{-1}(\lambda) &= v_{-2}(\lambda) + v_1(\lambda), \\ (2 - \lambda)v_1(\lambda) &= v_2(\lambda) + v_{-1}(\lambda). \end{aligned} \quad \text{(A.18)}$$

These can be easily solved to get $v_{\pm 1}$ as a function of $v_{\pm 2}$. Iteration leads to the following form for the equation after n steps:

$$\begin{aligned} \xi_n(\lambda)v_{-n-1}(\lambda) &= v_{-n-2}(\lambda) + \eta_n(\lambda)v_{n+1}(\lambda), \\ \xi_n(\lambda)v_{n+1}(\lambda) &= v_{n+2}(\lambda) + \eta_n(\lambda)v_{-n-1}(\lambda). \end{aligned} \quad \text{(A.19)}$$

One can check that this form is stable under one step of iteration and the following recursion relations are obtained:

$$\begin{cases} \xi_0 = 2 - \lambda, \\ \eta_0 = 1, \\ \xi_{n+1} = 2 - \lambda - \dfrac{\xi_n}{\xi_n^2 - \eta_n^2}, \\ \eta_{n+1} = \dfrac{\eta_n}{\xi_n^2 - \eta_n^2}, \end{cases} \quad \text{(A.20)}$$

where the initial values are determined by consistency between (A.18) and (A.19) for $n = 0$. These recurrences are easily solved by introducing the two quantities $A_n = 1/(\xi_n - \eta_n)$ and $B_n = 1/(\xi_n + \eta_n)$ respectively; these satisfy the same recurrence in (A.11) except for the initial condition which is different and determined according to (A.20).

At the leading order in $n \to \infty$ and at first order in λ, we get

$$\xi_n(\lambda) = 1 + \frac{1}{2n} - \frac{2n}{3}\lambda, \qquad \eta_n(\lambda) = \frac{1}{2n} + \frac{2n}{6}\lambda. \quad \text{(A.21)}$$

Finally, one obtains from this procedure a coupled equation for $v_{-(M-1)}$ and v_{M-1} where $x_1 \equiv v_{-M}$ and $x_2 \equiv v_M$ also appear.

A.2.2. Coupled effective equations. To obtain the coupled effective equations, one starts from the following system:

$$\begin{cases} -\gamma_m^U \dfrac{N}{3} \dfrac{\lambda}{\tau_m^V} x_1 = -\dfrac{k_m^U}{N} x_1 + k_m^V(v_{-M+1} - x_1), \\ \xi_{M-2}(\lambda)v_{-M+1}(\lambda) = x_1 + \eta_{M-2}(\lambda)v_{M-1}(\lambda), \\ \xi_{M-2}(\lambda)v_{M-1}(\lambda) = x_2 + \eta_{M-2}(\lambda)v_{-M+1}(\lambda), \\ (1 - \lambda)x_2 = v_{M-1} + f, \end{cases} \quad \text{(A.22)}$$

where the first equation is just the Laplace transform of equation (A.17) (recall that we use the definition of Laplace transform (A.7) with $\tau_m = \tau_m^V$), the second and third equations are equation (A.19) for $n = M - 2$ and the last equation is the Laplace transform of the equation for x_2, which in the time domain reads as $\gamma_m^V \dot{x}_2 = -k_m^V (x_2 - v_{M-1}) + f$.

Eliminating v_{-M+1} and v_{M-1} from these equations, using the recurrence equations (A.20) and the result (A.21) we finally get the coupled equations:

$$\begin{cases} \left(\gamma_m^U \dfrac{N}{3} + \gamma_m^V \dfrac{2M}{3} \right) \dot{x}_1 + \gamma_m^V \dfrac{2M}{6} \dot{x}_2 \\ \quad = -\dfrac{k_m^U}{N} x_1 + \dfrac{k_m^V}{2M} (x_2 - x_1), \\ \gamma_m^V \dfrac{2M}{3} \dot{x}_2 + \gamma_m^V \dfrac{2M}{6} \dot{x}_1 = -\dfrac{k_m^V}{2M} (x_2 - x_1) + f. \end{cases} \quad (A.23)$$

At this point we reintroduce the free energy of the polymer chain, defining $N_1 \equiv N$ and $N_2 \equiv 2M - 1 \sim 2M$:

$$F(x_1, x_2) = \frac{k_m^U}{2N_1} x_1^2 + \frac{k_m^V}{2N_2} (x_2 - x_1)^2, \quad (A.24)$$

and a matrix

$$\Gamma \equiv \begin{pmatrix} \gamma_m^U \dfrac{N_1}{3} + \gamma_m^V \dfrac{N_2}{3} & \gamma_m^V \dfrac{N_2}{6} \\ \gamma_m^V \dfrac{N_2}{6} & \gamma_m^V \dfrac{N_2}{3} \end{pmatrix} \quad (A.25)$$

so that we can write the above system as

$$\Gamma_{ij} \dot{x}_j = -\frac{\partial F}{\partial x_i} + f_i + \eta_i, \quad (A.26)$$

where $\vec{f} = (0, f)$ is the external force vector and we reintroduced the noise term $\vec{\eta}$ that we neglected before.

The correlation function of the noise at this point is determined by the requirement that the fluctuation–dissipation relation is verified. This imposes that

$$\langle \eta_i(t) \eta_j(0) \rangle = 2 k_B T \Gamma_{ij} \delta(t). \quad (A.27)$$

A.3. Beads

At this point, we should add the beads that are used for the optical manipulation of polymers. These beads are optically tweezed or subjected to magnetic fields in order to apply forces to the polymers. In the former case, the force acting on the bead is a harmonic force $f = -k(x - X)$, while in the latter it is constant, $f = f_{ext}$. Each bead is characterized by a friction coefficient that can be computed using the Stokes law; we denote it by γ. Typically they are of the order of 10^{-5} pN s nm^{-1}, i.e. much bigger than the microscopic viscosity of the polymers $\gamma_m \sim 10^{-8}$ pN s nm^{-1}.

In the presence of a bead attached to the endpoint of a polymer, the equations of motion (A.1), (A.18), etc, remain valid, but one should add the contribution of γ to the viscosity of the coordinate describing the position of the bead. For instance, if there is a bead attached to the endpoint u_N, the last equation of (A.1) reads as

$$(\gamma + \gamma_m) \dot{u}_N = -k_m u_N + k_m u_{N-1} + f + \eta_N. \quad (A.28)$$

Then the above derivation still holds because the last equation is not used until the end. The only modification will be the

inclusion of γ on the diagonal element Γ_{ii} corresponding to the coordinate of the bead.

Therefore to describe the beads attached to the end of the molecular construction in figure 1, we modify the matrix Γ as above, and in case A, we add to the free energy a term $\frac{1}{2} k (x_4 - X)^2$, while in case B we add a term $-f_{ext} x_3$.

In the case of figure 1(A), one also has to include the left bead. In this case, if we call V the first polymer after the bead, we can start from a system of equations identical to (A.22), but with the first equation replaced by

$$-\gamma \dot{x}_1 = -k x_1 + k_m^V (v_{-M+1} - x_1). \quad (A.29)$$

This will again lead to (A.26) with

$$\Gamma \equiv \begin{pmatrix} \gamma + \gamma_m^V \dfrac{N_2}{3} & \gamma_m^V \dfrac{N_2}{6} \\ \gamma_m^V \dfrac{N_2}{6} & \gamma_m^V \dfrac{N_2}{3} \end{pmatrix} \quad (A.30)$$

and

$$F(x_1, x_2) = \frac{k}{2} x_1^2 + \frac{k_m^V}{2N_2} (x_2 - x_1)^2. \quad (A.31)$$

A.4. Description of a generic setup

The arguments of the previous section suggest that in the general case, a bead can be treated 'as a particular instance of a polymer'. In other words, we can consider the setups in figure 1 as chains of p joint elements $U = U_1, U_2, \ldots, U_p$; each element can be an 'optical trap' (i.e. a spring) or a polymer of $N_1, N_2, \ldots, N_p$ monomers respectively (in the case of an optical trap, we set by default $N_i = 1$). The endpoint of each element is denoted by x_i, and $\vec{x} \equiv (x_1, x_2, \ldots, x_p)$ is the state vector of the system (we also define $x_0 \equiv 0$).

Then, the total free energy is $F(\vec{x}) = \sum_{i=1}^{p} W_{U_i}(x_i - x_{i-1})$ where $W_{U_i}(x) = \frac{1}{2} k x^2$ for an optical trap of stiffness k. Then equation (A.26) holds, with i, j running from 1 to p and the noise correlation matrix is given by (A.27).

The matrix Γ must be constructed as follows. Each diagonal term Γ_{ii}, related to x_i, is the sum of a Stokes term coming from a bead possibly attached to x_i and the contribution coming from the two elements adjacent to x_i (except for $i = p$ when there is only one contribution):

$$\Gamma_{ii} = \gamma + \gamma_m^{U_i} \frac{N_i}{3} + \gamma_m^{U_{i+1}} \frac{N_{i+1}}{3} (1 - \delta_{ip}) \quad (A.32)$$

(the first term is present only if there is a bead attached to x_i). All the off-diagonal elements are zero except those adjacent to the diagonal (i.e. connecting x_i and $x_{i\pm1}$) which get a contribution from the polymer connecting these two ends:

$$\Gamma_{i,i+1} = \Gamma_{i+1,i} = \gamma_m^{U_{i+1}} \frac{N_{i+1}}{6}, \qquad i = 1, \ldots, p-1. \quad (A.33)$$

Note that this final formulation is independent of the Gaussian form of $F(\vec{x})$ that we assumed in the derivation; therefore, we will also use it for non-Gaussian polymers substituting the appropriate form of $F(\vec{x})$ in equation (A.26).

To conclude this section, note that a further check of the quality of the first-order approximation can be done as follows. If we consider a single polymer made of $N_1 + N_2$ bases, the corresponding relaxation time is predicted to be $\tau = \tau_m (N_1 + N_2)^2 / 3$. On the other hand, we could

consider two coupled polymers of N_1 and N_2 bases following equation (A.23) for $y_m^{U,V} = \gamma_m$ and $k_m^{U,V} = k_m$ respectively. The coupled equation can be exactly solved and yields two distinct relaxation times (that typically differ by a factor of 10); the slowest relaxation time can be compared with $\tau = \tau_m(N_1 + N_2)^2/3$. We found that the difference is at most 20%, and the error is maximal for $N_1 \sim N_2$ while it decreases when one of the two polymers is much longer than the other.

Appendix B. Transition rates for the fork dynamics

We now consider a fork n attached to the polymers. For simplicity, we consider the case of a single polymer whose extension is x and free energy is $W(x, n)$. We want to construct a stochastic process that samples the equilibrium distribution $P_{eq}(x, n) = e^{-\beta W(x,n) - G(n;B)}/Z$, where $-G(n; B)$ is the free energy gain in closing the first n bases of DNA, as defined in equation (6).

The random process is constructed as follows. The Langevin equation discussed in the previous section is discretized with time step Δt. If at a given time t the system is in a state (x, n), we allow three possible transitions:

- $(x, n) \to (x + \Delta x, n)$ with rate $H^s(x, n, \Delta x)$,
- $(x, n) \to (x + \Delta x, n + 1)$ with rate $H^o(x, n, \Delta x)$,
- $(x, n) \to (x + \Delta x, n - 1)$ with rate $H^c(x, n, \Delta x)$.

We must have

$$\int \mathrm{d}\Delta x\, H^s(x, n, \Delta x) + H^o(x, n, \Delta x) + H^c(x, n, \Delta x) = 1. \tag{B.1}$$

Moreover we can define rates $r^{s,o,c}(x, n) = \int \mathrm{d}\Delta x\, H^{s,o,c}(x, n, \Delta x)$ that represent the rates to stay, open or close n independent of Δx. In a practical implementation we first decide whether to open, close or stay according to $r^{s,o,c}$, and then extract Δx from the distribution $H^{s,o,c}(x, n, \Delta x)/r^{s,o,c}(x, n)$.

The detailed balance conditions read as

$$P(n, x)H^o(x, n, \Delta x)$$
$$= P(n + 1, x + \Delta x)H^c(n + 1, x + \Delta x, -\Delta x)$$
$$P(n, x)H^c(x, n, \Delta x)$$
$$= P(n - 1, x + \Delta x)H^o(n - 1, x + \Delta x, -\Delta x) \tag{B.2}$$
$$P(n, x)H^s(x, n, \Delta x)$$
$$= P(n, x + \Delta x)H^s(n, x + \Delta x, -\Delta x).$$

We assume that the rate for opening is given by the product of a term that only depends on the binding free energy as in equation (21) and a term corresponding to a standard Langevin step:

$$H^o(x, n, \Delta x) = r\Delta t\, e^{G(n;B) - G(n+1;B)} \sqrt{\frac{4\pi T \Delta t}{\gamma_n}}$$
$$\times \exp\left[-\frac{\gamma_n}{4T\Delta t}\left(\Delta x - \frac{f(x, n)\Delta t}{\gamma_n}\right)^2\right]. \tag{B.3}$$

Note that integrating over Δx we find $r^o(x, n) = r\Delta t\, e^{G(n;B) - G(n+1;B)} = r\Delta t\, e^{-g_0(b_{n+1}, b_{n+2})}$, consistent with equation (21).

Now it is easy to show that the following expression for $H^c(x, n, \Delta x)$ follows from the second detailed balance condition:

$$H^c(x, n, \Delta x) = r\Delta t\, e^{\beta W(x,n) - \beta W(x+\Delta x, n-1)} \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}}$$
$$\times \exp\left[-\frac{\gamma_{n-1}}{4T\Delta t}\left(\Delta x + \frac{f(x + \Delta x, n - 1)\Delta t}{\gamma_{n-1}}\right)^2\right] \tag{B.4}$$

and that the first condition is then automatically satisfied. Up to now, we did not specify the form for $f(x, n)$. However for a generic $f(x, n)$, the above rate is not Gaussian. To obtain a Gaussian rate, we assume that

$$f(x, n) = -\frac{\partial W(x, n)}{\partial x}, \tag{B.5}$$

and perform the following simplifications assuming that Δt is small:

$$H^c(x, n, \Delta x) = r\Delta t\, e^{\beta W(x,n) - \beta W(x, n-1)}$$
$$\times e^{\beta W(x, n-1) - \beta W(x+\Delta x, n-1) - \beta f(n-1, x+\Delta x)} \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}}$$
$$\times \exp\left[-\frac{\gamma_{n-1}}{4T\Delta t}\left(\Delta x - \frac{f(x + \Delta x, n - 1)\Delta t}{\gamma_{n-1}}\right)^2\right]$$
$$\sim r\Delta t\, e^{\beta W(x,n) - \beta W(x, n-1)} \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}}$$
$$\times \exp\left[-\frac{\gamma_{n-1}}{4T\Delta t}\left(\Delta x - \frac{f(x + \Delta x, n - 1)\Delta t}{\gamma_{n-1}}\right)^2\right.$$
$$\left. + \frac{\beta}{2}\frac{\partial^2 W(x, n-1)}{\partial x^2}\Delta x^2\right]. \tag{B.6}$$

Neglecting $O(\Delta x^3)$ one obtains a Gaussian distribution for Δx, and computing the first and second moments of the Gaussian one can see that at the lowest order in Δt it is equivalent to

$$H^c(x, n, \Delta x) = r\Delta t\, e^{\beta W(x,n) - \beta W(x, n-1)} \sqrt{\frac{4\pi T \Delta t}{\gamma_{n-1}}}$$
$$\times \exp\left[-\frac{\gamma_{n-1}}{4T\Delta t}\left(\Delta x - \frac{f(x, n - 1)\Delta t}{\gamma_{n-1}}\right)^2\right]. \tag{B.7}$$

From the above expression, we deduce that the rate for closing is $r^c(x, n) = r\Delta t\, e^{\beta W(x,n) - \beta W(x, n-1)}$, and one *first* has to close and then perform a Langevin step with force $f(x, n - 1)$ and friction γ_{n-1}.

Finally, the rate at constant n is simply given by

$$H^s(x, n, \Delta x) = [1 - r^o(x, n) - r^c(x, n)] \sqrt{\frac{4\pi T \Delta t}{\gamma_n}}$$
$$\times \exp\left[-\frac{\gamma_n}{4T\Delta t}\left(\Delta x - \frac{f(x, n)\Delta t}{\gamma_n}\right)^2\right], \tag{B.8}$$

and it is easy to see that this verifies the third detailed balance equation if equation (B.5) holds and higher orders in Δt are neglected.

To resume, the implementation of the algorithm is as follows.

(1) Choose whether to stay, open or close, with rates $r^{s,o,c}(x, n)$ respectively.

(2) If open, *first* perform a Langevin step at n and *then* increase n by 1.

(3) If close, *first* decrease n by 1 and *then* perform a Langevin step at $n - 1$.

(4) If stay, just perform a Langevin step at n.

(5) Go to 1.

The extension of the above derivation to a case where many polymers are present is straightforward, since the only polymers whose rates are coupled with n are the two adjacent ones. All the other polymers are not influenced by n, and one can use standard discretized Langevin dynamics.

References

[1] Turner P C, McLennan A G, Bates A D and White M R H 2000 *Molecular Biology* (Berlin: Springer)

[2] Bloomfield V A, Crothers D M and Tinoco I 2000 *Nucleic Acids: Structures, Properties, and Functions* (Mill Valley, CA: University Science Books)

[3] Bustamante C, Bryant Z and Smith S B 2003 Ten years of tension: single-molecule DNA mechanics *Nature* **421** 423–7

[4] Marko J F and Cocco S 2003 The micromechanics of DNA *Phys. World* **16** 37–41

[5] Smith S B, Finzi L and Bustamante C 1992 Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads *Science* **258** 1122–6

[6] Cluzel P, Lebrun A, Heller C, Lavery R, Viovy J L, Chatenay D and Caron F 1996 DNA: an extensible molecule *Science* **271** 792

[7] Smith S B, Cui Y and Bustamante C 1996 Overstretching B-dNA: the elastic response of individual double-stranded and single-stranded DNA molecules *Science* **271** 795

[8] Essevaz-Roulet B, Bockelmann U and Heslot F 1997 Mechanical separation of the complementary strands of DNA *Proc. Natl Acad. Sci.* **94** 11935–40

[9] Bockelmann U, Essevaz-Roulet B and Heslot F 1998 DNA strand separation studied by single molecule force measurements *Phys. Rev.* E **58** 2386–94

[10] Bockelmann U, Thomen P, Essevaz-Roulet B, Viasnoff V and Heslot F 2002 Unzipping DNA with optical tweezers: high sequence sensitivity and force flips *Biophys. J.* **82** 1537–53

[11] Thomen P, Bockelmann U and Heslot F 2002 Rotational drag on DNA: a single molecule experiment *Phys. Rev. Lett.* **88** 248102

[12] Bockelmann U, Thomen P and Heslot F 2004 Dynamics of the DNA duplex formation studied by single molecule force measurements *Biophys. J.* **87** 3388–96

[13] Manosas M, Collin D and Ritort F 2006 Force-dependent fragility in RNA hairpins *Phys. Rev. Lett.* **96** 218301

[14] Manosas M, Wen J D, Li P T X, Smith S B, Bustamante C, Tinoco I Jr and Ritort F 2007 Force unfolding kinetics of RNA using optical tweezers: II. Modeling experiments *Biophys. J.* **92** 3010

[15] Liphardt J, Onoa B, Smith S B, Tinoco I and Bustamante C 2001 Reversible unfolding of single RNA molecules by mechanical force *Science* **292** 733–7

[16] Danilowicz C, Coljee V W, Bouzigues C, Lubensky D K, Nelson D R and Prentiss M 2003 DNA unzipped under a constant force exhibits multiple metastable intermediates *Proc. Natl Acad. Sci.* **100** 1694–9

[17] Danilowicz C, Kafri Y, Conroy R S, Coljee V W, Weeks J and Prentiss M 2004 Measurement of the phase diagram of DNA unzipping in the temperature–force plane *Phys. Rev. Lett.* **93** 78101

[18] Weeks J D, Lucks J B, Kafri Y, Danilowicz C, Nelson D R and Prentiss M 2005 Pause point spectra in DNA constant-force unzipping *Biophys. J.* **88** 2752–65

[19] Harlepp S, Marchal T, Robert J, Léger J F, Xayaphoummine A, Isambert H and Chatenay D 2003 Probing complex RNA structures by mechanical force *Eur. Phys. J.* E **12** 605–15

[20] van Oijen A M, Blainey P C, Crampton D J, Richardson C C, Ellenberger T and Xie X S 2003 Single-molecule kinetics of λ exonuclease reveal base dependence and dynamic disorder *Science* **301** 1235–8

[21] Perkins T T, Dalal R V, Mitsis P G and Block S M 2003 Sequence-dependent pausing of single lambda exonuclease molecules *Science* **301** 1914–8

[22] Wuite G J, Smith S B, Young M, Keller D and Bustamante C 2000 Single-molecule studies of the effect of template tension on T7 DNA polymerase activity *Nature* **404** 103–6

[23] Maier B, Bensimon D and Croquette V 2000 Replication by a single DNA polymerase of a stretched single-stranded DNA *Proc. Natl Acad. Sci.* **97** 12002

[24] Levene M J, Korlach J, Turner S W, Foquet M, Craighead H G and Webb W W 2003 Zero-mode waveguides for single-molecule analysis at high concentrations *Science* **299** 682–6

[25] Lang M J, Fordyce P M and Block S M 2003 Combined optical trapping and single-molecule fluorescence *J. Biol.* **2**

[26] Sauer-Budge A F, Nyamwanda J A, Lubensky D K and Branton D 2003 Unzipping kinetics of double-stranded DNA in a nanopore *Phys. Rev. Lett.* **90** 238101

[27] Mathé J, Visram H, Viasnoff V, Rabin Y and Meller A 2004 Nanopore unzipping of individual DNA hairpin molecules *Biophys. J.* **87** 3205–12

[28] Lionnet T, Dawid A, Bigot S, Barre F X, Saleh O A, Heslot F, Allemand J F, Bensimon D and Croquette V 2006 DNA mechanics as a tool to probe helicase and translocase activity *Nucl. Acids Res.* **34** 4232

[29] Lionnet T, Spiering M M, Benkovic S J, Bensimon D and Croquette V 2007 Real-time observation of bacteriophage T4 gp41 helicase reveals an unwinding mechanism *Proc. Natl Acad. Sci.* **104** 19790

[30] Lubensky D K and Nelson D R 2002 Single molecule statistics and the polynucleotide unzipping transition *Phys. Rev.* E **65** 31917

[31] Cocco S, Marko J F, Monasson R, Sarkar A and Yan J 2003 Force–extension behavior of folding polymers *Eur. Phys. J.* E **10** 249–63

[32] Mangeol P, Côte D, Bizebard T, Legrand O and Bockelmann U 2006 Probing DNA and RNA single molecules with a double optical tweezer *Eur. Phys. J.* E **19** 311–7

[33] Greenleaf W J, Woodside M T, Abbondanzieri E A and Block S M 2005 Passive all-optical force clamp for high-resolution laser trapping *Phys. Rev. Lett.* **95** 208102

[34] Cocco S, Monasson R and Marko J F 2002 Force and kinetic barriers to initiation of DNA unzipping *Phys. Rev.* E **65** 41907

[35] Collin D, Ritort F, Jarzynski C, Smith S B, Tinoco I Jr and Bustamante C 2005 Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies *Nature* **437** 231

[36] Woodside M T, Behnke-Parks W M, Larizadeh K, Travers K, Herschlag D and Block S M 2006 Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins *Proc. Natl Acad. Sci.* **103** 6190–5

[37] Baldazzi V, Cocco S, Marinari E and Monasson R 2006 Inference of DNA sequences from mechanical unzipping: an ideal-case study *Phys. Rev. Lett.* **96** 128102

[38] Harris T D *et al* 2008 Single-molecule DNA sequencing of a viral genome *Science* **320** 106

[39] Thompson R E and Siggia E D 1995 Physical limits on the mechanical measurement of the secondary structure of bio-molecules *Europhys. Lett.* **31** 335

[40] Hyeon C, Morrison G and Thirumalai D 2008 Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes *Proc. Natl Acad. Sci.* **105** 9604–9

[41] Zuker M 2000 Calculating nucleic acid secondary structure *Curr. Opin. Struct. Biol.* **10** 303–10

[42] SantaLucia J 1998 A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics *Proc. Natl Acad. Sci.* **95** 1460–5

[43] Doi M and Edwards S F 1986 *The Theory of Polymer Dynamics (International Series of Monographs on Physics* vol 73) (Oxford: Clarendon)

[44] Rouse P E Jr 1953 A theory of the linear viscoelastic properties of dilute solutions of coiling polymers *J. Chem. Phys.* **21** 1272–80

On the trajectories and performance of Infotaxis, an information-based greedy search algorithm

C. BARBIERI[2(a)], S. COCCO[1,2] and R. MONASSON[1,3]

[1] *Simons Center for Systems Biology, Institute for Advanced Study - Einstein Drive, Princeton, NJ 08540, USA*
[2] *Lab. Physique Statistique de l'ENS, CNRS and Univ. Paris 6 - 24 rue Lhomond, 75231 Paris, France, EU*
[3] *Lab. Physique Théorique de l'ENS, CNRS and Univ. Paris 6 - 24 rue Lhomond, 75231 Paris, France, EU*

Abstract – We present a continuous-space version of Infotaxis, a search algorithm where a searcher greedily moves to maximize the gain in information about the position of the target to be found. Using a combination of analytical and numerical tools we study the nature of the trajectories in two and three dimensions. The probability that the search is successful and the running time of the search are estimated. A possible extension to non-greedy search is suggested.

Introduction. – Reaching a target with limited information is a fundamental task for living organisms. Small organisms, such as bacteria and eukaryotic cells, are thought to estimate and ascend the gradient of nutrient concentration, a process called chemotaxis [1,2]. At the scale of a larger organism the Reynolds number is higher [3]: most biologically relevant chemical fields become turbulent and dilute. As a result, the trajectories of insects following odor traces appear much more complex than those of smaller organisms [4]. The modeling of search processes in the presence of noisy information is important not only for biology, but also for robotics [5,6].

Assume that the search has proceeded for some time, and the searcher has received some hits, *i.e.* has detected some molecule of odor sent by the target, along the trajectory. How should the searcher move next? The timing and locations of the hits, as well as the absence of hits along the remaining parts of the trajectory all provide useful information about the location $\mathbf{y}$ of the target. In Bayesian terms, this defines a posterior probability $P_t(\mathbf{y})$ for the position of the target. Going towards the maximum of P_t is not an optimal strategy as the maximum does generally not coincide with the target, especially in the initial stage of the search process. Recently, Vergassola, Villermaux and Shraiman proposed an alternative strategy, called Infotaxis [7]. The searcher moves to maximize the (expected) gain in information about the location of the target, that is, the loss in the entropy of the distribution $P_t(\mathbf{y})$. As the search goes on, the entropy typically decreases, until the source is finally located. The strategy naturally balances the needs for exploration (harvesting more information about the target location) and exploitation (going towards the maximum of P_t). Infotaxis was implemented and tested on two-dimensional square lattices[1]: the target was almost always found, and the distribution of search time appeared to decay exponentially.

Yet some important questions about Infotaxis remain open. First, how well does the algorithm perform in three dimensions? In addition to its practical interest, this question arises naturally in the context of the Brownian motion theory. Purely random walks are space-filling in two dimensions, and transient in higher-dimensional spaces. Finding a target in three dimensions is therefore much harder, and constitutes a real test for the capabilities of Infotaxis. Secondly, how dependent on the underlying lattice are the results reported in [7]? Realistic descriptions of animal behavior or implementations in biomimetic robots require us to consider continuous spaces. In addition the presence of a lattice introduces anisotropies, while the odor propagation model used in [7] was isotropic. Thirdly, two-dimensional trajectories seem to exhibit spiral-like shapes. How precisely can we

---

[(a)]E-mail: barbieri@lps.ens.fr

[1]Few short trajectories were obtained on small three-dimensional lattices in [8].

characterize those spirals, and what are their counterparts in three dimensions?

In this letter, we derive the equation of motion for the Infotaxis searcher in the continuous space. We then introduce an algorithm to solve this equation[2]. The performances of Infotaxis, *i.e.* the probability of success and the distribution of the search times are studied in $D = 2$ and 3 dimensions. The spiral- and coil-like shape of the search trajectories for, respectively, $D = 2$ and 3, and the pinning of the trajectory around the received hits are investigated analytically and numerically. We show that the motion of a searcher receiving an average and deterministic signal is a good predictor of the typical properties of the motion in the presence of stochastic hits. Finally we discuss a possible extension to a non-greedy search strategy, which could help reduce pinning effects.

Equation of motion for the searcher. – A point-source in $\mathbf{y}^*$ emits particles, which diffuse in space, and have a finite lifetime. In the stationary regime, the probability per unit of time to encounter a particle in $\mathbf{x}$ is denoted by $R(\mathbf{y}^* - \mathbf{x})$ (for supplementary information see ref. [7]). Function R has an integrable divergence at the origin ($R(u) \sim -\log u$ in $D = 2$, $\sim \frac{1}{u}$ in $D = 3$ dimensions, when $u \to 0$), and exponentially decreasing tails for large distances u. In the following distances are measured in the unit of the decay length of R. The unit of time is the inverse of the rate of emission of particles by the source, divided by the (dimensionless) linear size, a, of the searcher for $D = 3$, or multiplied by $\log(1/a)$ for $D = 2$ [7].

Let $\mathbf{x}(t)$ be the position of the searcher at time t. We denote by N_H the number of particles detected (called hits) at earlier times, $0 \leqslant t_i \leqslant t$, with $i = 1, \ldots, N_H$. Based on those hits, the searcher can draw a probabilistic map over the possible locations $\mathbf{y}$ of the source. In the Bayesian framework, the posterior probability density $P_t(\mathbf{y})$ for the location of the source is the (normalized) product of the probabilities of having detected the N_H particles at locations $\mathbf{x}(t_i)$, times the probability of not having detected any particle at other locations along the trajectory, times the prior probability density P_0 over $\mathbf{y}$,

$$P_t(\mathbf{y}) \propto \prod_{i=1}^{N_H} R(\mathbf{y} - \mathbf{x}(t_i)) \, e^{-\int_0^t \mathrm{d}t' R(\mathbf{y} - \mathbf{x}(t'))} \, P_0(\mathbf{y}). \quad (1)$$

Hence, P_t diverges where the hits have been received and vanishes in the other places along the trajectory. In the following we will use brackets to denote averages over this posterior distribution:

$$\langle f(\mathbf{y}) \rangle_{\mathbf{y};t} = \int \mathrm{d}\mathbf{y} \, P_t(\mathbf{y}) \, f(\mathbf{y}). \quad (2)$$

Assume now that the searcher stays in $\mathbf{x}(t)$ during an infinitesimal time δt. The number of hits, n, received during this time interval is a stochastic variable equal to

---

zero or one, with probabilities $p(0|\mathbf{y}) = 1 - \delta t \, R(\mathbf{y} - \mathbf{x})$ and $p(1|\mathbf{y}) = \delta t \, R(\mathbf{y} - \mathbf{x})$, depending on the location $\mathbf{y}$ of the source. In the language of information theory, the particle emission and detection system can be thought as a noisy channel, and n is the output message associated to the input codeword $\mathbf{y}$. The mutual information δI between n and $\mathbf{y}$ is

$$\delta I = \sum_{n=0,1} \left\langle P(n|\mathbf{y}) \log \left(\frac{P(n|\mathbf{y})}{\langle P(n|\mathbf{y}') \rangle_{\mathbf{y}';t}} \right) \right\rangle_{\mathbf{y};t} = -\delta t \, V_t(\mathbf{x}(t)) \quad (3)$$

up to $O(\delta t^2)$, where

$$V_t(\mathbf{x}) = \left\langle R(\mathbf{y} - \mathbf{x}) \log \left(\frac{\langle R(\mathbf{y}' - \mathbf{x}) \rangle_{\mathbf{y}';t}}{R(\mathbf{y} - \mathbf{x})} \right) \right\rangle_{\mathbf{y};t}, \quad (4)$$

is the entropy rate of the posterior distribution P_t. Infotaxis stipulates that δI should be maximized, or, equivalently that V_t should be minimized, *i.e.* made as negative as possible. In other words, interpreting V_t as a potential, the searcher should descend the gradient of V_t. A natural equation of motion is then

$$\gamma(t) \, \dot{\mathbf{x}}(t) = -\nabla_{\mathbf{x}} V_t(\mathbf{x}(t)), \quad (5)$$

where $\gamma(t)$ plays the role of a friction coefficient. A possible choice is $\gamma(t) = |\nabla_{\mathbf{x}} V_t(\mathbf{x}(t))|/v_0$, to keep the modulus of the velocity fixed and equal to v_0. This is close to the lattice version of [7], where the searcher could either stay immobile or move by one lattice site, and the velocity could take only one non-zero value. In general, $\gamma(t)$ can be any positive function of the time. In the following, we will first consider the case $\gamma(t) = \gamma$ const. We will later discuss to what extent the trajectories and the performances change when γ varies with time.

Numerical integration. – The equation of motion (5) is highly non-linear and depends on the whole history of the search process through the posterior probability (1). To solve eq. (5) numerically we discretize the time with a step Δt. The positions $\mathbf{x}_\ell$ of the searcher at the instants $t_\ell = \ell \, \Delta t$, where ℓ is a positive integer, are memorized, as well as the occurrences (times) of the hits. The amplitude of the ℓ-th elementary move is estimated from (5): $\mathbf{x}_{\ell+1} - \mathbf{x}_\ell = -\frac{\Delta t}{\gamma(t_\ell)} \nabla_{\mathbf{x}} V_t(\mathbf{x}_\ell)$. The calculation of the gradient of the potential requires to estimate averages over the space $\mathbf{y}$ with measure P_t (2). To do so, we use the importance sampling Monte Carlo method [9]. The term inside the bracket in (4), and its gradient with respect to $\mathbf{x}$ are exponentially decreasing functions of the radial distance $u = |\mathbf{x}_\ell - \mathbf{y}|$. We thus perform the change of variable $u = u_0(1 - v)/v$, where $v \in]0; 1]$, and u_0 is a scale parameter. We draw N_{MC} values of the new variable, v_a, $a = 1, \ldots, N_{MC}$, uniformly and at random, and calculate the corresponding u_a. Angular variables $\mathbf{\Omega}_a$ are uniformly sampled on the unit sphere or circle to obtain the points $\mathbf{y}_a = \mathbf{x}_\ell + u_a \mathbf{\Omega}_a$ in the original space. The corresponding probabilities $P_t(\mathbf{y}_a)$ are computed
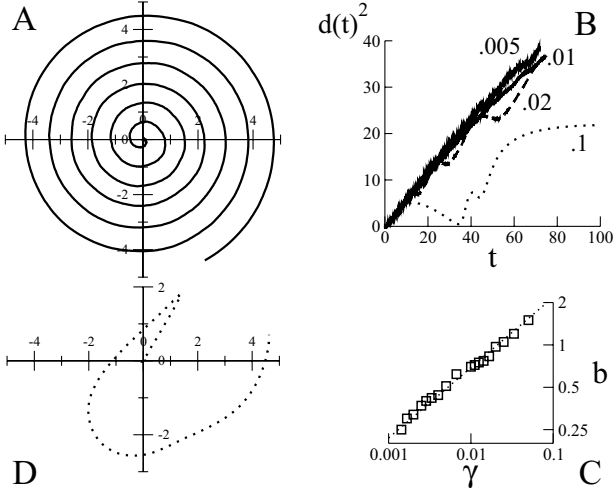
Fig. 1: Two-dimensional trajectories for $\gamma = .01$ (A) and $\gamma = .1$ (D), after one hit is received at time $t = 0$ ($P_0 = R$). B: squared distance to the origin, $d(t)^2$, vs. time t for four values of γ. C: spacing b between turnings vs. γ. The dotted line has slope $\frac{1}{2}$. Integration is done with $N_{MC} = 10^4$ points.

using Simpson's method to calculate the integral over time in (1).

The algorithm of [7] stored and updated P_t, which made the computational time linear in t and in the size of the lattice. Our procedure recalculates P_t at each time step, which makes the computational time quadratic in t, and independent of the a priori infinite size of the space. Errors on P_t do not accumulate with time, and the accuracy is directly controlled by the number of Monte Carlo sampling points, N_{MC}. In addition, the map P_t is guaranteed to be accurately determined where it really matters, i.e. in the vicinity of the searcher.

Initial stage of the search. – We first focus on the initial stage of the search process, which strongly depends on the a priori distribution, P_0. A possible choice is the one-hit prior, $P_0 = R$, which means that the search starts only when a first hit is received at time $t = 0$ [7]. In two dimensions and before subsequent hits are detected, the search trajectories are Archimedean spirals [8,10] for a large range of values of γ (fig. 1A). The squared distance of the searcher to the origin at time t, $d(t)^2$, increases linearly with t, and is, to a large extent, independent of γ (fig. 1B). The spacing b between successive turnings is independent of time and increases as $\sqrt{\gamma}$ (fig. 1C); the velocity $|\dot{\mathbf{x}}| \sim \frac{1}{b}$ decreases with the friction γ as expected.

The increase of b can be intuitively understood: the larger γ, the longer the searcher spends along the trajectory without receiving hits, and the more likely is the source to be located far away. For values of $\gamma (> .08)$ such that b would exceed the length ($= 1$) over which R decreases, the spiral nature of trajectories breaks down (fig. 1D). We have run simulations with a modified potential V_t, where the arguments $\mathbf{y} - \mathbf{x}$ and $\mathbf{y}' - \mathbf{x}$ of the

functions R in (4) were divided by a large factor ($= 10$). The regular spirals then disappeared, and looked like the trajectory in fig. 1D, even for small values of γ.

Simulations with other prior distributions show that the long-distance behavior of P_0 is critical to the existence of spiral trajectories, while the behavior of P_0 close to the origin is irrelevant. Choosing $P_0(\mathbf{y}) \sim \exp(-|\mathbf{y}|/y_0)$ we obtain spirals as long as y_0 is not too large. The reason is that the spacing b is proportional to y_0: spirals explore as much space as allowed by the prior. Spirals breakdown for the (γ-dependent) value of y_0 such that b exceeds 1.

Search trajectories in three dimensions display a more complex structure than their two-dimensional counterparts. Figure 2 shows the motion of the searcher with the one-hit prior, $P_0 = R$. Roughly speaking, the trajectory is constituted of subsequent shells of increasing radii, which are densely covered before a new shell is built. The distance to the origin, $d(t)$, is compatible with $t^{1/3}$ at large times, but grows faster at smaller times (fig. 2). To better understand how trajectories develop in three dimensions, we have resorted to a small-$\mathbf{x}$ expansion of the potential V_t. The relevant contributions to the equation of motion are, up to cubic order,

$$\gamma \dot{\mathbf{x}}(t) = \alpha_1(t)\,\mathbf{x}(t) + \alpha_2(t) \int_0^t \mathrm{d}t'\, \mathbf{x}(t')$$
$$+ \int_0^t \mathrm{d}t'\, \mathbf{x}(t') \left[\beta_1(t)|\mathbf{x}(t')|^2 + \beta_2(t)\,\mathbf{x}(t') \cdot \int_0^t \mathrm{d}t''\mathbf{x}(t'') \right], \quad (6)$$

where the coefficients $\alpha_i(t)$ have explicit analytical expressions. When $\mathbf{x}$ is very small, only the linear terms matter. As $\alpha_1 \simeq \frac{\sqrt{2}}{3e}\frac{\log t}{t} > 0$ for large t, the trajectory tends to follow a line radiating from the origin. However, the straight line is unstable against local bending since $\alpha_2 \simeq -\frac{3\sqrt{3}}{e^2}\frac{\log t}{t^2} < 0$. The trajectory thus acquires a spiral shape confined within the plane spanned by $\mathbf{x}(0)$ and $\dot{\mathbf{x}}(0)$. The presence of cubic terms ($\beta_1, \beta_2 > 0$), which are not constrained to lie in this plane, eventually lead to a cross-over from the quasi-bidimensional spiral to a fully three-dimensional trajectory (fig. 2).

Replacing coefficients α_1, α_2 with the smaller values $\alpha_1 = \frac{a_1}{t}, \alpha_2 = -\frac{a_2}{t^2}$, with $a_1, a_2 > 0$, allows for an exact resolution of (6). A spiral is found when $a_1 < a_2$, with a radius and an angle growing as, respectively, t^ω and $\eta \log t$, where $\omega = \frac{a_1 - \gamma}{2\gamma}$ and $\eta = \sqrt{4a_2\gamma - (\gamma + a_1)^2}$. As γ increases so does the angular velocity ($\propto \eta$), while the growth exponent ω diminishes: spirals stop growing if γ is too large, a fact reminiscent of fig. 1D. Numerical resolution of (6) shows that this scenario is qualitatively unchanged when the logarithmic factors in α_1, α_2 are taken into account.

Pinning after a hit. – Examples of trajectories where the searcher receives hits at times $t_i > 0$ are shown in fig. 3. Generally speaking, the trajectories are denser when more hits are received. After each hit i, the posterior distribution P_t is considerably reinforced in $\mathbf{x}(t_i)$ and its
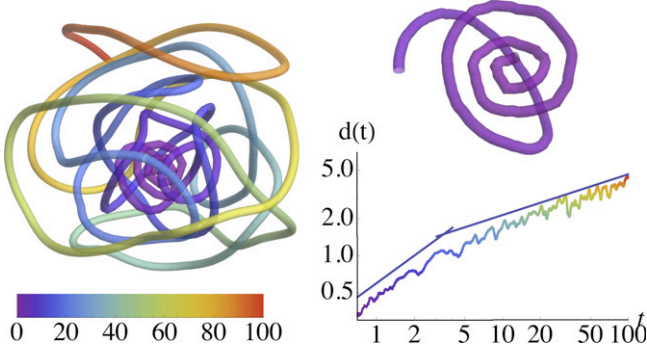
Fig. 2: A three-dimensional trajectory in the absence of hit for $\gamma = .01$ (left), with its quasi–two-dimensional initial portion (top); the time axis is color coded. Bottom: distance to the origin, $d(t)$, compared to the power laws $t^{.75}$, then $t^{1/3}$.

neighborhood. The searcher remains in the immediate vicinity for a certain time, t_w, until the search resumes. Informally speaking, the searcher makes sure that the source is not in $\mathbf{x}(t_i)$ before looking elsewhere. Imagine that the searcher has not moved at all for a period of time t_w after the hit at time t_i. The posterior distribution is then

$$P_{t_i+t_w}(\mathbf{y}) \propto P_{t_i^-}(\mathbf{y})\, R(\mathbf{y} - \mathbf{x}(t_i))\, e^{-t_w\, R(\mathbf{y}-\mathbf{x}(t_i))}. \quad (7)$$

Assuming the posterior distribution right before the hit, $P_{t_i^-}$, is smooth in the vicinity of $\mathbf{x}(t_i)$, the potential $V_{t_i+t_w}(\mathbf{x}(t_i) + \mathbf{u})$ is a function of the small displacement $u = |\mathbf{u}|$ and t_w only. There are a local maximum in $u = 0$, since $\alpha_1 > 0$ in (7), and a global minimum in $u_m(t_w) > 0$. For $t_w < .4$, $u_m < .01$ is smaller than the error on the position deriving from our Monte Carlo integration, while for $t_w > .5$, $u_m > .1$, and the displacement of the searcher is easily seen.

The pinning effect is an important feature of the continuous formulation of Infotaxis, and was not observed on a lattice. The reason is that, at each time step, a whole lattice site probability was set to zero in [7], which created a strong repulsive effect for the searcher and prevented pinning. In the continuous space, however, the trajectory has zero measure and the repulsion is too weak to override the pinning. As the searcher gets closer to the source, the average delay τ between successive hits gets smaller. When $\tau \simeq t_w$, the searcher could, in principle, come to a complete halt. The distance from the source for which this happens, d_{halt}, depends on the dimension: $d_{\mathrm{halt}} = .1$ for $D = 2$, $d_{\mathrm{halt}} = .3$ for $D = 3$.

Performances. – We now introduce a source and observe the trajectory of the searcher reacting to hits (fig. 3). We are interested in the probability that the search process is successful as a function of the initial distance to the source, d_0. If the searcher reaches the neighborhood of the source of radius d_{halt} defined above the source is declared found. If the searcher misses this neighborhood and reaches a distance $d_{\mathrm{fail}} \gg 1$ to the source such that
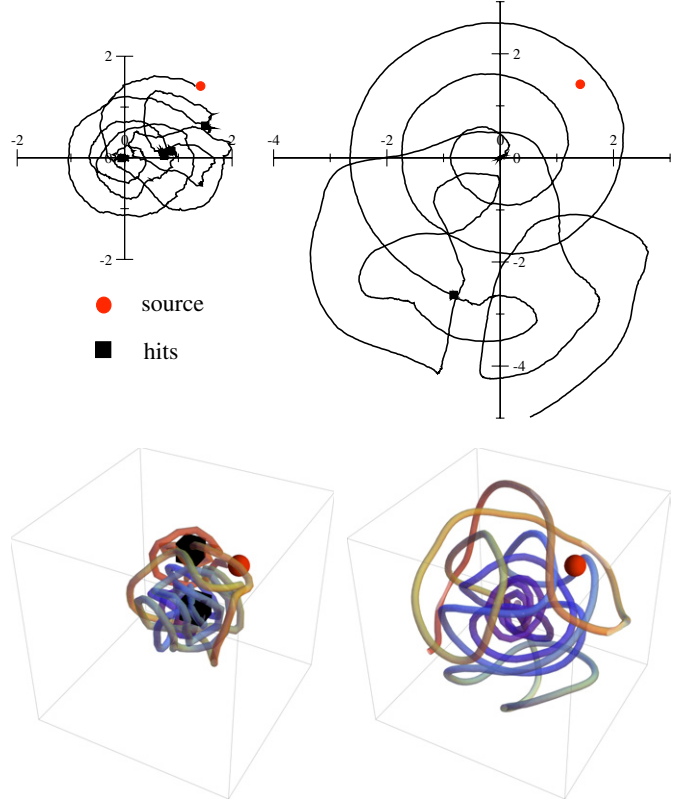
Fig. 3: Examples of search trajectories with hits in $D = 2$ (top, $\gamma = .02$) and $D = 3$ (bottom, $\gamma = .01$) dimensions. Trajectories on the left find the source, while searches on the right are not successful. The initial distance to the source is $d_0 = 2$. Red disks or spheres represent points at distance $< d_{\mathrm{halt}}$ to the source. Black squares or cubes locate the hits; their size corresponds to the amplitude of the erratic motion of the searcher during the pinning after a hit.
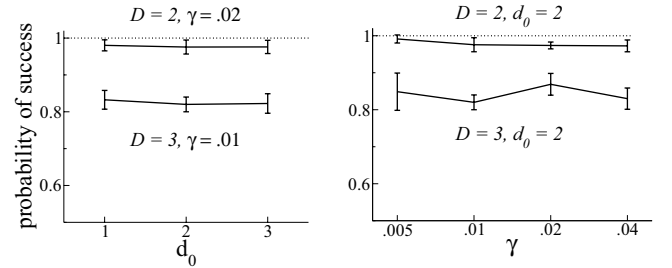
Fig. 4: Probability of success of Infotaxis as a function of the initial distance to the source, d_0 (left), and of the friction γ (right). Top points correspond to $D = 2$ dimensions, bottom points to $D = 3$. The numbers of runs is of about 200 for each point. All probabilities where obtained with $d_{\mathrm{fail}} = 8$.

new hits are highly unlikely, the searcher is declared to be lost. Examples of successful and unsuccessful trajectories are shown in fig. 3.

Figure 4 (left) shows that the probability of success in dimension $D = 2$ is compatible with unity for all distances d_0 smaller than a few units, in agreement with the findings of [7]. On the contrary, in dimension $D = 3$, the probability
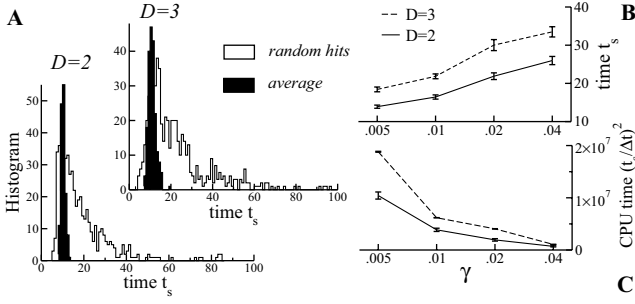
Fig. 5: A: histograms of the search times t_s in $D = 2$ ($\gamma = .02$) and $D = 3$ ($\gamma = .01$) dimensions for an initial distance $d_0 = 2$ to the source. Full histograms correspond to the average trajectories, contour histograms to trajectories with random hits. B: average search time t_s as a function of γ. C) Total CPU time as a function of γ, calculated as $(t_s/\Delta t)^2$.

of success is definitely smaller than one, and is about .8 for distances d_0 ranging from 1 to 3 and for $\gamma = .01$. We have also measured the probability of success at fixed d_0 over a range of values of γ and found little variation (fig. 4, right).

Figure 5A shows that the distribution of the search times t_s of successful runs has a positive skew and a roughly exponential tail not only in dimension $D = 2$ [7] but also in dimension $D = 3$. The exponential nature of the tail was checked for various values of the distance d_0. Figure 5B shows that the average time to find the source, t_s, decreases with γ. The CPU time scales as $A(t_s/\Delta t)^2$, where $\Delta t = \gamma$ is chosen to obtain numerical stability, *i.e.* small enough local moves at any step ℓ. Hence, the CPU time is much larger for small friction (fig. 5C). We find $A \simeq 3\,\text{ms}$ on one core of a 2.4 GHz Intel Core 2 Quad desktop computer, and for $N_{MC} = 10^4$ Monte Carlo steps. The CPU time can be decreased by choosing a smaller value for N_{MC}, or by increasing γ (and Δt) with time.

Case of a time-dependent friction $\gamma(t)$. – The results reported so far correspond to constant frictions γ. We have generated trajectories with various time-dependent functions $\gamma(t)$, *e.g.* slowly increasing to reduce the scaling of the CPU time with t (fig. 6A), or with the fixed-velocity requirement (fig. 6B). From a qualitative point of view, we observe no drastic difference with the constant γ case, provided that $\gamma(t)$ does not exceed the maximal value ($\simeq .08$) at which $b \sim 1$ and spirals break down. For instance, the distance between turns in the spiral region of the trajectory in fig. 6A can be deduced from the value of $b(\gamma(t))$ in fig. 1C. In the fixed-velocity case of fig. 6B, before the first hit is detected, $\gamma(t)$ shows regular oscillations and the trajectory has a spiral-like shape with b corresponding to the maximum value of $\gamma(t)(\simeq .011)$. Larger fluctuations are visible during the erratic motion after each hit, but $\gamma(t)$ remains smaller than the arrest value $\gamma \simeq .08$.

We have seen in fig. 4 that the probability of success is essentially independent of d_0 and γ. This key result can be understood as follows. Far away from the source, hits are
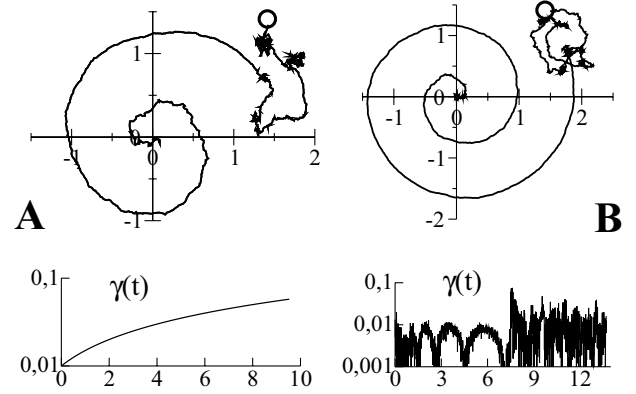

Fig. 6: A: a search trajectory for the time-dependent $\gamma(t)$ shown below. B: a trajectory with fixed velocity $v_0 = 2$, and $\gamma(t) = |\nabla_\mathbf{x} V_t|/v_0$. The source is located in $(\sqrt{2}, \sqrt{2})$ (circle).

very unlikely and the trajectory develops as a spiral. As the area spanned by the trajectory is roughly independent of γ (fig. 1B), the time it takes for the searcher to reach the D-dimensional sphere of radius 1 and centered in the source will be roughly independent of γ and will strongly increase with d_0, while the time spent in the sphere and the number of hits received will be essentially independent of both γ and d_0. We conclude that varying γ has little consequence on the performance of Infotaxis, as long as the spiral-like motion is possible.

Motion in the presence of the "average" signal. – Certain characteristics of the search time distribution, such as the typical (most probable) value of t_s, can be assessed from the study of an abstract searcher receiving an average signal rather than discrete and sparse hits. To define an average posterior density $P_t^{av}(\mathbf{y})$ we remark that $P_t(\mathbf{y})$ contains the product of N_H stochastic R factors over the hits in (1). It is natural to define $P_t^{av}(\mathbf{y})$ through the average value of $\log P_t(\mathbf{y})$ over the hits:

$$P_t^{av}(\mathbf{y}) \propto e^{-\int_0^t \mathrm{d}t'\, R(\mathbf{y}-\mathbf{x}(t')) + R(\mathbf{y}^*-\mathbf{x}(t'))\log R(\mathbf{y}-\mathbf{x}(t'))} P_0(\mathbf{y})$$
$$(8)$$

up to a multiplicative normalization constant; here, $\mathbf{y}^*$ denotes the location of the source as usual. We define the average search trajectory as the solution of eq. (5) with the average $\langle\cdot\rangle$ (2) calculated over the measure (8). Average search trajectories are obviously smoother than trajectories with random hits, but have common features, such as a possible return towards the origin after having been close to the source.

Figure 5A shows that the search time for the average motion is in very good agreement with the typical search time for trajectories with random hits. This coincidence has been observed for all the frictions γ and distance d_0 to the source we have tested. Note that, while the average motion is fully deterministic, some noise is introduced to break the rotational invariance and select the initial phase of the spiral; this noise is responsible for the small width of the full histograms shown in fig. 5A.
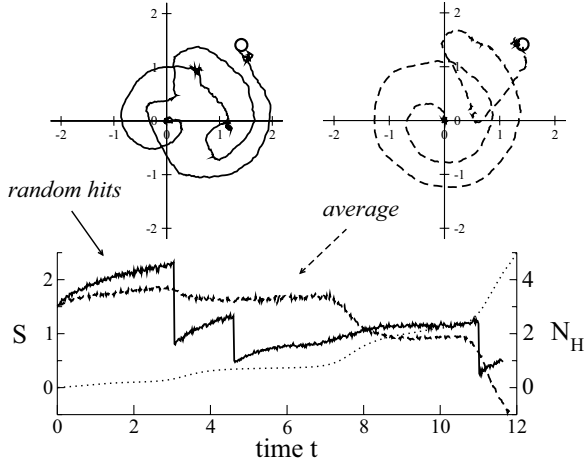
Fig. 7: Entropy $S(t)$ (bottom, left scale) for one trajectory $\mathbf{x}(t)$ obtained with random hits (top left, full curve, 3 hits are received) and the average trajectory (top right, dashed curve). The dotted line shows the average number of hits N_H (right scale) received along the average trajectory. The source is located in $(\sqrt{2}, \sqrt{2})$ (circle).

Figure 7 shows the average search trajectory and a random trajectory in $D = 2$ dimensions, together with the entropies of their posterior distributions. In the random case, the entropy abruptly decreases right after a hit, then increases until the next hit is received (due to the exponential time decay in P_t (1)). As for the average motion, the entropy shows weak oscillations (due to the spiral motion) superimposed to a smooth trend, which decreases as the searcher gets close to the source.

Conclusion. – In this letter we have presented a continuous-space version of Infotaxis, and have analyzed its behavior in two and three dimensions. When the initial distance to the source, d_0, is of the order of the decay length of R, the probability that Infotaxis finds the target is essentially equal to unity in $D = 2$, and is smaller ($\simeq .8$) in $D = 3$ dimensions. The probability of success is roughly independent of $\gamma(t)$ in (5), while the search time and the CPU time strongly depend on the friction. The quadratic increase of the CPU time and the presence of the pinning effect make the computational cost increase as the searcher gets closer to the source. Note that the CPU time could be made linear in t_s (instead of quadratic) if the integral in (1) were restricted to the recent past, *i.e.* if the search had a finite memory.

The pinning of the trajectory by the hits is a direct consequence of the greedy nature of Infotaxis: the searcher moves to maximize the immediate gain in information, irrespectively of what could be gained on a longer-time horizon. To overstep this greedy strategy consider the expected gain in information, $I[\mathbf{x}(t'); t < t' < t + \tau]$, when the searcher plans to move along a portion of trajectory $\mathbf{x}(t')$ during the current time t and the time $t + \tau$. The best portion of trajectory is determined through the maximization of I, minus a quadratic term $\propto \gamma \dot{\mathbf{x}}^2$ penalizing

large velocities. While this variational calculation appears intractable for general τ, a systematic expansion in powers of τ is possible. To the lowest orders in τ the equation of motion becomes

$$\tau^2 \big(\nabla_{\mathbf{x}} \nabla_{\mathbf{x}} V_t(\mathbf{x})\big) \ddot{\mathbf{x}} + \gamma \dot{\mathbf{x}} = -\nabla_{\mathbf{x}} V_t(\mathbf{x}) \qquad (9)$$

up to $O(\tau)$ corrections to the friction γ and to the force on the right-hand side. The introduction of a finite-time horizon, τ, gives birth to an inertial term, with an effective mass tensor proportional to the curvature matrix of the potential (4). This inertial motion could help reduce the pinning following a hit, and avoid the slowing-down of the searcher close to the source. The analysis of (9) and of the search trajectories is left for a future work.

Last of all, the shapes of the trajectories observed in two and three dimensions result from a trade-off between the self-repulsion of the trajectory (the searcher does not come again close to a point where the source was not detected) and the confinement due to the hits or to the prior (the source is likely to be close to a detection). This trade-off is present in physical systems such as polyelectrolytes (charged polymers) confined in a volume or on a surface [11]. It is however unclear how far the analogy between the out-of-equilibrium process generated by Infotaxis and equilibrium polyelectrolytes could be pursued [12].

$* * *$

REFERENCES

[1] ADLER J., *Science*, **153** (1966) 708.
[2] BERG H. C., *Annu. Rev. Biophys. Bioeng.*, **4** (1975) 119.
[3] BALKOVSKY E. and SHRAIMAN B. I., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 12589.
[4] VICKERS N. J. and BAKER T. C., *Proc. Natl. Acad. Sci. U.S.A.*, **91** (1994) 5756.
[5] LOCHMATTER T. and MARTINOLI A., *Exp. Robot.*, **54** (2009) 473.
[6] MORAUD E. M. and MARTINEZ D., *Front. Neurorobot.*, **4** (2010) 1.
[7] VERGASSOLA M., VILLERMAUX E. and SHRAIMAN B. I., *Nature*, **445** (2007) 406.
[8] MASSON J.-B., BAILLY BECHET M. and VERGASSOLA M., *J. Phys. A*, **42** (2009) 434009.
[9] HAMMERSLEY J. M. and HANDSCOMB D. C., *Monte Carlo Methods* (Taylor & Francis) 1964.
[10] BARBIERI C., *Infotassi: un algoritmo di ricerca senza gradienti,* Master Thesis, University of Rome "La Sapienza" (2007).
[11] ANGELESCU D. G., LINSE P., NGUYEN T. T. and BRUINSMA R. F., *Eur. Phys. J. E*, **25** (2008) 323.
[12] AMIT D., PARISI G. and PELITI L., *Phys. Rev. B*, **27** (1983) 1635; PELITI L. and PIETRONERO L., *Riv. Nuovo Cimento*, **10** (1987) 1.

Appendix A

Inference of couplings for a set of leaky integrate and fire neurons

A.1 Introduction

Recent advances in experimental techniques and in the miniaturization of components have permitted the recording of the activity of several neurons at the same time thorugh the use of multi-electrode recordings [Taketani 06].

The observation of substantial correlations in the firing activities of neurons has raised fundamental issues on their functional role [Averbeck 06]. However the problem of inferring the structure of the network and the interaction between different neurons has only recently been attacked (Fig. A.1). The problem is not easy to tackle, because data sets are already quite big and can contain millions of spiking events from up to a hundred neurons. A classical approach to infer functional neural connectivity is through the analysis of pairwise cross-correlations. The approach is versatile and fast, but cannot disentangle direct correlations from common or correlated inputs. Alternative approaches assume a particular dynamical model for the spike generation, such as the generalized linear model, which represents the generation of spikes as a Poisson process with a time-dependent rate, and the Integrate-and-Fire (IF) model, where spikes are emitted according to the dynamics of the membrane potential [Jolivet 04].

While the problem of estimating the model parameters (external current, variance of the noise, capacitance and conductance of the membrane, ...) of a single stochastic IF neuron from the observation of a spike train has received a lot of attention [Paninski 04, Lansky 08], few studies have focused on the inference of interactions in an assembly of IF neurons. Cocco and Monasson have recently proposed a Bayesian algorithm to infer the intereactions and the external currents of a set of leaky integrate and fire neurons[Cocco 09, Monasson 11]. They applied their approach to data coming from real experiments on salamander retinas and validated their results with artificial data and cross-checking with another algorithm based on the Ising model.

An interesting problem they've come across is the disentanglement of the correlations already present in the stimulus and the correlations that come from the topology of the network itself. This has been discussed by comparing datasets coming from the same retina with different stimuli.
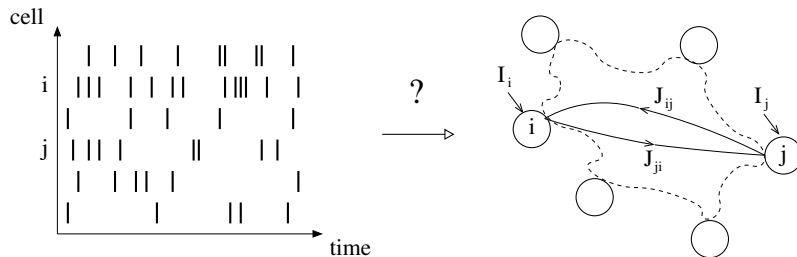
Figure A.1: Left: times $t_{i,k}$ of spikes emitted by a set of neurons (raster plot). Right: network of LIF neurons with couplings J_{ij} and external currents I_i. Given the set of spikes we want to infer the values of the couplings and currents.

A.2   Integrate and fire neurons

Each neuron is represented by the Leaky Integrate-and-Fire (LIF) model (see [Jolivet 04] and references therein). The membrane potential obeys the differential equation,

$$C\frac{dV_i}{dt}(t) = -g\,V_i(t) + \sum_{j(\neq i)} J_{ij} \sum_k \delta(t - t_{j,k}) + I_i + \eta_i(t) \;, \qquad (A.1)$$

where C, g are, respectively, the membrane capacitance and conductance. J_{ij} is the strength of the connection from neuron j onto neuron i and $t_{j,k}$ the time at which cell j fires its kth spike; we assume that synaptic inputs are instantaneously integrated *i.e.* the synaptic integration time is much smaller than the membrane leakage time, C/g, and the typical inter-spike interval. I_i is a constant external current flowing into cell i, and $\eta_i(t)$ is a fluctuating current, modeled as a Gaussian white noise process with variance σ^2. Neuron i remains silent as long as V_i remains below the threshold potential V_{th} (set to unity in the following). If the threshold is reached at some time then a spike is emitted, and the potential is reset to its rest value (which can be set to zero without loss of generality), and the dynamics resumes.

The above model (A.1) implicitly defines the likelihood P of the spiking times $\{t_{j,k}\}$ given the currents I_i and synaptic couplings J_{ij}. If we are given the spiking times $\{t_{j,k}\}$ we will infer the couplings and currents by maximizing P. In principle P can be calculated through the resolution of Fokker-Planck equations (one for each inter-spike interval) for a one-dimensional Orstein-Uhlenbeck process with moving boundaries. However this approach, or related numerical approximations [Paninski 04], are inadequateis too slow to treat data sets with hundreds of thousands of spikes.

In Cocco's and Monasson's approach P is approximated from the contribution coming from the most probable trajectory for the potential for each cell i, referred to as $V_i^*(t)$. This semi-classical approximation is exact when the amplitude σ of the noise is small. The determination of $V_i^*(t)$ was done numerically by Paninski for one cell in [Paninski 06]. What they proposed is a fast algorithm to determine $V_i^*(t)$ analytically in a time growing linearly with the number of spikes and quadratically with the number of neurons, which allows them to process recordings with tens of neurons easily. The algorithm is based on a detailed and analytical resolution of the coupled equations for the optimal potential $V_i^*(t)$ and the associated optimal noise $\eta_i^*(t)$ through (A.1), since this work has not been performed during this thesis and a full explanation would take many pages we will not discuss the details of the algorithm. The interest reader can find an explanation of the approach in a recent publication [Monasson 11].

Once the optimal paths for the potential and noise have been determined, one cancalculate the log-likelihood of the corresponding couplings and currents through the integral of

the squared optimal noise. This log-likelihood is a concave function of the currents and couplings and can be easily maximized using convex optimization procedures. Measure of the curvature of the log-likelihood allows us to estimate the error bars on the inferred parameters.

A.3 Limitations of the original implementation

Even though the conception of the algorithm and the subsequent testing performed by Cocco and Monasson have been very thorough, distribution of software through its source code can scare all but the most tech savy scholars in the field.

Moreover the algorithm was originally written in non-standard Fortran 77 that was only compatible with g77 and not with gfortran. As of version 3.4 of GCC (released in May 2006) development of g77 has stopped and users are encouraged to use gfortran.

Because of this most modern Linux distributions do not come with g77 preinstalled and users who need g77 have to install it separately sometimes compiling the compiler itself.

Another important hurdle to overcome before the public release of this software was the fact that it originally contained parts of code that were covered by copyright [Press 86] and could not be reused freely.

An important part of the implementation of the original program relied on the implementation of the classical Newton method for multidimensional minimzation. Knowing that the function to minimize is convex in the parameters guarantees the convergence of the algorithm, howver it is not at all clear that among all minimization techniques Newton's would be the faster in all cases.

In fact Newton's method relies on the exact computation of the Hessian matrix which is computatinally taxing and sometimes possible only in an approximate form, because of this we have chosen to reimplement the software in a way that permitted a modular change of minimzation techniques.

Further limitations included the lack of inference of certain parameters of the model such as the leaking constant g, which was fixed at the beginning of the inference and considered equal for all neurons.

Because of this we have decided to translate the problem in standard C which is a far more widespread programming language, arguably the most common. Compilers in C are available on virtually every architecture, and there are a variety of free open-source numerical libraries that can be effortlessly used for the implementation of minimization techniques and special functions.

Morevore standard C code is very easily integrated in more higher level computational software such as Matlab which is very widely used in the computational biology community.

We believe that these contributions, even though it is not an algorithmic effort, but of a more mundane nature can be of great use in the diffusion of this algorithm and its use.

A.4 Description of the software package

The software package is written in standard C and the source code will soon be available for download. C was the natural choice for a program that can be used either as a stand-alone executable or called from widely used computational software such as Matlab [MathWorks 11] and Mathematica [Wolfram Research 11].

The software is composed of several functions:

- A function that reads data in the form of spike trains, *i.e.* a two column file where the

first column is the time at which the spike was emitted and the second is an integer value that identifies the neuron responsible for that spike. Data are then stored in data structures of variable size, to be conveniently accessed by other functions.

- A function that goes through the data and identifies all spikes incoming to a cell in the time interval between two successive spikes of that cell. This function also performs a check on collision, that is, multiple spikes emitted by the same cell at the same time.

- A function that computes the log-likelihood and its first and second derivatives with respect to the interactions and the current. Note that the time-consuming calculation of second derivatives can be switched off if the minimization algorithm employed does not require them.

These functions have appropriate wrappers that allow for use with the minimization routines available from the GNU Scientific Library (GSL) [GSL 11] and Matlab. The choice of minimization technique and related parameter is left to the user, but we have observed the gnewton technique of the GSL to be the fastest in most cases.

The user chooses the values of the parameters of the LIF model (A.1); when the leaking constant g is set to zero, that is when the integrator is non-leaky, a specific and much faster program is used. Otherwise the most likely value for g can be inferred from the data for each neuron. Also available to the user specifications is a choice of priors over the coupling values, based on the L_1 and L_2 norms, which can be used to ensure convergence to realistic values and/or to eliminate couplings which are very close to zero. The program can be easily modified to add further specific priors. The user can further improve upon the instantaneous synaptic integration assumption in the model (A.1). To do so an option allows the user to introduce a synaptic reweighing function, replacing J_{ij} with $J_{ij} \times K(t_{i,k'} - t_{j,k})$, where $t_{j,k}$ is the time of the spike fired by cell j and entering cell i, and $t_{i,k'}$ is the next spiking time of cell i; $K(x) = 0$ for $x = 0$ and $K(x) \simeq 1$ for $x > \tau_s$, the synaptic integration time.

The output data can be printed to a file or to a Matlab array. The file is composed of three columns, the first two denote the indices of the coupling matrix J_{ij} and the third the value of the coupling constant. The diagonal elements of the matrix J_{ii} are the currents I_i.

Bibliography

[Adler 66]   J. Adler. *Chemotaxis in Bacteria*. Science, vol. 153, pages 708–716, 1966.

[Angelescu 08]   D. G. Angelescu, P. Linse, T. T. Nguyen & R. F. Bruinsma. *Structural transitions of encapsidated polyelectrolytes*. The European Physical Journal E - Soft Matter and Biological Physics, vol. 25, no. 3, pages 323–334, 2008.

[Ashkin 70]   A. Ashkin. *Acceleration and trapping of particles by radiation pressure*. Physical Review Letters, vol. 24, no. 4, pages 156–159, 1970.

[Ashkin 86]   A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm & S. Chu. *Observation of a single-beam gradient force optical trap for dielectric particles*. Optics letters, vol. 11, no. 5, pages 288–290, 1986.

[Atkinson 69]   M. R. Atkinson, M. P. Deutscher, A. Kornberg, A. F. Russell & J. G. Moffatt. *Enzymic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2', 3'-dideoxyribonucleotide*. Biochemistry, vol. 8, no. 12, pages 4897–4904, 1969.

[Averbeck 06]   B. B. Averbeck, P. E. Latham & A. Pouget. *Neural correlations, population coding and computation*. Nature Reviews Neuroscience, vol. 7, no. 5, pages 358–366, 2006.

[Baldazzi 06]   V. Baldazzi, S. Cocco, E. Marinari & R. Monasson. *Inference of DNA Sequences from Mechanical Unzipping: An Ideal-Case Study*. Physical Review Letters, vol. 96, no. 12, page 128102, 2006.

[Baldazzi 07]   V. Baldazzi, S. Bradde, S. Cocco, E. Marinari & R. Monasson. *Inferring DNA sequences from mechanical unzipping data: the large-bandwidth case*. Physical Review E, vol. 75, no. 1, page 011904, 2007.

[Balkovsky 02]   E. Balkovsky & B. I. Shraiman. *Olfactory search at high Reynolds number*. Proceedings of the National Academy of Sciences, vol. 99, no. 20, page 12589, 2002.

[Barany 91]   F. Barany. *The ligase chain reaction in a PCR world*. Genome Research, vol. 1, no. 1, page 5, 1991.

[Barbieri 07]   C. Barbieri. Infotassi: un algoritmo di ricerca senza gradienti. Master's thesis, Università di Roma "La Sapienza", December 2007.

[Barbieri 09]   C. Barbieri, S. Cocco, R. Monasson & F. Zamponi. *Dynamical modeling of molecular constructions and setups for DNA unzipping.* Physical Biology, vol. 6, page 025003, 2009.

[Barbieri 11]   C. Barbieri, S. Cocco & R. Monasson. *On the trajectories and performance of Infotaxis, an information-based greedy search algorithm.* Europhysics Letters, vol. 94, no. 2, page 20005, 2011.

[Bayes 63]   T. Bayes. *An Essay towards Solving a Problem in the Doctrine of Chances.* Philosophical Transactions of the Royal Society, vol. 53, pages 370–418, 1763.

[Bayes 58]   T. Bayes. *An Essay towards Solving a Problem in the Doctrine of Chances. Reprint in modern notation of [Bayes 63].* Biometrika, vol. 45, pages 293–315, 1958.

[Berg 75]   H. C. Berg. *Chemotaxis in Bacteria.* Annual Review of Biophysics and Bioengineering, vol. 4, pages 119–136, 1975.

[Berg 88]   H. C. Berg. *A physicist looks at bacterial chemotaxis.* In Cold Spring Harbor Symposium on Quantitative Biology, volume 53, pages 1–9, 1988.

[Bouchiat 99]   C. Bouchiat, M. D. Wang, J. F. Allemand, T. Strick, S. M. Block & V. Croquette. *Estimating the persistence length of a worm-like chain molecule from force-extension measurements.* Biophysical journal, vol. 76, no. 1, pages 409–413, 1999.

[Breslauer 86]   K. J. Breslauer, R. Frank, H. Blöcker & L. A. Marky. *Predicting DNA duplex stability from the base sequence.* Proceedings of the National Academy of Sciences, vol. 83, no. 11, page 3746, 1986.

[Cerdà 05]   J. J. Cerdà, T. Sintes & A. Chakrabarti. *Excluded volume effects on polymer chains confined to spherical surfaces.* Macromolecules, vol. 38, no. 4, pages 1469–1477, 2005.

[Clarke 09]   J. Clarke, H. C. Wu, L. Jayasinghe, A. Patel, S. Reid & H. Bayley. *Continuous base identification for single-molecule nanopore DNA sequencing.* Nature nanotechnology, vol. 4, no. 4, pages 265–270, 2009.

[Cocco 01]   S. Cocco, R. Monasson & J. F. Marko. *Force and kinetic barriers to unzipping of the DNA double helix.* Proceedings of the National Academy of Sciences, vol. 98, no. 15, page 8608, 2001.

[Cocco 03]   S. Cocco, J. F. Marko & R. Monasson. *Slow nucleic acid unzipping kinetics from sequence-defined barriers.* The European Physical Journal E - Soft Matter and Biological Physics, vol. 10, no. 2, pages 153–161, 2003.

[Cocco 09]   S. Cocco, S. Leibler & R. Monasson. *Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods.* Proceedings of the National Academy of Sciences, vol. 106, no. 33, page 14058, 2009.

[Crothers 64]   D. M. Crothers & B. H. Zimm. *Theory of the melting transition of synthetic polynucleotides: Evaluation of the stacking free energy**. Journal of Molecular Biology, vol. 9, no. 1, pages 1–9, 1964.

[Curtis 02]   J. E. Curtis, B. A. Koss & D. G. Grier. *Dynamic holographic optical tweezers*. Optics Communications, vol. 207, no. 1-6, pages 169–175, 2002.

[Danilowicz 03]   C. Danilowicz, V. W. Coljee, C. Bouzigues, D. K. Lubensky, D. R. Nelson & M. Prentiss. *DNA unzipped under a constant force exhibits multiple metastable intermediates*. Proceedings of the National Academy of Sciences, vol. 100, no. 4, page 1694, 2003.

[Doi 86]   M. Doi & S. F. Edwards. The Theory of polymer dynamics. Numeéro 73 in International Series of Monographs on Physics. Clarendon Press, 1986.

[Elowitz 02]   M.B. Elowitz, A.J. Levine, E.D. Siggia & P.S. Swain. *Stochastic gene expression in a single cell*. Science, vol. 297, no. 5584, page 1183, 2002.

[Eyring 35]   H. Eyring. *The activated complex in chemical reactions*. The Journal of Chemical Physics, vol. 3, page 107, 1935.

[Flory 53]   P. J. Flory. Principles of polymer chemistry. George Fisher Baker non-resident lectureship in chemistry at Cornell University. Cornell University Press, 1953.

[Friedman 06]   J. M. Friedman, Á. Baross, A. D. Delaney, A. Ally, L. Arbour, J. Asano, D. K. Bailey, S. Barber, P. Birch, M. Brown-John *et al.* *Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation*. The American Journal of Human Genetics, vol. 79, no. 3, pages 500–513, 2006.

[Gal 80]   S. Gal. Search games. Academic Press, 1980.

[Glessner 10]   J. T. Glessner, M. P. Reilly, C. E. Kim, N. Takahashi, A. Albano, C. Hou, J. P. Bradfield, H. Zhang, P. Sleiman, J. H. Flory *et al.* *Strong synaptic transmission impact by copy number variations in schizophrenia*. Proceedings of the National Academy of Sciences, vol. 107, no. 23, page 10584, 2010.

[Gosse 02]   C. Gosse & V. Croquette. *Magnetic tweezers: micromanipulation and force measurement at the molecular level*. Biophysical Journal, vol. 82, no. 6, pages 3314–3329, 2002.

[GSL 11]   GSL. *GSL documentation – Multidimensional Root-finding.* http://www.gnu.org/software/gsl/manual/html_node/ Multidimensional-Root_002dFinding.html, 2011.

[Huelsenbeck 01]   J.P. Huelsenbeck, F. Ronquist, R. Nielsen & J.P. Bollback. *Bayesian inference of phylogeny and its impact on evolutionary biology*. Science, vol. 294, no. 5550, page 2310, 2001.

[Huguet 09]   J. M. Huguet, N. Forns & F. Ritort. *Statistical properties of metastable intermediates in DNA unzipping.* Physical review letters, vol. 103, no. 24, page 248106, 2009.

[Huguet 10]   J. M. Huguet, C. V. Bizarro, N. Forns, S. B. Smith, C. Bustamante & F. Ritort. *Single-molecule derivation of salt dependent base-pair free energies in DNA.* Proceedings of the National Academy of Sciences, vol. 107, no. 35, page 15431, 2010.

[Iafrate 04]   A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer & C. Lee. *Detection of large-scale variation in the human genome.* Nature genetics, vol. 36, no. 9, pages 949–951, 2004.

[James 43]   H. M. James & E. Guth. *Theory of the elastic properties of rubber.* The Journal of Chemical Physics, vol. 11, page 455, 1943.

[Jolivet 04]   R. Jolivet, T. J. Lewis & W. Gerstner. *Generalized integrate-and-fire models of neuronal activity approximate spike trains of a detailed model to a high degree of accuracy.* Journal of Neurophysiology, vol. 92, no. 2, page 959, 2004.

[Keller 70]   E. F. Keller & L. A. Segel. *Initiation of Slime Mold Aggregation Viewed as an Instability.* Journal of Theoretical Biology, vol. 26, no. 3, pages 399–415, 1970.

[Kerker 69]   M. Kerker. The scattering of light and other electromagnetic radiation. Academic Press, 1969.

[Koike 11]   A. Koike, N. Nishida, D. Yamashita & K. Tokunaga. *Comparative analysis of copy number variation detection methods and database construction.* BMC genetics, vol. 12, no. 1, page 29, 2011.

[Kramers 40]   H. A. Kramers. *Brownian motion in a field of force and the diffusion model of chemical reactions.* Physica, vol. 7, no. 4, pages 284–304, 1940.

[Kuhn 42]   W. Kuhn & F. Grün. *Beziehungen zwischen elastischen Konstanten und Dehnungsdoppelbrechung hochelastischer Stoffe.* Colloid & Polymer Science, vol. 101, no. 3, pages 248–271, 1942.

[Lansky 08]   P. Lansky & S. Ditlevsen. *A review of the methods for signal estimation in stochastic diffusion leaky integrate-and-fire neuronal models.* Biological cybernetics, vol. 99, no. 4, pages 253–262, 2008.

[Levene 03]   M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead & W. W. Webb. *Zero-mode waveguides for single-molecule analysis at high concentrations.* Science, vol. 299, no. 5607, page 682, 2003.

[MacKay 05]   D. J. C. MacKay. Information theory, inference, and learning algorithms. Cambridge University Press, 4th edition, 2005.

[Mangeol 08]   P. Mangeol & U. Bockelmann. *Interference and crosstalk in double optical tweezers using a single laser source.* Review of Scientific Instruments, vol. 79, page 083103, 2008.

[Manosas 05]   M. Manosas & F. Ritort. *Thermodynamic and kinetic aspects of RNA pulling experiments*. Biophysical journal, vol. 88, no. 5, pages 3224–3242, 2005.

[Marko 94]   J. F. Marko & E. D. Siggia. *Bending and twisting elasticity of DNA*. Macromolecules, vol. 27, no. 4, pages 981–988, 1994.

[Marko 95]   J. F. Marko & E. D. Siggia. *Stretching dna*. Macromolecules, vol. 28, no. 26, pages 8759–8770, 1995.

[Masson 09]   J. B. Masson, M. B. Bechet & M. Vergassola. *Chasing information to search in random environments*. Journal of Physics A: Mathematical and Theoretical, vol. 42, page 434009, 2009.

[MathWorks 11]   MathWorks. *Matlab support – MEX-files guide*. `http://www.mathworks.com/support/tech-notes/1600/1605.html`, 2011.

[Maxam 77]   A. M. Maxam & W. Gilbert. *A new method for sequencing DNA*. Proceedings of the National Academy of Sciences, vol. 74, no. 2, page 560, 1977.

[McKernan 09]   K. J. McKernan, H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, Jo. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De La Vega & A. P. Blanchard. *Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding*. Genome Research, vol. 19, no. 9, pages 1527–1541, 2009.

[Monasson 11]   R. Monasson & S. Cocco. *Fast inference of interactions in assemblies of stochastic integrate-and-fire neurons from spike recordings*. Journal of Computational Neuroscience, pages 1–29, 2011. 10.1007/s10827-010-0306-8.

[Moroz 97]   J. D. Moroz & P. Nelson. *Torsional directed walks, entropic elasticity, and DNA twist stiffness*. Proceedings of the National Academy of Sciences, vol. 94, no. 26, page 14418, 1997.

[Mossa 10]   A. Mossa, J. M. Huguet & F. Ritort. *Investigating the thermodynamics of small biosystems with optical tweezers*. Physica E: Low-dimensional Systems and Nanostructures, vol. 42, no. 3, pages 666–671, 2010.

[Mullis 86]   K. B. Mullis, F. A. Faloona, S. J. Scharf, R. K. Saiki, G. T. Horn & H. Erlich. *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction*. In Cold Spring Harbor Symposia on Quantitative Biology, volume 51, page 263. Cold Spring Harbor Laboratory Press, 1986.

[Mullis 94]   K. B. Mullis, F. Ferré & R. A. Gibbs. *The polymerase chain reaction*. Birkhauser Boston Inc., 1994.

[Odijk 95]   T. Odijk. *Stiff chains and filaments under tension*. Macromolecules, vol. 28, no. 20, pages 7016–7018, 1995.

[Paninski 04]   L. Paninski, J. W. Pillow & E. P. Simoncelli. *Maximum likelihood estimation of a stochastic integrate-and-fire neural encoding model*. Neural Computation, vol. 16, no. 12, pages 2533–2561, 2004.

[Paninski 06]   L. Paninski. *The most likely voltage path and large deviations approximations for integrate-and-fire neurons*. Journal of Computational Neuroscience, vol. 21, no. 1, pages 71–87, 2006.

[Press 86]   W. H. Press, S. A. Teukolsky, W. T. Wetterling & B. P. Flannery. Numerical recipes in fortran 77. the art of scientific computing. Cambridge University Press, 2 edition, 1986.

[Raper 35]   K. B. Raper. Dictyostelium discoideum*, a new species of slime mold from decaying forest leaves*. Journal of Agricultural Research, vol. 50, pages 135–147, 1935.

[Raper 40]   K. B. Raper. *Pseudoplasmodium formation and organization in* Dictyostelium discoideum. Journal of the Elisha Mitchell Scientific Society, vol. 56, pages 241–282, 1940.

[Redner 01]   S. Redner. A guide to first-passage processes. Cambridge University Press, 2001.

[Ronaghi 96]   M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén & P. Nyrén. *Real-time DNA sequencing using detection of pyrophosphate release*. Analytical biochemistry, vol. 242, no. 1, pages 84–89, 1996.

[Ronaghi 98]   M. Ronaghi, M. Uhlén & P. Nyrén. *A sequencing method based on real-time pyrophosphate*. Science, vol. 281, no. 5375, pages 363–365, 1998.

[Rouse Jr 53]   P. E. Rouse Jr. *A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers*. Journal of Chemical Physics, vol. 21, pages 1272–1280, 1953.

[Sanger 75]   F. Sanger & A. R. Coulson. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase*. Journal of Molecular Biology, vol. 94, no. 3, pages 441–446, 1975.

[Sanger 77]   F. Sanger, S. Nicklen & A. R. Coulson. *DNA sequencing with chain-terminating inhibitors*. Proceedings of the National Academy of Sciences, vol. 74, no. 12, page 5463, 1977.

[Sebat 04]   J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi*et al. Large-scale copy number polymorphism in the human genome*. Science, vol. 305, no. 5683, page 525, 2004.

[Sebat 07]   J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall *et al. Strong association of de novo copy number mutations with autism.* Science, vol. 316, no. 5823, page 445, 2007.

[Segall 86]   J. E. Segall, S. E. Block & H. C. Berg. *Temporal Comparisons in Bacterial Chemotaxis.* Proceedings of the National Academy of Sciences, vol. 83, no. 23, pages 8987–8991, 1986.

[Shaffer 53]   B. M. Shaffer. *Aggregation in cellular slime molds:* in vitro *isolation of acrasin.* Nature, vol. 171, pages 975–977, 1953.

[Shlien 10]   A. Shlien & D. Malkin. *Copy number variations and cancer susceptibility.* Current opinion in oncology, vol. 22, no. 1, page 55, 2010.

[Slosar 06]   A. Slosar & R. Podgornik. *On the connected-charges Thomson problem.* Europhysics Letters, vol. 75, page 631, 2006.

[Smith 92]   S. B. Smith, L. Finzi & C. Bustamante. *Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads.* Science, vol. 258, no. 5085, pages 1122–1126, 1992.

[Smith 96]   S. B. Smith, Y. Cui & C. Bustamante. *Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules.* Science, vol. 271, no. 5250, page 795, 1996.

[Smoluchowski 17] M. V. Smoluchowski. *Versuch einer mathematischen Theorie des Koagulationeskinetik kolloider Lösungen.* Zeitschrift für physicalische Chimie, vol. 92, pages 129–168, 1917.

[Staden 79]   R. Staden. *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Research, vol. 6, no. 7, page 2601, 1979.

[Stokes 51]   G. G. Stokes. *On the Effects of the Internal Friction of Fluids on the Motion of Pendulums.* Transactions of the Cambridge Philosophical Society, vol. IX, page 8, 1851.

[Sundaram 10]   S. K. Sundaram, A. M. Huq, B. J. Wilson & H. T. Chugani. *Tourette syndrome is associated with recurrent exonic copy number variants.* Neurology, vol. 74, no. 20, page 1583, 2010.

[Taketani 06]   M. Taketani & M. Baudry. Advances in network electrophysiology: using multi-electrode arrays. Springer, 2006.

[Tinoco 71]   I. Tinoco, O. C. Uhlenbeck & M. D. Levine. *Estimation of secondary structure in ribonucleic acids.* Nature, vol. 230, no. 5293, pages 362–367, 1971.

[Tinoco 73]   I. Tinoco, P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers & J. Gralla. *Improved estimation of secondary structure in ribonucleic acids.* Nature, vol. 246, no. 150, pages 40–41, 1973.

[Uhlenbeck 30]   G. E. Uhlenbeck & L. S. Ornstein. *On the theory of the Brownian motion.* Physical Review, vol. 36, no. 5, page 823, 1930.

[Vergassola 07a]   M. Vergassola. Private communication. 2007.

[Vergassola 07b]   M. Vergassola, E. Villermaux & B. I. Shraiman. *'Infotaxis' as a Strategy for Searching without Gradients.* Nature, vol. 445, pages 406–409, 2007.

[Vickers 94]   N. J. Vickers & T. C. Baker. *Reiterative responses to single strands of odor promote sustained upwind flight and odor source location by moths.* Proceedings of the National Academy of Sciences, vol. 91, no. 13, page 5756, 1994.

[Viterbi 67]   A. Viterbi. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Transactions on Information Theory, vol. 13, no. 2, pages 260–269, 1967.

[Wallmark 57]   J. T. Wallmark. *A new semiconductor photocell using lateral photoeffect.* Proceedings of the IRE, vol. 45, no. 4, pages 474–483, 1957.

[Wiedmann 94]   M. Wiedmann, W. J. Wilson, J. Czajka, J. Luo, F. Barany & C. A. Batt. *Ligase chain reaction (LCR)–overview and applications.* Genome Research, vol. 3, no. 4, page S51, 1994.

[Wolfram Research 11]   Wolfram Research. *Mathematica support C/C++ Language Interface.* http://reference.wolfram.com/mathematica/guide/CLanguageInterface.html, 2011.

[Woodside 06a]   M. T. Woodside, P. C. Anthony, W. M. Behnke-Parks, K. Larizadeh, D. Herschlag & S. M. Block. *Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid.* Science, vol. 314, no. 5801, page 1001, 2006.

[Woodside 06b]   M. T. Woodside, W. M. Behnke-Parks, K. Larizadeh, K. Travers, D. Herschlag & S. M. Block. *Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins.* Proceedings of the National Academy of Sciences, vol. 103, no. 16, pages 6190–6195, 2006.

[Wu 89]   D. Y. Wu & R. B. Wallace. *The ligation amplification reaction (LAR)–amplification of specific DNA sequences using sequential rounds of template-dependent ligation.* Genomics, vol. 4, no. 4, pages 560–569, 1989.

[Zou 05]   M. Zou & S.D. Conzen. *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.* Bioinformatics, vol. 21, no. 1, page 71, 2005.